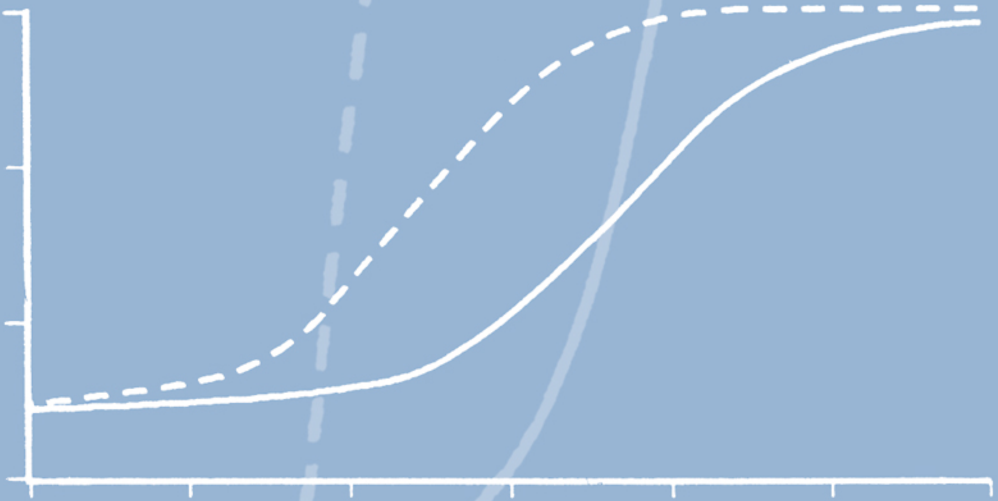
Adapting Educational and Psychological Tests for Cross-Cultural Assessment



Edited by

Ronald K. Hambleton

Peter F. Merenda

Charles D. Spielberger

Adapting Educational and Psychological Tests for Cross-Cultural Assessment

Adapting Educational and Psychological Tests for Cross-Cultural Assessment

Edited by

Ronald K.Hambleton
University of Massachusetts at Amherst

Peter F.Merenda
University of Rhode Island

Charles D.Spielberger
University of South Florida

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2005 Mahwah, New Jersey London

This edition published in the Taylor & Francis e-Library, 2009.

To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk.

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

Cover design by Sean Sciarrone

Library of Congress Cataloging-in-Publication Data

Adapting educational and psychological tests for
cross-cultural assessment/edited by Ronald K.Hambleton,
Peter F.Merenda, Charles D.Spielberger
p. cm.
Includes bibliographical references and index.
ISBN 0-8058-3025-1 (cloth: alk. paper)
1. Educational tests and measurements—Designs and construction—Cross-cultural studies. 2. Psychological tests—Design and construction—Cross-cultural studies. I. Hambleton, Ronald K. II. Merenda, Peter Francis, 1922– III. Spielberger, Charles Donald, 1927–

LB3060.65.A33 2005
371.26'1–dc22

2003060545
CIP

ISBN 1-4106-1175-2 Master e-book ISBN

Contents

Part I: Cross-Cultural Adaptation of Educational and Psychological Tests: Theoretical and Methodological Issues

Part II: Cross-Cultural Adaptation of Educational and Psychological Tests: Applications to Achievement, Aptitude, and Personality Tests

Preface

In 1989 I happened to read a report on the comparative levels of mathematics achievement of school children in five countries. The results surprised me, and so I began to wonder about the impact of a variety of methodological factors that might have influenced the results: the quality of sampling of students in each participating country, the particular choices of content and format for the test, but mostly, I wondered about the way that the test had been translated from English to the other languages in which the test was used in the study. International studies of educational achievement can be invaluable to policy makers and educators but not if methodological factors undermine the validity of the results. It struck me that possibly the surprising results were due to the fact that the test may have been made unintentionally easier or harder by the translators. What were their qualifications? How much time were they given to do the work? What empirical evidence was compiled to support the equivalence of the test in multiple languages? I called the testing agency responsible for conducting the study to discuss test translation methods. Unfortunately, I was not overly impressed with the details they provided for how the test had actually been translated and how they checked the linguistic, psychological, and statistical equivalence of the test in multiple language and cultural groups.

In my own subsequent checking for good test translation practices I was disappointed by the relatively low level of methodological sophistication that I found compared to the sophistication in the testing field for addressing other important topics such as test development, test score equating, and test score norming. This was my first serious exposure to the world of cross-cultural testing. I could see that there was important work to be done.

In 1991 I brought my concern about test translation methodology to the council of the International Test Commission (ITC). Today, the ITC is an organization of national psychological societies, testing agencies, and individual members, and is committed to improving testing practices around the world. The ITC council decided to form an international committee of scholars and practitioners to develop guidelines for test translation and adaptation, and we were fortunate to secure some financial assistance for the work of the committee from the National Center for Educational Statistics and the College Board in the United States. We were able to interest a number of international organizations in the work of the committee and they provided members. These organizations were the European Association of Psychological Assessment, European Test Publishers Group, International Association for Cross-Cultural Psychology, International Association of Applied Psychology, International Association for the Evaluation of Educational Achievement, International Language Testing Association, and the International Union of Psychological Science.

The committee members worked hard over 3 years and several meetings to organize the technical advances that had been made over the years on the topic of test translation and adaptation, and eventually the committee produced a final report that offered 22 guidelines (called the "International Test Commission Guidelines for Test Adaptation"). The guidelines themselves and the rationale for including each one in the collection is presented in chapter 1.

At about the time the Guidelines in draft form were being circulated around for comment, Tom Oakland from the University of Florida, in the United States, and an ITC council member, and I, decided to organize a conference that would introduce the Guidelines. This conference, sponsored by the ITC, was held at Georgetown University in the United States in the spring of 1999. Attendance at the Conference was high, and highlighted what the ITC knew, which was that a set of guidelines for test translation and adaptation would be well received by the testing field, and would be an important addition to the emerging literature.

At about the same time as the Conference, Professors Charles Spielberger and Peter Merenda came forward (Professor Spielberger had been a member of the committee that developed the Guidelines) and agreed to assist in the preparation of a book that would highlight important technical advances in the test translation and adaptation field. Professor Spielberger, himself, had been involved in more than

50 translations of his own instrument, State-Trait Anxiety Inventory, and Peter Merenda had been active in translations research for most of his career. The three of us teamed up to produce this book, which is a collection of many of the invited addresses from the ITC Conference at Georgetown University and invited chapters that were added to provide comprehensive coverage of the topic.

