Association of Test Publishers
1201 Pennsylvania Avenue, Suite 300, Washington, D.C. 20004 Tel: (866) - 240 -7909

Journal of Applied Testing Technology
August 1999 Issue

Increasing the Validity of Adapted Tests: Myths to be Avoided and Guidelines for Improving Test Adaptation Practices[1,2]

Ronald K. Hambleton and Liane Patsula
University of Massachusetts at Amherst

Adapting or translating achievement, ability, and personality tests and questionnaires prepared in one language and culture into other languages and cultures has had a long history in educational and psychological testing though this fact is not well-known among educational researchers and measurement specialists. At least five reasons can be found in the literature for adapting tests:

1. very often adapting a test is considerably cheaper and faster than constructing a new test in a second language,
2. when the purpose for the adapted test is cross-cultural or cross-national assessment (such as with many credentialing exams), an adapted test is the most effective way to produce an equivalent test in a second language,
3. there may be a lack of expertise for developing a new test in a second language,
4. there is a sense of security that is associated with an adapted test more so than a newly constructed test especially when the original test is well-known, and
5. fairness to examinees often results from the presence of multiple language versions of a test (see Hambleton & Patsula, 1998).

Unfortunately, though the practice of adapting or (simply) translating tests can be traced to the intelligence tests of the French psychologist Alfred Binet at the beginning of this century, there is substantial evidence to suggest that improved methods for adapting or translating tests from one language and culture to others are needed, and that considerably more attention should be given to this important task than it is typically given by researchers and/or test developers. Too often in practice the test adaptation process seems to be viewed as a routine task that can be completed by anyone who knows the relevant languages. One consequence is adapted tests in the target languages of interest with only superficial equivalence to the tests in the source language.

The purposes of this paper are two-fold: First, a set of myths or problems which need to be discarded/overcome about the test adaptation process will be presented. These myths are widespread and undermine effective test adaptation initiatives. Second, steps for adapting tests will be offered along with a discussion of the importance of each step in the test adaptation process.

Myths About Adapting Tests

There are a number of myths associated with adapting tests which appear in measurement practice and should be discarded as quickly as possible:

Myth 1. The preferable strategy is always to adapt an existing test rather than develop a new test for a second language group.

There are many good reasons for adapting a test, but there are reasons for not proceeding with a test adaptation as well. Especially when cross-cultural comparisons are not of interest, it may be substantially easier and more relevant to construct a new test for a second language group. This avoids any complications with copyright, insures that the format will be suitable, and any desired modifications in the definition of the construct of interest can be made at the outset of the test development process.

Sometimes, too, it may be desirable not to adapt a test but rather to require all examinees to take a test in a single language. For example, in the United States, there has been interest in some states in making high school graduation tests available in both English and Spanish. Technically this is possible, but the question of whether or not to make two language versions of a test available depends on many factors including the definition of the construct being measured. Is the language in which performance is to be demonstrated a part of the construct definition or not? In the case of reading, reading in the language of English is almost always part of the construct of interest. Producing a Spanish equivalent version of a reading test in English makes very little sense because inferences of English reading proficiency cannot be made from a test administered in Spanish.

The situation with a mathematics test may be very different. The construct of interest may be focused on computation skills, concepts, and problem-solving skills and here, the purpose of the test is to look for a demonstration of the skills, and the language in which the performance is assessed and demonstrated may be of little or no interest. Of course, if the desired inference is mastery of mathematics skills when the test questions are presented in English, then a Spanish version of the test would be inappropriate in this situation too.

Myth 2. Anyone who knows the two languages can produce an acceptable translation of a test.

This is one of the most troublesome myths because it results in unqualified persons adapting tests. There is considerable evidence suggesting that test translators need to be familiar with both source and target languages and the cultures, and they need to be generally familiar with the construct being assessed, and the principles of good test development practices. How, for example, can someone translate a high school physics test from English into Spanish without some knowledge of the content? Would a translator with little knowledge of test development principles be aware to preserve the relevant features of the original test in an adapted test such as clearly written item stems, a single correct or best answer, answer choices of approximately the same length, etc.?

