INTERNATIONAL TEST COMMISSION

ITC Guidelines for Translating and Adapting Tests (Second Edition)

Version 2.4

Please reference this document as:
International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition).* [www.InTestCom.org]

ACKNOWLEDGMENT OF APPRECIATION

The Council of the International Test Commission wishes to thank the six-person committee who worked for several years to produce the second edition of the Guidelines for Translating and Adapting Tests: David Bartram, SHL, UK; Giray Berberoglu, Middle East Technical University, Turkey; Jacques Grégoire, Catholic University of Louvain, Belgium; Ronald Hambleton, Committee Chairperson, University of Massachusetts Amherst, USA; Jose Muniz, University of Oviedo, Spain; and Fons van de Vijver, University of Tilburg, Netherlands.

Also, the International Test Commission wishes to thank Chad Buckendahl (USA); Anne Herrmann and her colleagues at OPP Ltd. (UK); and April Zenisky at the University of Massachusetts (USA) for their careful review of an earlier draft of the document. The ITC is grateful too to all of the other reviewers from around the world who directly or indirectly have contributed to the second edition of the ITC Guidelines for Translating and Adapting Tests.

SUMMARY

The second edition of the ITC Guidelines for Translating and Adapting Tests was prepared between 2005 and 2015 to improve upon the first edition, and to respond to advances in testing technology and practices. The 18 guidelines are organized into six categories to facilitate their use: Pre-condition (3), test development (5), confirmation (4), administration (2), scoring and interpretation (2), and documentation (2). For each guideline, an explanation is provided along with suggestions for practice. A checklist is provided to improve the implementation of the guidelines.

CONTENTS

BACKGROUND

The field of test translation and adaptation methodology has advanced rapidly in the past 25 years or so with the publication of several books and many new research studies and examples of outstanding test adaptation work (see, for example, van de Vijver & Leung, 1997, 2000; Hambleton, Merenda, & Spielberger, 2005; Grégoire & Hambleton, 2009; Rios & Sireci, 2014). These advances have been necessary because of the growing interest in (1) cross-cultural psychology, (2) large-scale international comparative studies of educational achievement (for example, TIMSS and OECD/PISA), (3) credentialing exams being used world-wide (for example, in the information technology field by companies such as Microsoft and Cisco), and (4) fairness in testing considerations by permitting candidates to choose the language in which assessments are administered to them (for example, university admissions in Israel with candidates being able to take many of their tests in one of six languages).

Technical advances have been made in the areas of qualitative and quantitative approaches for the assessment of construct, method, and item bias in adapted tests and questionnaires, including the uses of complex statistical procedures such as item response theory, structural equation modelling, and generalizability theory (see Hambleton et al., 2005; Byrne, 2008). New translation designs have been advanced by OECD/PISA (see, Grisay, 2003); steps have been offered for completing test adaptation projects (see, for example, Hambleton & Patsula, 1999; exemplary projects are available to guide test adaptation practices - e.g. OECD/PISA and TIMSS projects); and many more advances have been made.

The first edition of the Guidelines (see van de Vijver & Hambleton, 1996; Hambleton, 2005) started from a comparative perspective, which is the purpose of the test adaptation to permit or facilitate comparisons across groups of respondents. The implicit template for which the guidelines were intended used a successive instrument development in a comparative context (the existing instrument has to be adapted for use in a new cultural context). It is becoming increasingly clear, however, that test adaptations have a wider domain of applications. The most important example is the use of a new or existing instrument in a multicultural group, such as clients in counselling who come from different ethnic groups, educational assessment in ethnically diverse groups with a differential mastery of the testing language, and internationally oriented recruitment for management functions in multinational companies. This change in domain of applicability has implications for development, administration, validation, and documentation. For example, possible consequences could be that items of an existing test should be adapted in order to increase its comprehensibility for non-native speakers (e.g., by simplifying the language). Another important extension of the guidelines would be to accommodate simultaneous development (i.e., the combined development of source and target language questionnaires). Large-scale international projects increasingly use simultaneous development in order to avoid the problem that the version developed in one language cannot be translated/adapted to all the languages of the study.