Chapter 1, written by myself, was prepared to introduce the ITC Guidelines for Test Adaptation. In addition, many of the issues that arise in test translation and adaptation work are described. Chapter 2 was prepared by Professors Fons van de Vijver and Ype Poortinga from the University of Tilburg in the Netherlands on the topic of conceptual and methodological issues in test adaptation. Had it not been for the goal of introducing the Guidelines in the first chapter, this chapter would have been the first one in the book because the authors present a framework for understanding the process of translation and adaptation that is relevant for all of the chapters. Chapter 3 was prepared by Professor Tom Oakland and he tackles the all important question of ethics and test adaptation. At the core of his work is a concern for validity of test scores in cross-cultural contexts.

Chapters 4 to 7 provide a wonderful array of advances in test translation and adaptation methodology. Chapter 4 by Steve Sireci, Liane Patsula (now at the Educational Testing Service in the United States), and myself from the University of Massachusetts in the United States, provides a comprehensive review of approaches for statistically identifying flawed test items that occur during the test translation and adaptation process. Chapter 5 by Professor Sireci was prepared to address the issues, strengths, and weaknesses associated with the uses of bilingual participants in establishing equivalence of different language versions of a test. Chapters 6 and 7 by Dr. Linda Cook, Dr. Alicia Schmitt-Cascallar, and Catharine Brown (chapter 7 only) from the Educational Testing Service provide descriptions of important methodology for statistically comparing tests in multiple languages, and a review of important issues that arise in translating and adapting tests. We regret to announce the untimely passing of Alicia Schmitt-Cascallar in 2003. She was an invited speaker at the ITC Conference at Georgetown University and was an important contributor to the research on testing methodology, including test translation and adaptation methodology.

Chapters 8 to 14 in the book were intended to shift the focus from primarily presentations of issues and methodology to the complicated world of test translation and adaptation applications. The applications of test translation and adaptation methodology include credentialing exams, intelligence tests, cognitive tests, tests used in industrial and organizational settings, admissions tests, and personality tests. Dr. Cyndy Fitzgerald, formerly of Microsoft and now a consultant to Caveon, describes in chapter 8 the process Microsoft uses to translate and adapt their credentialing exams. The use of on-line systems to expedite the work of test translators appears exemplary in the profession. In chapter 9, Dr. Carlos Maldonado (from the Putnam/Northern Westchester BOCES in the United States) and Professor Kurt Geisinger (from The University of St. Thomas, in the United States) describe problems with the English to Spanish translation and adaptation of one of the most popular intelligence instruments in the world: Wechsler Adult Intelligence Scale. In chapter 10, Professor Norbert Tanzer, (from Alliant International University in the United States and the University of Graz in Austria) makes a strong argument for simultaneous development of some psychological tests, rather than translating and adapting tests across languages and cultures. Professor Fritz Drasgow and Tahira Probst from the University of Illinois in the United States describe in chapter 11 their important work in establishing test equivalence across language groups and cultures with tests that are primarily used in industrial/organizational settings. In chapter 12, Drs. Michal Beller (from the Educational Testing Service), and Naomi Gafni and Pnina Hanani (from the National Institute for Testing and Evaluation in Israel) describe their ambitious efforts to produce college admissions tests for use in Israel in six languages. Chapter 13 by Peter Merenda (from the University of Rhode Island in the United States) presents many of his observations and findings in the test translation and adaptation field over his career. Few researchers have worked longer and more successfully in the field. Finally, in chapter 14, Professor Spielberger from the University of South Florida, and two of his colleagues, Manolete Moscoso and Thomas Brunner, from the same university, provide a wealth of information on the issues and methods associated with translating and adapting personality tests.

On behalf of myself and my co-editors, Peter Merenda and Charles Spielberger, we hope that this collection of 14 chapters furthers the mission of the International Test Commission by providing direction and stimulating research on the ever-increasingly important topic of test translation and adaptation. The growth of this field has been tremendous since my first queries in 1989. Today, the field is better developed, guidelines for good practice are in place, methodology has been organized and extended, and there are a growing number of exemplary examples for practitioners to follow. At the same time, there is considerably more research that needs to be done, and so we hope this collection of chapters stimulates others to advance this work.

—Ronald K.Hambleton

I

Cross-Cultural Adaptation of Educational and Psychological Tests: Theoretical and Methodological Issues

1

Issues, Designs, and Technical Guidelines for Adapting Tests Into Multiple Languages and Cultures

Ronald K.Hambleton
University of Massachusetts at Amherst

Considerable evidence exists today to suggest that the need for multilanguage versions of achievement, aptitude, and personality tests, and surveys, is growing (see, e.g., Ercikan, 2002; Hambleton, 2002; Hambleton & de Jong, 2003; Harkness, 1998). For example, the International Association for the Evaluation of Educational Achievement (IEA) conducted the Third International Mathematics and Science Study (TIMSS) in over 45 countries, which involved preparing mathematics and science tests in over 30 languages. Prominent examples of new test adaptation projects in the United States include studies to prepare Spanish versions of College Board's *Scholastic Assessment Test* (SAT), American Council on Education's *General Educational Development* (GED) test, the U.S. Department of Education's *National Assessment of Educational Progress* (NAEP), and achievement tests of several state departments of education. Substantially more test adaptations can be expected in the future as (a) international exchanges of tests become more common, (b) more exams are used to provide international credentials, and (c) interest in cross-cultural research grows.

Although the many reasons for adapting tests from one language and culture to another are clear—for example, facilitating comparative studies of school achievement across cultural and language groups, saving money and time associated with preparing new tests, and achieving fairness in assessment—methods and guidelines for preparing test adaptations and establishing the equivalence of scores are not well known (Hambleton, 1993, 1994; Hui & Triandis, 1985; van de Vijver & Hambleton, 1996). Some cross-cultural researchers have even suggested that a high percentage of the research in their field is flawed to the point of being invalid because of poorly adapted tests.

The purposes of this chapter are (a) to review several sources of error or invalidity associated with adapting tests and to suggest ways to reduce those errors, and (b) to describe a set of practical guidelines for adapting tests prepared by the International Test Commission (ITC) with the assistance of seven other large international agencies (see Hambleton, 1994; van de Vijver & Hambleton, 1996).

Before proceeding, a distinction needs to be made between test adaptation and test translation. The term *test adaptation* is preferred to the more popular and frequently used term *test translation* in this chapter because the former term is broader and more reflective of what should happen in practice when preparing a test that is constructed in one language and culture for use in a second language and culture. Test adaptation includes all the activities from deciding whether or not a test could measure the same construct in a different language and culture, to selecting translators, to deciding on appropriate accommodations to be made in preparing a test for use in a second language, to adapting the test and checking its equivalence in the adapted form. Test translation is only one of the steps in the process of test adaptation and even at this step, adaptation is often a more suitable term than translation to describe the actual process that takes place. This is because translators are trying to find concepts, words, and expressions that are culturally, psychologically, and linguistically equivalent in a second language and culture, and so clearly the task goes well beyond simply preparing a literal translation of the test content.

