TO TRANSLATE, OR NOT TRANSLATE: THAT IS THE QUESTION[1]

Guidelines for test adaptation

By Norman R. Hertz

Introduction

For many credentialing programs, whether regulatory or certification, the question of whether or not to translate or adapt an examination may generate considerable discussion. For other programs, the question may never arise because a law or policy establishes the answer regarding whether the exam must be translated or administered only in English. There are valid arguments on both sides of the issue. This article is intended to provide assistance to those agencies that are about to embark on adapting an examination.

Policy makers often believe that exams translated into the native language of the test taker are more fair to those who may be competent to practice, but who have a limited language competency in the language in which the exam was originally written. It is believed that by "translating" tests for test takers who are not necessarily fluent in the language of the original test that the tests become more equivalent for all test takers. Although this belief may not necessarily be accurate, credentialing programs may be faced with the mandate to translate examinations into one or more alternate languages.

There are well-developed technical guidelines established that should be applied when adapting examinations. There are also a variety of statistical procedures that should be applied to adapted tests in order to evaluate the equivalence of the adapted forms – some of these procedures are fairly basic while others are more sophisticated and require a larger population of test-takers in order to obtain statistically significant information.

This article will review considerations that should be followed when initiating the test adaptation process as well as the standard guidelines and statistical analyses used to obtain an accurately developed adapted examination. It will also provide information to assist programs required to translate examinations and to help reduce some of the errors inherent in the translation or adaptation process. The references included at the end of this article provide other sources that may be used to assist agencies through this process.

---

[1] Translation is the term commonly associated with converting an examination from one language to another. More recently, however, the preferred term used for this activity has become "test adaptation" since the examination may not be directly translated from one language to another. Test adaptation does not imply a literal word-to-word translation (Robin, Sireci & Hambleton, 2003), but provides for a recognition of cultural, content and language differences. The adaptation process is typically more flexible and allows for more complex word and situational substitutions so that the intended meaning is retained across languages even though the translation is not completely literal (Geisinger, 1994). Throughout this article, the two terms will be used interchangeably.

As with any test development activity, ethical standards (American Psychological Association, 1994) must be maintained throughout the adaptation process. The chapter by Oakland (2005) details selected standards of ethics that apply directly to test adaptation. Although both *Ethical principles* and Oakland's chapter specifically pertain to psychologists, many of the ethical standards are applicable across the psychometric field.

Points to Remember Before Embarking on Test Adaptation or Translation

Hambleton and Patsula (1999) provided the following caveats before beginning the translation or adaptation process:

- Adapting an existing examination rather than developing a new examination for a second language group is not necessarily the preferred strategy.

- An individual who knows the two languages may not necessarily be able to produce an acceptable translation of the examination.

- A well-translated examination does not guarantee that scores of candidates taking the examination in the second language will be valid.

- Test constructs (characteristics that an examination are designed to measure, which, in the case of credentialing examinations, is competency to practice a profession) are not universal. Therefore, not all examinations may be suitably translated into another language.

- Field testing should be implemented in all instances to ensure that the adaptation was appropriate.

Considerations to Follow When Adapting Examinations

The suggestions developed by Geisinger (1994) and Hambleton and Patsula (1999) for adapting examinations as well as the recommendations of Casillas and Robbins (2005) regarding the selection of individuals involved in the translation process have been combined into the following 14 steps. Although procedures for test adaptation continue to evolve, the following steps provide an excellent starting point for programs once the decision has been made to translate an examination.

1. Use translators who are fluent in both languages, who are extremely knowledgeable about both cultures, and who are familiar with the intended purpose of the examination. If translators are not competent with both cultures, hire cultural consultants to assist the translators. Provide the individuals who will be involved in the translation or adaptation with opportunities to enhance their cultural awareness and sensitivity on topics of historical and political determinants of the target culture.

2. Ensure that equivalence exists between the two languages and cultures for the constructs being measured.

3. Strive to portray different cultures in an accurate and positive light.

4. Review the translated version. Rather than just back-translating the exam, convene a group of individuals who are comparable to the original translators to review both versions of the examination to ensure comparability.

5. Revise the adapted examination based on the comments of the reviewers.

6. Pilot test the examination in order to determine any potential problems that may be encountered when the adapted examination is actually implemented.

7. Field test the examination. The examination should be administered to a representative sample of the population who will eventually be assessed with the instrument. Classical or item-response item analyses should be performed.

8. Conduct a fairness review to ensure that all examination materials provide an accurate reflection of the target population.

9. Consider the readability level of the adapted examination relative to the literacy rates as well as the technology literacy of the target population.

10. Standardize the scores. It is inappropriate to translate an examination into another language and then use statistical data based upon the original examination language results. Evidence of the comparability of both groups' performance is required before cross-cultural validity can be established.