Myth 3. A well-translated test guarantees that the test scores will be valid in a second language or culture for cross-language comparative purposes.

Van de Vijver and Poortinga (1997) make the point that not only should the meaning of a test be consistent across persons within a language group and culture but, that meaning, whatever it is, must be consistent across language groups and cultures. For example, if a test is more speeded in a second language version because of the nature of that language, then the two language versions of the test are not equally valid. We have encountered just such a problem in some German test translations we are currently working on. Quite simply, the German words are longer than English words and take correspondingly longer to read. The result is a slightly more speeded German version of the test. In this instance, the test may be equally valid in each language group and culture, but still not be suitable for cross-cultural comparisons.

Many other examples could be introduced. For one, the non-equivalent familiarity of students in different cultures with certain item formats, e.g., the multiple-choice format, places examinees from this second cultural group at a serious disadvantage. The translation could be excellent, but the scores from the two language versions are not equally valid.

Myth 4. Constructs are universal, and therefore all tests can be translated into other languages and cultures.

One of the best counter examples of this myth concerns intelligence tests. The Western notion of intelligence places considerable emphasis on speed of response. In some cultures, speed of response is of minor importance as a operating characteristic for life, and members of these cultural groups often score lower on Westernized intelligence tests because of a failure to perform quickly. But, it only in this limited sense of the Western definition of the construct of intelligence that these cultural groups appear of less intelligence. By another definition, perhaps one that devalues speed of response and emphasizes other human attributes of intelligence (see Sternberg and Gardner (1983) for broader definitions of intelligence which incorporate, for example, social and artistic skills) the results would be opposite.

There is currently considerable interest in cross-cultural comparisons of quality of life. It is interesting to discover that the construct associated with quality of life in this country is often very different in other countries and this makes cross-cultural comparisons very different. Televisions, portable telephones, personal computers, the great outdoors, and college sports are of no importance and do not affect the quality of life for persons in many other cultures. Cross-cultural comparisons of quality of life are difficult to carry out because the construct may have very different meanings across cultures. Myth 5. Translators are capable of finding flaws in a test adaptation. Field testing is not usually necessary. This is another of the major myths about adapting tests. There are literally thousands of examples of poorly adapted test items in the literature, and many of the items in these tests were approved by translators. The fact is translators are not able to anticipate all of the problems encountered by examinees taking a test in a second language.

One of the best examples because it was discovered on an international comparative study of reading achievement (and a study where the American students were about the middle of 20 countries) is the following:

Determine whether these two words are similar or different--

pessimistic - sanguine

In the English version of the test item, only about 54% of the American students were able to determine the correct response (a performance level slightly above chance) which is that the two words have a different meaning. In a second language version, the item was adapted as follows—

pessimistic - optimistic

In the foreign language version of the test item, almost 100% of the examinees answered the item correctly. Clearly, a poor translation had made the test item considerably easier. The reason given was that the word "sanguine" had no equivalent word in the second language and therefore another word was chosen which too, had a different meaning to pessimistic. Interestingly, this easier version of the test item was used in the country which finished number one among the 20 countries. One wonders what role this item and other improperly adapted test items played in the final rankings of the 20 countries.

In summary, all of the myths can seriously compromise the validity of a test in a second language or cultural group, or negatively influence the validity of adapted tests for use in cross-language comparison studies. Fortunately, each myth is straightforward to address in practice. What follows are steps for adapting tests which should eliminate all of the myths and other shortcomings in test adaptation methodology.

Steps for Adapting Tests

The International Test Commission (ITC) guidelines (Hambleton, 1994; van de Vijver & Hambleton, 1996) provide an excellent framework to guide researchers in the test adaptation process. Appendix A contains a copy of those guidelines. The following steps for adapting a test from one culture and/or language for use in another are a mixture of findings and recommendations from the ITC guidelines and many empirical studies (e.g., Angoff & Cook, 1988, Prieto, 1992; Hambleton, 1994). Geisinger's (1994) work in crosscultural assessment was especially influential in our thinking about the topic of steps for adapting tests. The steps are still evolving. Through the application of the steps indifferent contexts new insights will be gained and certain additions, deletions, and clarifications may be necessary.