For our purposes too we use the term “test” throughout the chapter to include all types of educational and psychological instruments, and even surveys and questionnaires.

SOURCES OF ERROR OR INVALIDITY IN TEST ADAPTATION

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) *Standards for Educational and Psychological Testing* (1985) provides careful directions for educational measurement specialists and psychologists who select, develop, administer, and use educational and psychological tests. Three of the standards in this publication are especially relevant in the context of test adaptation:

> *Standard 6.2.* When a test user makes a substantial change in test format, mode of administration, instructions, language, or content, the user should revalidate the use of the test for the changed conditions or have a rationale supporting the claim that additional validation is not necessary or possible.
>
> *Standard 13.4.* 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.
>
> *Standard 13.6.* When it is intended that the two versions of dual-language tests be comparable, evidence of test comparability should be reported.

These standards provide a framework for considering sources of error or invalidity that might arise in efforts to adapt a test from one language and culture to another. For our purposes, sources of error or invalidity that arise in test adaptation can be organized into three broad categories: (a) cultural/language differences, (b) technical issues, designs, and methods, and (c) interpretation of results. Failure to attend to the sources of error in each of these categories can result in an adapted test that is not equivalent in the two language and cultural groups for which it is intended. Nonequivalent tests, when they are assumed to be equivalent, can only lead to errors in interpretation and faulty conclusions about the groups involved.

A good example of the misinterpretation that can follow from poor test adaptation is the following (the example was passed on by Richard Wolf of Columbia Teachers College, a leader during his career in the field of international assessment). In an international comparative study of reading (around 1990), American students were asked to consider pairs of words and identify them as similar or different in meaning: "Sanguine—pessimistic" was one of the pairs of words where American student performance was only slightly above chance (or about 54% of the American students answered the question correctly). In the non-English-speaking country ranked first in performance, about 98% of the students answered the question correctly In the process of attempting to better understand the reason for the huge difference in performance, it was discovered that the word *sanguine* had no equivalent word in the language of this top-performing country and so the equivalent of the English word *optimistic* was used. This substitution made the question considerably easier and would have been answered correctly by a high percentage of the American students as well had they been presented with the pair of words "optimistic—pessimistic." The point of this example is to highlight the danger in drawing conclusions from international comparative studies of achievement without strong evidence that the test adaptation process resulted in two equivalent tests. Prior to 1990 many of the test adaptation initiatives for international studies involved little more than using a couple of good translators. This must be contrasted with the high level of test adaptation sophistication that is seen today in both TIMSS and Organization for Economic Cooperation and Development's Programme for International Student Assessment (OECD/PISA; see, e.g., Grisay, 2003; Hambleton, 2002).

What follows is a discussion of several common errors and how they might be addressed in practice.

Cultural/Language Differences Affecting Scores

The assessment and interpretation of cross-cultural results should not be viewed in the narrow context of just the translation or adaptation of tests (van de Vijver & Leung, 1997, 2000). Rather, this process should be considered for *all* parts of the assessment process, including construct equivalence, test administration, item formats used, and the influence of speed on examinee performance. These four factors are briefly considered next. They have received more attention in subsequent chapters.

Construct Equivalence. Construct equivalence encompasses both conceptual/functional equivalence as well as equivalence in the way the construct measured by the test is operationalized in each language/cultural group (Harkness, 1998). Determining that construct equivalence exists between different cultures under study is a prerequisite for doing any cross-national, cross-cultural, or cross-language comparisons. The use of nonequivalent constructs is one of the most serious errors in cross-language research. For example, it is of limited value to compare two countries in terms of their mathematics achievement when the content measured by the test is highly reflective of the mathematic curriculum in one country but not the other. Another example might be the construct of "quality of life." In one country the construct might include having many material items such as cars, homes, and television sets, whereas in another it could be the construct would include little more than food for survival and a doctor nearby. A comparison of scores from a quality-of-life test produced in one country and adapted for use in the other would have little value.

Determining whether construct equivalence exists between two cultures involves primarily judgmental strategies. A researcher must begin by using his or her common sense to answer such questions as, Is it sensible to compare these two cultures on this construct? Does the construct that is being measured have similar meaning in all cultures being compared? Is the construct operationalized in the same way in all cultures being studied?

To be able to answer yes to these questions and thus ensure conceptual/functional equivalence and equivalence of construct operationalization, several approaches might be taken. This maybe done by interviewing or observing people from the cultures of interest, researching the cultures of interest, and asking others who know about the cultures. These ways are very subjective, and therefore, the use of multiple sources of evidence is highly recommended. Van de Vijver and Poortinga (see chap. 2, this volume) and Sireci, Patsula, and Hambleton (see chap. 4, this volume) have much more to say in their chapters about judging construct equivalence.

Test Administration. Communication problems between a test administrator and examinees can be a serious threat to the validity of test results. Perhaps the test directions are not clearly communicated because of adaptation problems. One way to circumvent problems, but always feasible, is to ensure that the instructions on the test itself are clear and self-explanatory, with minimal reliance on verbal communication (van de Vijver & Poortinga, 1991). Special problems can be expected with directions for rating scales used in attitude measurement too because they are not common in many countries (see Harkness, 1998).

The proper selection of test administrators can be helpful too. They should (a) be drawn from the target communities, (b) be familiar with the culture, language, and dialects, (c) have adequate test administration skills and experience, and (d) know the importance of following any standardized procedures associated with the test. Additionally, consistency in test administration across different groups can be improved by providing (basic) training to all test administrators. Training sessions should be preplanned as part of the test development process, stressing clear, unambiguous communication, the importance of following instructions, strictly following time limits, the influence of test administrators on reliability and validity, and so on.

Test Format. Differential familiarity with particular item formats presents another source of invalidity of test results in cross-cultural studies. In the United States, selected response items such as multiple-choice items have been used extensively in assessment (though that practice has been changing in the last 10 years, and today, we see more use of performance assessments). In cross-cultural studies, it cannot be assumed that everyone is as familiar with multiple-choice items as American students. Nationalities that follow the British system of education, historically at least, have placed much greater emphasis on essays and short-answer questions, compared to multiple-choice items. Thus, students from these countries are placed at a possible disadvantage when compared to their American counterparts. When constructed response formats such as essay questions are emphasized or serve as the dominant mode of assessment, persons with more experience with selected response formats such as multiple-choice items will be placed at a disadvantage. Sometimes a balance of item formats may be the best solution to ensure fairness and reduce sources of invalidity in the assessment process. This strategy has been adopted in recent international studies of achievement (e.g., TIMSS and OECD/PISA).