The first edition of the ITC Guidelines for Translating and Adapting Tests was published by van de Vijver and Hambleton (1996), and by Hambleton (2002), and Hambleton, Merenda and Spielberger (2005). Only minor editorial changes were seen in the publication of the guidelines between 1996 and 2005. In the meantime, many advances have taken place since 1996. First, there have been a number of useful reviews of the ITC Guidelines. These include papers by Jeanrie and Bertrand (1999), Tanzer and Sim (1999), and Hambleton (2002). All the authors highlighted the value of the guidelines but then they offered a series of suggestions for improving them. Hambleton, Merenda, and Spielberger (2005) published the main proceedings of an ITC international conference held in 1999 at Georgetown University in the USA. Several of the chapter authors advanced new paradigms for test adaptations and offered new methodology including Cook and Schmitt-Cascallar (2005), and Sireci (2005). In 2006, the ITC held an international conference in Brussels, Belgium, to focus on the ITC Guidelines for Translating and Adapting Tests. More than 400 persons from over 40 countries focused on the topic of test adaptation and many new methodological ideas were advanced, new guidelines were suggested, and examples of successful implementations were shared. Papers presented in symposia at international meetings from 1996 to 2009 were plentiful (see, for example, Grégoire & Hambleton, 2009) and see Muniz, Elosua, and Hambleton (2013) for an early version of the second edition of the ITC Guidelines in Spanish.

In 2007, the ITC Council formed a six-person committee and assigned them the task of updating the ITC Guidelines to emphasize the new knowledge that was being advanced and the many experiences that were being gained by researchers in the field. These advances include (1) the development of structural equation modelling for identifying factorial equivalence of a test across language groups, (2) expanded approaches for identifying differential item functioning with polytomous response rating scales across language groups, and (3) new adaptation designs being pioneered by international assessment projects such as OECD/PISA and TIMSS. The committee, too, provided presentations and drafts of the new guidelines at international meetings of psychologists in Prague (in 2008) and Oslo (in 2009) and received substantial feedback on them.

The Administration Guidelines section was retained in the second edition, but overlapping guidelines were combined and the total number was reduced from six to two. "Documentation / score interpretations" was the final section in the first edition. In the second edition, we split this into two separate sections - one focused on score scales and interpretations, and the other focused on documentation. In addition, two of the four original guidelines in this section were substantially revised.

As in the first edition, we want readers to be clear on our distinction between test translation and test adaptation. Test translation is probably the more common term, but adaptation is the broader term and refers to moving a test from one language and culture to another. Test adaptation refers to all of the activities including: deciding whether or not a test in a second

language and culture could measure the same construct in the first language; selecting translators; choosing a design for evaluating the work of test translators (e.g., forward and backward translations); choosing any necessary accommodations; modifying the test format; conducting the translation; checking the equivalence of the test in the second language and culture and conducting other necessary validity studies. Test translation, on the other hand, has a more limited meaning restricted to the actual choosing of language to move the test from one language and culture to another to preserve the linguistic meaning. Test translation is only a part of the adaptation process, but can be, on its own, a very simplistic approach to transporting a test from one language to another with no regard for educational or psychological equivalence.

THE GUIDELINES

Introduction

A guideline is defined in our work as a practice that is important for conducting and evaluating the adaptation (also sometimes called "localisation") or simultaneous development of psychological and educational tests for use in different populations. In the text that follows, 18 guidelines are organized around six broad topics: Pre-Condition (3), Test Development (5), Confirmation [Empirical Analyses] (4), Administration (2), Score Scales and Interpretation (2), and Documentation (2).

The first section named "Pre-Condition" highlights the fact that decisions have to be made before the translation/adaptation process ever begins. The second section "Test Development Guidelines" is focused on the actual process of adapting a test. The third section "Confirmation" includes those guidelines associated with the compilation of empirical evidence to address the equivalence, reliability and validity of a test in multiple languages and cultures. The final three sections are related to "Administration", "Score Scales and Interpretation", and "Documentation". Documentation has been a particularly neglected topic in test adaptation initiatives in psychology and education and we would like to see journal editors and funding agencies demand more when it comes to documentation of the test adaptation process.

For each guideline, we have offered an explanation and suggestions for implementing it in practice.

Pre-Condition Guidelines

PC-1 (1) Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation.

Explanation. Intellectual property rights refer to a set of rights people have over their own creations, inventions, or products. These protect the interest of creators by giving them moral and economic rights over their own creations. According to the World Intellectual Property Organization (www.wipo.int), *"intellectual property relates to items of information or knowledge, which can be incorporated in tangible objects at the same time in an unlimited number of copies at different locations anywhere in the world."*

There are two branches of intellectual property: Industrial property and copyright. The first one refers to patents protecting inventions, industrial designs, trademarks and commercial names. Copyright refers to artistic and technology-based creations. The creator (the author) has specific rights over his/her creation (e.g., prevention of some distortions when it is copied or adapted). Other rights (e.g., making copies) can be exercised by other persons (e.g., a publisher) who have obtained a license from the author or copyright holder. For many tests, as with other written works, copyright is assigned by the author to the publisher or distributor.