Step 1 – Ensure that construct equivalence exists in the language and cultural groups of interest.

Assess whether construct equivalence exists between the cultures of interest and if it does not, either consider "decentering" (that is, revising the definition of the construct to be equally equivalent in each language and cultural group) or discontinue the project. The publication by Harkness (1998) is especially helpful in the study of construct equivalence.

Central questions are as follows:

- ✓ Does the particular construct that a researcher (e.g., the content domain for a credentialing exam) is interested in measuring exist in both cultures?
- ✓ Does it make sense to compare these two cultures on this construct?
- ✓ Would any cross-cultural comparison on this construct be meaningful?
- ✓ Does the construct that is being measured mean the same thing in all cultures being compared?

Researchers familiar with both languages and cultures are in a strong position to make judgments about construct equivalence between cultures. One can also judge whether cross-cultural construct equivalence exists by interviewing or observing people from the cultures of interest, researching the cultures of interest, asking others who know about the cultures, or visiting people in the culture.

Suggestions:

- ✓ Through discussions with psychologists and other knowledge persons in each culture, determine if the construct exists, and if the same definition applies equally well in both language and cultural groups.

Step 2 – Decide whether to adapt an existing test or develop a new test.

Consider the purpose of the adapted test, and the advantages and disadvantages of adapting an existing test rather than developing a new test. It is clear too that some tests will be more amenable to translation into certain languages than others (Ahluwalia, 1990, p. 20). The more similar the target language and/or culture are to the source language and/or culture, the easier will be the adaptation (thus, English to Spanish adaptations may make more sense than English to Arabic or English to Chinese adaptations). With tests intended for cross-cultural comparisons, test adaptation (possibly with some decentering) may be the only option. But when cross-cultural comparisons are not of interest, it may be easier to actually produce a new test that meets the cultural parameters in the second language group, than to adapt an already existing test which may have a number of shortcomings (e.g., a less than satisfactory definition of the construct, inappropriate item formats, use of some cultural specific content, etc.).

The standards with which to evaluate whether to adapt an existing test require some level of expertise in measurement, some knowledge of the relevant literature of the original test, and some knowledge of the language and culture to which the test is being adapted.

Suggestions:

- ✓ Consider the purpose of the adapted test, and carefully consider the advantages and disadvantages of adapting a test versus constructing a new test.

Step 3 – Select well-qualified translators.

This is often one of the major shortcomings of a test adaptation project. Two points can be made:

First, in selecting translators, search for persons who are fluent in both languages and who are very familiar with the cultures under study, and who have some knowledge of test construction and the construct being measured. As knowledge of test construction practices is not common among translators, this may be addressed with some training prior to initiating the test adaptation process. Adding a psychometrician to the mix may be desirable, too.

Second, some researchers have found that panels or committees of people translate the test better than individuals. Committees produce pooled adaptations that are often more accurate than translations from a single translator.

Suggestions:

- ✓ Seek out translators with language proficiency, knowledge of the relevant cultures, and some subject matter knowledge/knowledge of the construct of interest.
- ✓ Involve more than one translator in the process to provide a mix of perspectives and to enable checking to be conducted.

Step 4 – Translate and adapt the test.

One approach to increasing the likelihood of a valid test adaptation is to adopt one of the two (or both) standard designs: forward- and back-translation. Forward translation designs are the most technically sound because the focus of the review is on both the source and target language versions of the test. Backward translation designs can also be revealing of poor translations but without a focus on the target language version of the test, problems in the adaptation can be missed. For example, with a hard-to-translate concept like "ice hockey" into Chinese, these English words may be used in the adapted version. They are very easy to back translate, but they may be quite meaningless in the target language version of the test.

Suggestions:

- ✓ Use a forward translation design but a backward translation design can be useful too, but not as the only design.

Step 5 – Review the adapted version of the test and make necessary revisions.

In a forward translation design, another set of translators examine the adapted version of the test for any errors that may lead to differences in meaning between the two language versions. The group of translators' focus at this point would be on the quality of the translation or adaptation of the test. As Geisinger (1994) suggests, this review can be accomplished in a group meeting, individually, or by some combination of individual and group work. Geisinger believes that the most effective strategy is to first have the translators review the items and react in writing and then to have the individuals share their comments with one another and to reconcile any differences in opinion and make any changes in the original and/or adapted language versions as necessary.