Another solution to the potential biasing effect associated with a particular item format is to include only those formats with which all groups being assessed are experienced. Whenever it can be demonstrated that examinees are not placed at a disadvantage, and when all variables of interest can still be measured, it would seem that multiple-choice items or simple rating scales should be preferred. The major advantage is that multiple-choice items or simple rating scales can be objectively scored. Thus, complications in scoring associated with open-ended responses are avoided. This is especially relevant in cross-cultural studies where it may be *more difficult* to translate the scoring rubrics than the test items! In addition, extensive, unambiguous instructions including examples and exercises help to reduce differential familiarity (van de Vijver & Poortinga, 1992). At the same time, adopting a single item format for a test runs the danger of having to narrow the intended construct of interest to those parts that can be measured with the single item format, and this too can distort the findings from comparative studies across national boundaries.

Speededness. It is often assumed that examinees will work fast on "speeded" tests (van de Vijver & Poortinga, 1991). But to know to work quickly is a test-taking skill that may not be known or understood by examinees in different cultures. In a study comparing Dutch and other ethnic students in the Netherlands, van Leest and Bleichrodt (1990) found that the speed factor increased score bias. Not all cultural groups have had the same experiences with speeded tests, and those that had not were placed at a serious disadvantage. There are numerous other studies highlighting item and test bias due to the role of test speededness (see, e.g., studies on ethnic bias on the *SATs* in the United States). For example, it is common to find items appearing late in a test to show more bias than items appearing earlier in a test. The bias is against poor readers, and often the problem is due to the role of speed in test performance. The best solution would seem to be to minimize test speededness as a factor in cognitive test performance unless it is a relevant part of the construct being measured. The last point is important because sometimes speed of performance is an integral part of the construct being measured such as it is with the ability to solve analytic reasoning problems. Then, speed is an important part of the construct, so examinees need to understand the need to work quickly.

Technical Issues, Designs, and Methods

There are five technical factors that can influence the validity of tests adapted for use in other languages and cultures: the test itself, selection and training of translators, the process of translation, judgmental designs for adapting tests, and data collection designs and data analysis for establishing equivalence. Each of these factors is considered briefly next. More extensive discussions of these factors appear in subsequent chapters.

The Test Itself. If a researcher knows that he or she will be using a test in a different language or culture, it is advantageous to take this into account at the outset of the test development process. Failure to do so can introduce problems later in the adaptation process that will reduce the validity of the adapted test (Hambleton & Patsula, 1999). Choice of item formats, stimulus material for the test, vocabulary, sentence structure, and other aspects that might be difficult to translate well can all be taken into account in preparing the test specifications. Such preventive actions can minimize later problems. For example, questions about money might be eliminated because currencies are different around the world and equivalent adaptations may be difficult to produce. Also, reading passages about country-specific topics such as "ice hockey" that would be unfamiliar in many cultures could be rejected in favor of passages about walking through a park or other activities that would have meaning across many language and cultural groups. Another problem that arises in adaptation of passages from English to other languages is the presence of the "passive tense" in the text. Whereas this tense is common in English writing, it does not exist in some other languages (e.g., Spanish).

With personality scales, for example, care must be taken to choose situations, vocabulary, and expressions that will adapt easily across language groups and cultures. For example, behaviors that may be common in the Western world may have a very different meaning or not be meaningful at all in some other cultures. A statement such as "I like to start conversations at a party" has little meaning in a culture where parties are unknown, or where women do not go to parties, or where approaching others maybe perceived as inappropriate behavior. This is simply one of many examples that could be offered.

Selection and Training of Translators. The importance of obtaining the services of competent translators is obvious. Too often though, researchers have tried to go through the translation process with a single translator selected because he or she happened to be available—a friend, a wife of a colleague, someone who could be hired cheaply, and so on. Competent translation work cannot be assumed. Also, the use of a single translator, competent or not, does not permit valuable interactions among independent translators to take place to resolve different points that arise in preparing a test adaptation. A single translator brings, for instance, a perspective, a preference for certain words and expressions, which may not be the most suitable for producing a good adaptation of a test. Multiple translators can protect against the dangers of a single translator and his or her preferences and peculiarities.

At the same time, translators should be more than persons familiar and competent with the languages involved in the translation. They should know the cultures very well, especially the target culture (i.e., the culture associated with the language of the adapted test). This knowledge is often essential for an effective adaptation. Also, subject matter knowledge in the adaptation of achievement tests is highly desirable. The nuances and subtleties of a subject area will be lost on a translator unfamiliar with the subject matter. Too often, translators without technical knowledge will resort to literal translations that are often problematic to target-language examinees and threaten test validity. For example, the sentence, “Je ne suis pas une valise,” has an easy literal translation in English (I am not a suitcase) but the actual meaning of the sentence in French is “I am not that stupid.” A literal translation from French to English would totally distort the meaning.

Finally, test translators would benefit from some training in test construction. For example, test translators need to know that when doing adaptations of achievement or aptitude tests they should not create clang associations that might lead test-wise examinees to the correct answers, or translate distractors in multiple-choice items unknowingly so that they have the same meaning. A test translator without knowledge of the principles of test and scale construction could easily make test material more or less difficult unknowingly, and correspondingly, lower the validity of the test in the target population.

Process of Translation. The problem of dialects within a language can become a threat to the validity of adapted tests. Which dialect is of interest, or is the goal to produce an adaptation that would apply across dialects within a language? This problem should be settled before the test adaption begins, and should be used in the selection and training of translators.

Frequency counts of words can be valuable in producing valid test adaptations. In general, it is best to translate words and expressions with words and expressions with approximately the same frequencies in the two languages in an effort to control for the difficulty of words across languages. A problem is that these frequency lists of words and expressions are not always available. This again is a reason for preferring translators who are familiar with both the source and target cultures and not just the languages.

“De-centering” is sometimes used in adapting tests. It maybe that some words and expressions do not have equivalent words and expressions in the target language. It is even possible that the words and expressions do not exist in the target language. De-centering involves making revisions to the source-language test so that equivalent material can be used in both the source- and target-language versions. De-centering is possible when the source-language test is under development at the same time as the target-language version. This is the situation with tests intended for use in international assessments, and some credential tests (e.g., those produced by Microsoft) intended for worldwide use.

Judgmental Designs for Adapting Tests. The two most popular designs are *forward translation* and *backward translation*. With a forward-translation design, a single translator, or preferably, a group of translators adapt the test from the source language to the target language. Then, the equivalence of the two versions of the test is judged by another group of translators. Revisions can be made to the target-language version of the test to correct problems identified by the translators. Sometimes as a final step, yet another person, though not necessarily a translator, will take the target-language version of the test and edit the test to "smooth out" the language. Choppiness can result when translations from different individuals or groups are merged into a single version.