As educational and psychological tests are clearly creations of the human mind, they are covered by intellectual property rights. Most of the time the copyright does not refer to specific content of items (e.g., no one has rights on items such as *"1+1 = ..."* or *"I feel sad"),* but to the original organization of the test (structure of the scales, scoring system, organization of the material, etc.). Consequently, mimicking an existing test, i.e., keeping the structure of the original test and its scoring system but creating new items, is a breach of the original intellectual property rights. When authorized to carry out an adaptation, the test developer should respect the original characteristics of the test (structure, material, format, scoring . . .), unless an agreement from the holder of the intellectual property allows modifications of these characteristics.

Suggestions for practice. Test developers should respect any copyright law and agreements that exist for the original test. They should have a signed agreement from the intellectual property owner (i.e., the author or the publisher) before starting a test adaptation. The agreement should specify the modifications in the adapted test that will be acceptable regarding the characteristics of the original test and should make clear who would own the intellectual property rights in the adapted version.

PC-2 (2) Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use (or uses) of the scores.

Explanation. This guideline requires that what is assessed should be understood in the same way across language and cultural groups, and this is the foundation of valid cross-cultural comparisons. At this stage in the process, the test or instrument has not even been adapted so compilation of previous empirical evidence with similar tests, and judgements of construct-item match and suitability for the language groups involved in the study would be desirable. Ultimately, however, this important guideline must be assessed with empirical data along the lines of evidence required in C-2 (10). The goal of any analyses is not to establish the structure of a test, though that is a by-product of any analyses, but to confirm the equivalence of the structure of the test across multiple language versions.

Suggestions for Practice. Individuals who are experts with respect to the construct measured, and who are familiar with the cultural groups being tested, should be recruited to evaluate the legitimacy of the construct measured in each of the cultural/linguistic groups. They can try and answer the following question: Does the construct make sense in the cultures of both groups? We have seen many times in educational testing, for example, that a committee has judged the construct measured by a test to lack meaning or have diminished meaning in a second culture (for example, quality of life, depression or intelligence). Methods such as focus groups, interviews and surveys can be used to obtain structured information about the degree of construct overlap.

PC-3 (3) Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest.

Explanation. The cultural and linguistic characteristics irrelevant to the variables that the test is intended to measure should be identified at the early stage of the project. They can be related to the item format, material (e.g. use of computer, pictures or ideograms…), time limits, etc.

An approach to the problem has been to assess the 'linguistic and cultural distance' between the source and target language and cultural groups. Assessment of linguistic and cultural distance might include considerations of differences in language, family structure, religion, lifestyle, and values (van de Vijver & Leung, 1997).

This guideline relies mainly on qualitative methods and specialists familiar with the research on specific cultural and language differences. It places special pressure on the selection of test translators and requires that translators be native to the target language and culture, since knowing the target language only is not sufficient for identifying possible sources of method bias. For example, in the Chinese-American comparative study of eighth-grade mathematics achievement carried out by Hambleton, Yu, and Slater (1999), format and test length problems were identified, along with a host of cultural features associated with the eighth-grade mathematics test.

Suggestions for practice. This is a difficult guideline to address with empirical data at any time. It is especially difficult at the early stages of test adaptation. At the same time, qualitative evidence can often be collected:

- Either by observation, interview, focus group, or survey, determine motivational levels of participants, their understanding of the instructions, their experience with psychological tests, the speediness associated with test administration, familiarity with the rating scales, and cultural differences (but even these comparisons could be problematic because of cultural differences in understanding the variables themselves). When collecting such research data from participants is problematic, obtain as much information as possible from the translators. Some of this work could be done prior to any progress with the test adaptation.

- It may be possible to control for these 'nuisance variables' in any subsequent empirical analysis once the test has been adapted and is ready for validation studies via the use of analysis of covariance or other analyses, that match participants across language/cultural groups on variables such as motivational level or familiarity with a particular rating scale (e.g., Johnson, 2003; Javaras & Ripley, 2007).

Test Development Guidelines

TD-1 (4) Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise.

Explanation. This has been, over the years, one of the most impactful guidelines because there is considerable evidence suggesting that it has been influential in getting testing agencies to look for translators with qualifications beyond knowledge of the two languages involved in the test adaptation (see, for example, Grisay, 2003). Knowledge of the cultures, and at least general knowledge of the subject matter and test construction, has become part of the selection criteria for translators. Also, this guideline appears to have been influential in encouraging agencies translating and adapting tests to use at least two translators in various designs (e.g., forward and backward translation designs). The old practice of relying on a single translator for all decisions, however well qualified that person might be, has been eliminated from the list of acceptable practices today.