The National Institute for Testing and Evaluation in Israel is responsible for adapting college admissions tests into five languages from the original Hebrew-language version. One special feature in their process is that their translators work from the translated version first and attempt to determine the validity of the questions:

For example, is the stem clear? Is there a single correct answer? Are there grammatical clues that lead the test-wise candidate to the correct answer? After it is determined that the test items can stand on their own merits, then the equivalence of the adapted version and the original Hebrew version are compared. Translators look at several features of the adapted items:

- ✓ accuracy of the translation as well as the clarity of the sentences,
- ✓ the level of difficulty of the words, and
- ✓ the fluency of the translation.

With a backward translation design, translators would take the adapted version of the test, back translate to the source language, and then judgments would be made about the equivalence of the original and back-translated versions of the test. Where nonequivalence is identified, changes in the adapted version of the test are considered. The idea is that if the adaptation has been effective, the back-adapted version of the test should look very much like the original. Of course, when the adaptation involves format changes, time changes, and other changes, the target language version of the test may be fine, but a back-translated test may not look at all like the original. In general, backtranslation designs seem like an excellent supplement to the forward translation design, but they are not likely to be able to stand on their own. The information they provide about the validity of the adapted test is limited.

Based on the comments of the reviewers, changes can be made in the original and/or adapted version of the test, as necessary. Of course, if many changes are made, there may be advantages to repeating step 4 and 5.

Suggestions:

- ✓ Review and revision of the adapted test is absolutely necessary, following the initial translation. In most cases, the adapted test is too important to be dependent on the insights of a single translator or group of translators.

Step 6 – Conduct a small tryout of the adapted version of the test.

It is at this step that many studies seem to go wrong. Too many researchers and test developers feel that judgmental review is sufficient evidence to establish the validity of a test in a second language. But validity evidence for using a test in a second language depends on stronger evidence than that the test seems to look acceptable to translators and/or reviewers. Not only is empirical evidence needed to support the validity of inferences from an adapted version of a test, but perhaps two or more empirical studies are needed. A good example of what researchers might learn from a tryout of test items in a second language and culture is clearly highlighted in the paper by Allalouf and Sireci (1998).

Beginning with a small tryout of the adapted test seems to be prudent before investing considerable resources in a more ambitious field test. Pilot test the instrument using a small sample of individuals representative of the eventual target population and compare the results to results obtained from a source sample. The pilot test should consist of administering the test, as well as interviewing the individuals to obtain their criticisms of the test itself, instructions, time limits, etc. These findings form the basis for revising the test. One good suggestion from Ellis and Mead (1998) might be carried out at this point.

Ellis and Mead suggest that when there are disagreements about the best adaptation of a test item, these variations might all be field tested, and the results used to make the final decision about which adaptation is best.

Suggestions:

- ✓ Conduct a pilot test to gain preliminary information about the test, and revise accordingly.

Step 7 – Carry out a more ambitious field test.

This is one of the most important steps in the total test adaptation process. Good translators are often capable of identifying and fixing many shortcomings in adapted tests. But many problems go unidentified until test items are field tested. For example, in a recent study by Hambleton, Slater, and Yu (in press) in which National Assessment of Educational Progress (NAEP) mathematics items were adapted into Chinese, the NAEP test item went unidentified by the translators. A field test revealed a major problem with the item which could not be identified by the translators because it was a curriculum issue. Chinese students at the eighth grade were unfamiliar with the concept of estimation.

Field test the adapted test using a larger sample of individuals representative of the eventual target population and conduct preliminary statistical analyses, such as a reliability analysis and a classical item analysis. In addition, check for construct equivalence using factor analysis should be carried out.

Suggestions:

- ✓ Design and carry out an ambitious field test to check out test items (using classical or modern item analysis procedures), test and subtest reliabilities, and the factor structure of the test (factor analysis or structural equating modeling are popular for this analysis). Compare findings to those obtained with the source language version of the test.