The main advantage of the forward-translation design is that judgments are made directly about the equivalence of the source- and target-language versions of the test. The validity of the judgments about the equivalence of the two versions can be enhanced by having a small group of examinees provide translators with their interpretations of the test or questionnaire directions, content, and formats. This can be done in what are called "think-aloud" studies.

The main weakness of the forward-translation design is associated with the high level of inference that must be made by the translators about the equivalence of the two versions of the test. Other weaknesses include (a) translators may be more proficient in one language than the other, (b) ratings of test equivalence involve judgments by persons who are bilingual, and so they may use insightful guesses based on their knowledge of both languages, (c) translators may be better educated than the monolingual examinees for whom the test is intended and so they miss some problems that would be confronted by the examinees, and (d) (the monolingual) test developers are not in a position to judge test equivalence themselves.

The back-translation design is the best known and most popular of the judgmental designs. In its most popular version, one or more translators adapts a test from the source language to the target language. Different translators take the adapted test (in the target language) and adapt it back to the source language. Then, the original and the back-translated versions of the test are compared and judgments are made about their equivalence. To the extent that the two versions of the test in the source language look similar, support is provided for the equivalence of the source and target versions of the test. The back-translation design can be used to provide a general check both on the quality of the translation and to detect at least some of the problems associated with poor translations or adaptations. Researchers especially like this design because it provides them with an opportunity to judge the original and back-translated versions of the test so that they can form their own opinions about the adaptation process. This is not a possibility for them with the forward-translations design unless they are proficient in the languages.

Although the back-translation design has merit and often can identify problems in an adaptation process, it would rarely provide a sufficient amount of evidence to support the valid use of an adapted test. Evidence of test equivalence provided by a back-translation design is only one of many types of evidence that should be compiled in a test adaptation study. One of the main shortcomings is that the comparison of two or more language versions of a test is carried out only in the source language. It is possible that the test adaptation could be poor although the evidence on the comparability of the original test and the back-translated test would suggest otherwise. This could happen if the translators used a shared set of adaptation rules that ensured that the back-translated test looked like the original test. A second shortcoming is that the adaptation could be poor because it retained inappropriate aspects of the source-language test such as the same grammatical structure and spelling. Such errors would facilitate back-translations but this design would hide serious shortcomings in the target version of the test. For example, the game "ice hockey" may be retained when adapting a test into Spanish and the words then would be easy to back-translate. Unfortunately, the game may have little meaning to many persons who speak only the Spanish language, and so the validity of the Spanish version of the test would be lowered.

Finally, this and other judgmental designs have drawbacks because samples of the intended populations for the tests never actually take the tests under testlike conditions (or, for that matter, any other conditions). There is ample evidence to suggest that reviewers are not able to identify all the flaws in test items and this is why test items are routinely field-tested prior to their use. Adapted tests need to be field-tested too to uncover problems that go unidentified by the translators even when a combination of optimal translation designs and excellent translators are used (see, e.g., Hambleton & Patsula, 1999).

Data Collection Designs and Data Analysis for Establishing Test and Item Equivalence. Three data collection designs are commonly used to evaluate the equivalence of factor structure of the test and of the test items (or rating scales) in different languages. Evaluation of these designs follows (substantially more details about the designs and appropriate statistical methods can be found in subsequent chapters of the book):

1. *Bilingual examinees take source and target versions of the test.* In this design, the same examinees take both the source and target versions of the test. The advantage of this design is that differences in examinee characteristics on the test (e.g., demographic characteristics) can be controlled (see Sireci, chap. 5, this volume; Sireci, 1997). Various item and test statistics can be compiled from the administration of each version of the test and compared to determine equivalence. However, the design is based on the assumption that bilingual examinees are equally proficient in each of the languages. This is highly unlikely to occur for a substantial number of examinees (Cziko, 1987; Rosansky, 1979) and so the assumption should be checked whenever possible. For a bilingual data collection design to be effective, it is often best that it be implemented with another data collection design so that convergent validity of results can be investigated.

A second major problem with this data collection design is that statistical results obtained from data collection may not be generalizable to the intended populations of monolinguals as bilingual examinees tend to be, on the average, different in important ways from their monolingual counterparts (Hambleton, 1993). In one study by Hulin, Drasgow, and Komocar (1982) with the *Job Descriptive Index,* these researchers learned that only 4% of the items in their attitude scale were identified as poorly translated with a bilingual sample of examinees. Over 30% of the items were identified as poorly translated when monolingual samples of examinees from the source- and target-language populations were used.

A variation of this bilingual design, which has the same limitations but is easier to implement, involves randomly assigning bilingual examinees to take one of the language versions of the test. In this case, a randomly equivalent populations design is in effect.

2. *Source-language monolinguals take the original and back-translated versions.* This design involves the administration of the original and back-adapted versions of the test to a sample of monolingual examinees in the source language. Item equivalence is identified by comparing

participant performance on the original and back-translated version of each item. Factor analysis might be applied to the data collected from each version of the test, and factor structures compared. The advantage of this design is that by using one sample of participants, the resulting scores are not confounded by differences in examinee characteristics (Hambleton & Bollwark, 1991).

Two major shortcomings, however, weaken the usefulness of this data collection design. First, no empirical data are collected from the target-language version of the test. That is, no target-language monolinguals are used, although the aim of the research is to apply the findings to the target-language version of the test and the target-language monolinguals. Second, the results that are obtained may not be independent because it cannot be ruled out that learning results from administering the first original-source language version of the test and that the learning affects examinee performance on the back-translated version of the test. Counterbalancing can reduce the significance of practice effects but it does complicate the analyses.

3. *Source-language monolinguals take source language and target-language monolinguals take target language.* A more suitable data collection design would involve monolinguals taking the source-language version of the test and a second sample of monolinguals taking the target-language version of the test. An assumption of equal ability distributions across the two groups is not usually tenable and, fortunately, such an assumption does not need to be made if the analyses are carried out within an item response theory (IRT) framework (Ellis, 1989, 1991; Ellis & Kimmel, 1992; Hambleton, Swaminathan, & Rogers, 1991; van de Vijver & Leung, 1997, 2000) and/or item equivalence studies are carried out using conditioning procedures (Holland & Wainer, 1993). The advantage of this design is that samples of the source and target populations are used in the analyses and therefore findings about the equivalence of the two language versions of the test are generalizable to the populations of interest.