11. Develop manuals and other documents for the users of the adapted examination and train the examination users.

12. Collect reactions from individuals who use these examinations.

13. Monitor the performance of the adapted examinations.

14. Perform validation research as appropriate.

Guidelines to Implement When Adapting Examinations

Researchers engaged in translation, adaptation, and cross-cultural studies report that there is no single source of information available about the methodology that should be applied when converting an examination from one language to another. Nevertheless, the International Testing Commission's (ITC) *International guidelines for test use*[2] provides the most comprehensive list of guidelines and should be considered the standards when adapting examinations. The entire set of guidelines, which must play a significant role during the adaptation process, is included as it serves as the criteria for evaluating the process of test adaptation.

[2] Copyright© 2001 International Test Commission. All rights reserved. Used with the permission of the International Test Commission. Personal communication with Jose Muniz, ITC President, November 15, 2004.

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 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.

Professional Standards

The defensibility of an examination is contingent upon it being developed and administered in accordance with the standards detailed in the 1999 *Standards for educational and psychological testing* (*Standards*). While the ITC guidelines provide sound methodology and represent the current level of practice, the requirements included in the *Standards* (1999) must also be satisfied in order for test developers and publishers to ensure that adapted tests are equivalent across the forms of the tests. The *Standards* include requirements for test development in general as well as three specific requirements for the translation of examinations.

- Testing practices should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences (Standard 9.1).

- When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguist groups to be tested (Standard 9.7).

- When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability (Standard 9.9).

Evaluating Adapted Examinations

Considerable information about test translation/adaptation can be found in articles in which the authors provide advice or recommendations based upon their research. Sireci (1997) reviewed the literature and reported that it has long been argued that the translation of a test from one language to another does not result in tests that are psychometrically equivalent. He states that translated items should not be considered equivalent without empirical evidence and further suggests that nonverbal items, or items minimally associated with linguistic content, provide an appealing source for identifying items that could be used to link tests. Likewise, Ellis (1989) stated that when cultural differences are under investigation, language differences create a measurement problem that may prohibit valid conclusions based on test scores. Because of these concerns, statistical analyses must be performed in order to ensure the integrity of the examination.

Robin, Sireci, and Hambleton (2003) suggested that the first step is to perform descriptive statistics to evaluate examinee performance across groups and technical characteristics of each language version of the test. Summary statistics such as the average and standard deviation of test scores, internal consistency reliability, and average item discrimination were shown to be valuable data in their research. If large differences are found in the summary statistics, the authors suggest that further investigations are warranted.

If the data permits, another way to evaluate the equivalence of a test is to examine differential item functioning (DIF) of the items. In general, DIF is a statistical procedure used to determine whether any differences in performance could be the result of socioeconomic background, age, gender, etc. It is also an appropriate measure to use with translated examinations in order to determine whether there are any differences in item performance between the original and the adapted examination using two comparable groups of examinees that are matched with respect to the construct being measured by the examinations (Dorans and Holland, 1993). The results from the DIF studies can be very informative in guiding the translation of tests even when there are an insufficient number of candidates to yield statistically meaningful results. The article by Gierl and Khaliq (2001) discusses some of the reasons why original and adapted tests may have differences in the way the items perform.

Summary

To ensure a valid second-language examination requires considerable time and funding and should not be undertaken without a thorough evaluation of the potential outcomes, both positive and negative. Once a decision has been made to adapt an examination and sufficient resources have been allocated, this article can assist the credentialing program to achieve a valid adapted examination that will determine an individual's competency to practice in a safe and effective manner.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (1994). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060-1073.

Casillas, A. & Robbins, S. B. (2005). Test adaptation and cross-cultural assessment from a business perspective: Issues and recommendations. *International Journal of Testing*, *5*, 5-21.

Dorans, N. J. & Holland, P. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds). *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

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

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments*. International Journal of Testing, 2*, 199-215.

Ercikan, K., Gierl, M. J., McCreith, Puhan, G. & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301-321.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 4*, 304-312.

Gierl, M. J. & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*, 164-187.

Hambleton, R. K., Sireci, S. G., & Robin, F. (1999). Adapting credentialing exams for use in multiple languages. *CLEAR Exam Review, 10*, 24-28.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*, 164-172.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1*, 93-114.

Oakland, T. & Lane H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4*, 239-252.

Oakland, T. (2005). Selected ethical issues relevant to test adaptations. In R. K. Hambleton, Merenda, P. F. & Spielberger, C. D., (Eds). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 65-92). Mahwah, NJ: Lawrence Erlbaum.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different versions of a credentialing examination. *International Journal of Testing, 3*, 1-20.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*, 12-19, 29.

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

Van de Vijver, F. J .R. & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, Merenda, P. F. & Spielberger, C. D., (Eds). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 65-92). Mahwah, NJ: Lawrence Erlbaum.