Step 8 – Choose a statistical design for connecting scores on the source and target language versions of the test.

This step is necessary when cross-cultural comparisons are of interest, or the test score norms or performance standards (i.e., the passing score on a credentialing exam) with the source language version of the test are of interest with the target language version of the test. At this step (which might be combined with step 7), a linking design is needed to place the test scores from the different versions of the test on a common scale. There are three popular linking designs:

1. bilingual group design,
2. matched monolingual group design, and
3. monolingual group design.

All three designs are popular, though the third design may be the easiest to implement in practice (see, for example, Angoff & Cook, 1988). For a worked example based on item response modeling of the data, studies by Angoff and Cook (1988) or Woodcock and Munoz-Sandoval (1993) would be of special interest.

Suggestions:

- ✓ Choose a linking design to equate scores from the source and target language versions of the test. Item response modeling is a standard way to proceed. Large samples are highly desirable at this step to produce a stable linking of scores from one test to the other.

Step 9 – If cross-cultural comparisons are of interest, ensure equivalence of the language versions of the test.

This step, too, may be combined with steps 7 and 8. We have highlighted this activity as a step because of its central importance in the test adaptation process. Administer the source version of the test to a large sample of the source population and perform statistical analyses to determine whether or not the items function similarly in both the adapted and source language versions of the test. This is accomplished through the use of an item bias study (often called a "differential item functioning" or DIF study). If there are items that function differently for each group, rewrite or retranslate, readminister, and reanalyze those items to determine whether they function the same for both groups. The Muniz, Hambleton, and Xing (1998) study highlights the fact that even small samples (i.e., 50 candidates per group) can be useful in detecting flaws in the translation/adaptation process.

Suggestions:

- ✓ Conduct a DIF study using one or more of the standard statistical procedures--Mantel-Haenszel statistic, logistic regression, IRT-based area procedures, etc.

Step 10 – Perform validation research as appropriate.

Regardless of the interest in cross-cultural comparisons of scores from the two language versions of the test, and the related research generated by that concern, there is also a need to ensure that the test scores of the newly adapted test are valid and reliable. Step 1 involved judgmental strategies for collecting evidence of construct equivalence, as there was no data available with which to conduct statistical analyses. Now that the test has been administered, there are data available and so evidence of construct-related validity can be compiled. This may be compiled from factor analytic, experimental, or other correlational information (e.g., predictive or concurrent validity studies). Again, this step may be combined with steps 7 to 9.

Suggestions:

- ✓ Conduct empirical studies which address the equivalence of the multilanguage versions of the test in the populations where the test will be used. Evidence of construct equivalence as well as the absence of method and item bias are important.

Step 11 – Document the process and prepare a manual for the users of the adapted test.

Document results obtained from steps 1 to 10 and prepare a manual for the users of the adapted test. The manual should include specifics regarding the administration of the test, as well as how to interpret the test scores. This is a very important step, yet often overlooked.

Suggestions:

- ✓ Document the full process of adapting a test. Everything from the persons involved, and designs used, to the findings and the nature of the changes which were made needs to be compiled and placed in a technical manual for future reference.

Step 12 – Train users.

Where possible, train the users of the test. Although documentation and a manual will assist users of the adapted instrument, training will further assist them.

Suggestions:

- ✓ Train test administrators to follow the directions and to answer any questions appropriately which may arise. Especially when cross-cultural comparisons are being made, or the norms for the target language version of the test are being used, standardized test administrations are essential across language groups.

Step 13 – Ongoing monitoring of the adapted test.

Often cross-cultural studies are a "one-shot affair." But some tests are adapted for ongoing use in a second language group. Popular intelligence, credentialing, aptitude, and personality tests would be ones which are adapted and intended for ongoing use. Researchers should remain vigilant to potential flaws in their adapted tests, and this means that ongoing monitoring of adapted tests is needed. Re-investigation and reevaluation of the reliability and validity of test scores should be ongoing.

Suggestions:

- ✓ Continue to monitor the evaluation of adapted tests and assess their reliability and validity on a regular basis. The reliability and validity of all tests can be expected to change over time due to changes in curriculum, values, experiences, exposure to the test, etc.