One of the major investigations for establishing item equivalence proceeds like item bias studies (Hambleton et al., 1991; Sireci & Allalouf, 2003). Comparisons of the item statistics in the two language versions of the test (or more, if available) are made controlling for any ability differences in the two groups (see Hambleton & Kanjee, 1995a). Items showing differences are identified and carefully studied to determine possible explanations for the differences (see, e.g., Ercikan, 2002). A poor adaptation is one explanation. Unfortunately, these studies are unable to disconfound cultural differences and adaptation problems but they are often revealing, generally, of potential problems with the adapted version of the test. Item bias analyses come from both classical and modern test theory and can be applied to both binary and polytomous response data (Sireci & Allalouf, 2003).

Factors Affecting Interpretation of Results

In large-scale cross-cultural studies, the purpose of the test is to provide a basis for making comparisons between various cultural/language groups, so as to understand the differences and similarities that exist (Hambleton, 1993, 2002). Sometimes cognitive variables are of interest and other times the focus maybe on the assessment of personality variables or general information (e.g., quality of life, health). It is hoped that results will be used for seeking ways of comparing groups and understanding the differences. Cross-cultural studies should not be used to support arguments about the superiority or exceptionality of nations as if the international comparative study is the equivalent of a horse race with winners and losers (Westbury, 1992). At best these studies provide only a "snapshot" of differences that exist, and provide only a limited basis for interpreting the results. In this context, to gain a better understanding when interpreting scores, other relevant factors external to the tests or assessment measures and specific to a nationality should be considered. Curricula, educational policies and standards, wealth, standard of living, cultural values, and so on, may all be essential factors for properly interpreting scores across cultural/language and/or national groups. A sampling of the factors that should normally be considered in interpreting test results across language and cultural groups is presented next.

Similarity of Curricula. To the extent that differences in curricula exist, achievement comparisons between different cultures will be tenuous if these curricula differences are not taken into account. Westbury (1992) noted that the results of the *Second International Mathematics Study* (SIMS) indicated that American students performed poorly in every grade and in every aspect of mathematics that was covered on the test. When comparing performance of Japanese and American students, major curricular differences between the two countries were noted. However, in areas of the curricula of the two countries that were similar, Westbury found no essential differences between student performance in the two countries. Analyses of curricula differences are obviously important in these international comparative studies of achievement, and this is why, despite some opposition (because of extra burden and cost), extensive questionnaire data are compiled along with the test data in each participating country.

Student Motivation. Wainer (1993) questioned whether demonstrated proficiency as measured by tests can be separated from motivation. He noted that in the *International Assessment of Educational Progress Study* (Lapointe, Mead, & Askew, 1992), all the (randomly) selected students from one participating country were made aware of the great honor of being chosen to represent their school and country, and thus had a responsibility to perform at their best. For students in some other countries, on the other hand, participation on this international comparative study was just another activity and not especially important to students because individual scores were not made available. For these students, the tests were "low-stakes." To interpret performance differences between countries with motivated students and those countries without motivated students without considering differential motivation to perform on the test could lead to a major misinterpretation of the findings.

Also, van de Vijver and Poortinga (1991) noted that it cannot be assumed that examinees will always try to achieve a high score. For example, it has been reported that for many Black South African students, the aim in tests was to achieve the minimum score needed to pass. This is because the imposed state education system at the time was perceived by many examinees to be detrimental to Blacks, and thus, students aspired only to the minimum required of them. In this context, it would not be unusual to expect levels of performance that may have little to do with true ability

Sociopolitical Factors. The meaning and interpretation of test scores can also differ even when the scores are the same. Consider comparing test scores between students from developed and developing nations, or industrialized and mainly rural societies. In this context, performance of students may not be related to ability at all. Rather, performance may be a reflection of the lack of access to adequate resources, or the different quality of educational services available.

The point is that, for any meaningful interpretations of the results, the different social, political, and economic realities facing nationalities, as well as the relevance of educational opportunities in the light of these realities, must be considered (Olmeda, 1981). Thus, it is important for test developers and policymakers to be aware of those specific cultural issues that might impact on test performance.

PRACTICAL GUIDELINES FOR ADAPTING TESTS

The technical literature for guiding the test adaptation process is definitely incomplete (from a measurement perspective), and scattered through a plethora of international journals, reports, and books. There has been no single complete source that practitioners could turn to for advice, nor was a set of guidelines for adapting tests ever formalized until recently (Hambleton, 1994; van de Vijver & Hambleton, 1996). Also, until recently, the more complex measurement methods (e.g., item response models and structural equation models), which are very useful in formally establishing the equivalence of scores obtained from tests adapted for use in multiple languages and cultures, have not been well known to researchers who do test adaptations (e.g., Hulin, 1987). But, as is clear from the chapters in this book (see also Hambleton & de Jong, 2003), the situation has improved substantially since the early 1990s. In fact, two of the purposes of the ITC conference held at Georgetown University in the United States in 1999, were to bring researchers from around the world together to share their knowledge and experience about test adaptation, and to unveil the final version of the ITC Guidelines for Test Adaptation. The purposes of this section of the chapter are to describe the motivation for the ITC to prepare the Guidelines, to provide some of the background for preparing the Guidelines, and then to describe the 22 Guidelines and the rationale for including each of them.

In view of the fact that "high-stakes" are often associated with the results from cross-cultural or international comparative studies of educational achievement (see, e.g., the high level of interest there is today in financially supporting international comparative studies of achievement), the need for professionally developed and validated practical guidelines for adapting tests and establishing score equivalence seemed clear to the ITC as early as 1992. Technical standards or guidelines for assessment practices concerning test development, reliability assessment, validity assessment, and reporting were available in many countries (see, e.g., AERA, APA, & NCME, 1985, 1999), but rarely had much attention been given to the preparation of guidelines for adapting tests and establishing score equivalence. For example, in the widely used AERA, APA, and NCME Test Standards published in 1985 (which were the most influential test standards in the United States until the 1999 Test Standards were published), only three standards directly address the topic of test adaptations. And in Canada, a bilingual country, only three standards that addressed test adaptation appeared in the Canadian Psychological Association's test standards (which were available in 1993).

The ITC addressed this shortcoming by preparing a set of practical guidelines for adapting tests (see Hambleton, 1994; van de Vijver & Hambleton, 1996), referred to as the ITC Guidelines for Test Adaptation. Table 1.1 identifies the eight organizations who came together to develop the Guidelines. Table 1.2 identifies the committee members, who worked for 3 years to produce them. The ITC Guidelines for Test Adaptation are organized into four sections: context, test development and adaptation, administration, and documentation/score interpretations. The thinking of the ITC committee who produced the Guidelines was that the Guidelines would be more convenient to use if they were organized into meaningful categories. Guidelines in the context category address concerns about construct equivalence in the language groups of interest. The test development and adaptation category includes guidelines that arise in the process of adapting a test, everything from choosing translators to statistical methods for analyzing empirical data to investigating score equivalence. The third category, administration, addresses guidelines having to do with the ways that tests are administered in multiple language groups, and this includes everything from selecting administrators, to the choice of item formats, to establishing time limits. The fourth category of guidelines concerns documentation and score interpretations. Typically, researchers have

TABLE 1.1
Participating Organizations in the Development of the International Test Commission Guidelines for Test Adaptation

International Test Commission (ITC)
European Association of Psychological Assessment (EAPA)
European Test Publishers Group (ETPG)
International Association for Cross-Cultural Psychology (IACCP)
International Association of Applied Psychology (IAAP)
International Association for the Evaluation of Educational Achievement (IEA)
International Language Testing Association (ILTA)
International Union of Psychological Science (IUPsyS)

TABLE 1.2
Committee Members and the Organizations They Represented

Chairperson

Ronald K. Hambleton (ITC)
University of Massachusetts at Amherst, USA

Committee Members

Glen Budgell (ITC)
Canadian Nurses Association, Canada

Rob Feltham (ETPG)
NFER-Nelson, England

Rocio Fernandez-Ballesteros (EAPA)
Universidad de Autonoma, Spain

John H. A. L. de Jong (ILTA)
Cito, The Netherlands

Ingrid Munck (IEA)
Statistics Sweden, Sweden

José Muñiz (ITC)
Universidad de Oviedo, Spain

Ype Poortinga (IACCP)
Tilburg University, The Netherlands

Isik Savasir (IUPsyS)
Hacettepe University, Turkey

Charles Spielberger (IAAP)
University of South Florida, USA

Fons van de Vijver (ITC)
Tilburg University, The Netherlands

Jac N. Zaal (ITC)
GITP International, The Netherlands

Research Associate

Anil Kanjee (ITC)
University of Massachusetts at Amherst, USA

provided very little documentation of the adaptation process to establish the validity of an adapted test, and misinterpretations of scores from tests in multiple languages have been common. The ITC Guidelines for Test Adaptation addressed concerns in this area.

The following was adopted by the ITC committee as a definition of a guideline for test adaptation: “A test adaptation guideline is a practice that is judged as important for conducting and evaluating the adaptation or parallel development of psychological and educational tests for use in different populations.” The 22 Guidelines advanced by the ITC committee are summarized in the following discussion and in Table 1.3 (and were published in draft form earlier in Hambleton, 1994, and van de Vijver & Hambleton, 1996). They appear in this chapter with only minor modifications. In the committee’s final report (ITC, 2001), each guideline was described by (a) a rationale for including the guideline, (b) steps for addressing the guideline in practice, (c) a list of common errors, and (d) a set of references. A complete example of one of the guidelines is provided in Table 1.4. What follows is a brief description of each guideline and the rationale for including the guideline on the list.

Context

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

Rationale/Explanation. There are many factors affecting crosscultural/language comparisons that need to be considered whenever two or more groups from different language/cultural backgrounds are compared, especially when a test is being developed or adapted, or scores are being interpreted. However, often it is necessary that some of these factors are not merely taken into account, but that practical steps be taken to either minimize or eliminate the unwanted effects of these factors on any cross-cultural/ language comparisons that are made. For example, the different levels of test motivation of participants in a recent International Assessment of Educational Progress study is one of the likely reasons for the very different performances of participants from these countries (Wainer, 1993).

2. C.2 The amount of overlap in the construct measured by the test in the populations of interest should be assessed.

Rationale/Explanation. Differences that exist between various cultural and language groups depend not only on different traditions, norms, and values, but also on different worldviews and interpretations. Thus,

TABLE 1.3
ITC Guidelines for Test Adaptation

Context

C.1 (1) Effects of cultural differences that are not important to the main purposes of the study should be minimized to the extent possible.

C.2 (2) The amount of overlap in the construct measured by the test in the populations of interest should be assessed.

Test Development and Adaptation

D.1 (3) Test developers/publishers should ensure that the adaptation process takes full account of linguistic and cultural differences in the intended populations.

D.2 (4) Test developers/publishers should provide evidence that the language used in the test directions, scoring rubrics, and the items themselves are appropriate for all cultural and language populations for whom the test is intended.

D.3 (5) Test developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and other procedures are familiar to all intended populations.

D.4 (6) Test developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.

D.5 (7) Test developers/publishers should compile 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 (8) Test developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish construct and item equivalence among the language versions of the test.

D.7 (9) Test developers/publishers should apply appropriate statistical techniques to (a) establish the equivalence of the language versions of the test, and (b) identify problematic components or aspects of the test that may be inadequate in one or more of the intended populations.

D.8 (10) Test developers/publishers should provide information on the validity of the adapted versions of the test in the intended populations.

D.9 (11) Test developers/publishers should provide statistical evidence about the equivalence of items in all intended populations.

D.10 (12) Non-equivalent items across the intended populations should not be used in "linking" adapted versions of the test to a common score reporting scale. However, these same items may be useful for reporting scores in each population, separately.

Administration

A.1 (13) Those aspects of the environment that influence the administration of a test should be made as similar as possible across populations for whom the test is intended.

A.2 (14) 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.3 (15) 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.4 (16) Test administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.

A.5 (17) The test manual should specify all aspects of the test and its administration that require scrutiny in the application of the test in a new cultural context.

A.6 (18) The administrator should be unobtrusive and the administrator-examinee interaction should be minimized. Explicit rules that are described in the test administration manual should be followed.

Documentation/Score Interpretations

I.1 (19) When a test is adapted for use in another population, documentation of the changes should be provided, along with evidence to support the equivalence of the adapted version of the test.

I.2 (20) Score differences among samples of populations administered the test should not be taken at face value. The researcher has the responsibility to substantiate the meaningfulness of the differences with other empirical evidence.

I.3 (21) 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 (22) 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.

TABLE 1.4
An Example of Guideline D.1 in Its Complete Form

Guideline D.1: General and Professional Requirements

Test developers/publishers should ensure that the adaptation process takes full account of linguistic and cultural differences in the intended populations.

Rationale/Explanation

The expertise and experience of translators are perhaps the most crucial aspects of the entire process of adapting tests as they can significantly affect the reliability and validity of the test (Bracken & Barona, 1991). For example, translators without domain specific or technical knowledge often resort to literal translations that may cause misunderstanding in the target population and threaten the validity of the test (Hambleton & Kanjee, 1995b). Consequently, the selection of appropriately qualified translators is an important aspect of the test adaptation process. Though expertise in both languages is a basic requirement, familiarity and experience with (a) both cultures, (b) the contents of the test, and (c) the principles of developing tests, especially item writing, should also be included as part of the essential requirements for the selection and/or training of translators. Because a single translator cannot be expected to have all of the required qualities and brings a single perspective to the task of translation, in general, it seems clear that a team of specialists is needed to accomplish an accurate adaptation.