Conclusions

An increasing number of educational, credentialing, and psychological tests are being adapted for use in other languages and cultures. At the same time, these adapted tests will have limited value unless they are adapted with a high degree of concern for issues of usability, reliability and validity. There is a rapidly emerging psychometric literature on the topic of test adaptation methodology, and more advances can be expected in the coming years as researchers respond to the expanding need for adapted tests of high technical quality. Avoiding the five myths and following the 13 steps introduced in this paper for the test adaptation process should go a long way toward improving current practices. In addition, the 13 steps provide a framework for incorporating new methodology into the process as it is developed.

References

Ahluwalia, N. T. (1990). Comparability of translated tests in occupational testing. **CLEAR Exam Review, 1,** 19-21.

Allalouf, A., & Sireci, S. G. (1998, April). **Detecting sources of DIF in translated verbal items**. Paper presented at the meeting of AERA, San Diego.

Angoff, W. H., & Cook, L. L. (1988). **Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test** (Report No. 88-2). New York, NY: College Entrance Examination Board.

Ellis, B., & Mead, A. (1998, August). **Measurement equivalence of a 16PF Spanish translation: An IRT differential item and test functioning analysis**. Paper presented at the 24th meeting of the International Association of Applied Psychology, San Francisco.

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

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. **European Journal of Psychological Assessment**, **10**, 229-244.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. **Social Indicators Research**, **45,** 153-171.

Hambleton, R. K., Slater, S. C., & Yu, J. (in press). Field test of the ITC guidelines for adapting psychological tests. **European Journal of Psychological Assessment.**

Harkness, J. (Ed.). (1998), **Cross-cultural equivalence**. Mannheim, Germany: ZUMA.

Muniz, J., Hambleton, R. K., & Xing, D. (1998). **Small sample studies to detect flaws in test translation.** Paper presented at the meeting of AERA, San Diego.

Prieto, A. J. (1992*)*. A method for translation of instruments to other languages. **Adult Education Quarterly, 43,** 1-14**.**

Sternberg, R. L., & Gardner, M. K. (1983). Unities in inductive reasoning. **Journal of Experimental Psychology: General**, **112,** 80-116.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. **European Psychologist**, **1,** 89-99.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. **European Journal of Psychological Assessment**, **13,** 29-37**.**

Woodcock, R. W., & Munoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. **European Journal of Psychological Assessment**, **9,** 233-241**.**

Appendix A
ITC Test Adaptation Guidelines

Context

C.1 Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.

C.2 The amount of overlap in the constructs in the populations of interest should be assessed.

Test Development and Adaptation

D.1 Test developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the instrument are intended.

D.2 Test developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the instrument is intended.

D.3 Test developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations

D.4 Test developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.

D.5 Test developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions.

D.6 Test developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the instrument.

D.7 Test developers/publishers should apply appropriate statistical techniques to

1. establish the equivalence of the different versions of the instrument, and
2. identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations.

D.8 Test developers/publishers should provide information on the evaluation of validity in all target populations for whom the adapted versions are intended.

D.9 Test developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations.

D.10 Non-equivalent questions between versions intended for different populations should not be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content validity of scores reported for each population separately.

Administration

A.1 Test developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions.

A.2 Test administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.

A.3 Those aspects of the environment that influence the administration of an instrument should be made as similar as possible across populations for whom the instrument is intended.

A.4 Test administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.

A.5 The test manual should specify all aspects of the instrument and its administration that require scrutiny in the application of the test in a new cultural context.

A.6 The administrator should be unobtrusive and the administrator-examinee interaction should be minimized. Explicit rules that are described in the manual for the test should be followed.

Documentation/Score Interpretations

I.1 When a test is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence.

I.2 Score differences among samples of populations administered the test should not be taken at face value. The researcher has the responsibility to substantiate the differences with other empirical evidence.

I.3 Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.

I.4 The test developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance on the test, and should suggest procedures to account for these effects in the interpretation of results.

1) August 1999, Journal of Applied Testing Technology Volume1, No.1, 1-30.
2) Paper presented at the annual meeting of CLEAR, Denver, September, 1998.