Steps to Meet the Guideline

1. As a basic minimum, ensure that translators are qualified and experienced in the source and target languages as well as in both cultures (Butcher & Garcia, 1978). Certification and/or prior experience is an important requirement. For instance, it cannot be assumed that bilinguals have equal command of both languages in all relevant domains or are equally familiar with both cultures.

2. Knowledge of the subject matter is an important requirement for any translator involved in adapting a test. Without at least some content knowledge, the subtleties and nuances of the subject matter can be lost. Prior familiarization with the subject matter for translators lacking domain-specific knowledge should be included as part of the test adaptation process.

Where is a bird with webbed feet most likely to live?

a. in the mountains
b. in the woods
c. in the sea
d. in the desert

When this question was translated from English into Swedish, "webbed feet" became "swimming feet," that then provided an obvious clue to Swedish children about the location of the correct answer. A translator with some knowledge of the principles of item writing would have noticed the flaw in the translation of the item stem and revised the translation.

4. A test adaptation project is best carried out by a team of specialists (see, for example, Grisay, 2003). Translators should participate in such a project team and be involved in the decision making process, and their opinions and views should be actively sought and acknowledged. According to Brislin (1986), this approach can greatly improve the quality of an adaptation. The teamwork approach can help to (1) enable the use of the back-translation methods (see step 5, below); (2) allow translators to compare and discuss their work and thus improve on the relevance and quality of translations; and (3) can help to ensure that specialist knowledge in all required fields is accessible.

5. One possible design is to use a team of translators working independently or in small groups to adapt the test. Later, independent evaluations of the test can be compared, and differences resolved to produce a single best translation. Another procedure is the use of monolingual test developers and translators simultaneously, where tests are first translated/adapted by a translator, edited by a monolingual test developer in the target language and then re-assessed by a bilingual (Brislin, 1986). Brislin (1986) noted that the advantage of this design is that monolingual test developers can rewrite tests so that they would be clear and technically acceptable for target language examinees, and this design minimizes situations where the target version is poor, but this problem might be missed because a highly skilled translator produced an excellent back-translated version of the flawed target version. In the case where only a single translator is available, the use of a member from the target language population to assist the translator is strongly recommended. In this situation, the translator can at least discuss the target language version with someone from the target language group who can indicate problem areas and may suggest revisions too.

Common Errors

1. Selection of translators or easily available individuals familiar to the test developer (i.e., friends or neighbors), simply because they are bilingual has been shown to be an unsuccessful practice (Brislin, 1986).

2. Failure to ensure that translators selected are familiar with the content area as well as experienced in test development. This problem has sometimes been reported by countries participating in TIMSS.

3. Translators are not given sufficient time to do their work. Again, this problem has sometimes been reported by countries participating in TIMSS.

References for Additional Study

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

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds), Field methods in cross-cultural psychology (pp. 137–164). Newbury Park, CA: Sage.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. Language Testing, 20(2), 225–240.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. The Personnel and Guidance Journal, 56(8), 472–475.

Hambleton, R. K., & Kanjee, A. (1995b). Translation of tests and attitude scales. In T. Husen & T. N. Postlewaite (Eds), International Encyclopedia of Education (2nd ed., pp. 6328-6334). Oxford, England: Pergamon.

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

it is possible for the same construct to be interpreted and understood in completely different ways by two cultures. For example, the concept of "intelligence" is known to exist in almost all cultures. However, in many Western cultures this concept is associated with producing answers quickly, whereas for many Eastern cultures, intelligence is often associated with thoughtfulness, reflection, and saying the right thing (Lonner, 1990). Cross-cultural researchers have to ensure that the construct measured by a test in the original source cultural/language group can be found in the same form and frequency in the other cultures that are being studied.

Test Development and Adaptation

1. D.1 Test developers/publishers should ensure that the adaptation process takes full account of linguistic and cultural differences in the intended populations.

Rationale/Explanation. The rationale for this guideline along with the other parts of this guideline description appear in Table 1.4. This one is used as an example of the information that is available for each guideline in the final report (see ITC, 2001).

2. D.2 Test developers/publishers should provide evidence that the language used in the test directions, scoring rubrics, and the items themselves are appropriate for all cultural and language populations for whom the test is intended.

Rationale/Explanation. One of the causes of poor test adaptation for cross-cultural research is that the source-language version of the test is often flawed, and therefore difficult to adapt. Another cause may be that concepts, expressions, and ideas used in the source-language version of the test do not have equivalents in the target language. One of many reasons for the success of recent TIMSS and OECD/PISA studies is the substantial effort that has gone into the source-language test development with clearly defined constructs and test specifications, careful item development and field-testing, and other activities associated with proper test development.

Also it is important to ensure that the vocabulary used for a test in two or more languages is comparable in terms of the level of difficulty of words, readability, grammar usage, writing style, and punctuation. In this context, the reasons for using the test, for example, assessment of adult literacy, and the reading level of participants (children vs. adults) should be carefully considered.

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

Rationale/Explanation. Specific formats (e.g., multiple choice, essay, 5-point rating scales) and certain conventions and procedures in giving instructions and presenting test items may not be equally familiar to all populations. Conventions and procedures range from language use in test rubrics, layout and use of graphics, and presentation mode (e.g., paper and pencil, computer). To ensure fairness it is important that all formats, conventions, and procedures be familiar to all populations for whom adaptations of the test are intended and this may involve the development of extensive practice materials to reduce bias due to unfamiliarity of some aspects of the assessment process.

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

Rationale/Explanation. Any adapted test that proves easier or more difficult to read or understand because of the specific content will introduce an additional source of bias. In some parts of the world, different units are used to express quantity in, for example, weight, length, and money. An adaptation of a test can be more difficult for the target population if the units used are less familiar or if they require different mathematical operations (see, Hambleton, Yu, & Slater, 1999). Also, certain stimulus material (diagrams, tables, figures, famous landmarks) may not be equally familiar to all populations.

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

Rationale/Explanation. The equivalence of meaning in questions/tasks/rating scales in different languages and cultures must be assessed. Judgmental methods of establishing translation equivalence are based on decisions by translators or groups of translators. The two most popular designs, forward translations and backward translations, were considered earlier in the chapter. But both designs have flaws, and so rarely would judgmental designs provide sufficient evidence to validate an adapted test.

6. D.6 Test developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish construct and item equivalence among the language versions of the test.