Knowledge/expertise in the target culture results from using translators who are native in the target language and are living in the target locale, with the former being essential and the latter highly desirable. A native speaker of the target language will not only produce an accurate translation, but also one that reads fluently and appears indigenous. Living in the target locale will ensure up-to-date knowledge of the current language use.

Our definition of an "expert", then, is a person or a team with sufficient combined knowledge of (1) the languages involved, (2) the cultures, (3) the content of the test, and (4) general principles of testing, to produce a professional quality translation/adaptation of a test. In practice it may be effective to use teams of people with different qualifications (for example, translators with and without expertise in the specific subject, a test expert, etc.) in order to identify areas that others may overlook. In all cases, knowledge of general principles of testing, in addition to knowledge of the test content, should form part of the training that translators receive.

Suggestions for practice. We would suggest the following:

- Choose translators who are native speakers of the target language and have an in-depth knowledge of culture into which a test is being adapted, preferably living in the target locale. A common mistake is to identify persons as translators who know the language, but not very well the culture, because an in-depth knowledge of the culture is often essential for maintaining cultural equivalence. Having the cultural knowledge will identify cultural references (e.g., cricket, Eiffel Tower, President Lincoln, kangaroo etc.), with which local participants may be unfamiliar.

- Choose translators, if possible, with experience in the content of the test, and with knowledge of assessment principles (e.g., with multiple-choice items, the correct answer

should be about as long and no longer or shorter than other answer choices; grammatical cues should not be helpful in locating the correct answer; and, in true-false items, true statements should not be notably longer than false statements).

- Translators with knowledge of test development principles may be nearly impossible, in practice, to find, and thus it would be essential to provide training for translators to provide them with the item writing principles for the formats with which they will be working. Without the training, sometimes overly conscientious translators will introduce sources of error, which can lower the validity of a translated test. For example, sometimes a translator might add a clarifying remark to ensure an intended correct answer is in fact the correct answer. In doing so, the translator may make the item easier than was intended, or the longer correct answer choice may provide a clue to the correct answer to test-wise candidates.

TD-2 (5) Use appropriate translation designs and procedures to maximize the suitability of the test adaptation in the intended populations.

Explanation. This guideline requires that decisions made by translators or groups of translators maximize suitability of the adapted version to the intended population. This means the language should feel natural and acceptable; focusing on functional rather than on literal equivalence. Popular translation designs to achieve these goals are forward translations and backward translations. Brislin (1986) and Hambleton and Patsula (1999) provide full discussions of the two designs, including their definitions, strengths and weaknesses. But it should be noted that both designs have flaws, and so rarely would these two designs provide sufficient evidence to validate a translated and adapted test. The main drawback of the backward translation design is that, if this design is implemented in its narrowest form, no review of the target language version of the test is ever done. The design too often results in a target language version of the test which maximizes the ease of back translation, but sometimes produces a rather awkward target language version of the test.
A double-translation and reconciliation procedure is aimed to address the shortcomings and risks of relying on idiosyncrasies of single translations. In this approach, a third independent translator or an expert panel identifies and resolves any discrepancies between alternative forward translations, and reconciles them into a single version. In large-scale cross-cultural assessment programmes such as PISA, two different language versions (for example, English and French), may be used as separate sources for translation, which are then reconciled into a single target language version (Grisay, 2003). This approach offers important advantages, such as possible discrepancies are identified and reviewed directly in the target language. In addition, using more than one source language helps minimize the impact of cultural characteristics of the source.

Differences in the language structure can cause problems in test translation. For instance, in a well-known scale developed by Rotter and Rafferty (1950) in English, examinees are required to fill in the blanks in incomplete item format such as, *"I like….."; "I regret….."; "I can't ....."*.

However, the same format is inappropriate in the Turkish language, where the object of a sentence must come before the verb and subject. The use of incomplete sentences as in the English version, therefore, would change the answering behaviour completely since the Turkish students should first look at the end of the statement before they fill out the beginning.
In any alternative solutions to this problem, the translated (i.e., target language) version will be somehow different than the source language version in terms of format specifications.

Suggestions for practice. The compilation of judgemental data from reviewers seems especially valuable for checking that this guideline is met:

- Use the rating scales advanced by Brislin (1986), Jeanrie and Bertrand (1999), or Hambleton and Zenisky (2010). Hambleton and Zenisky provide an empirically validated list of 25 different features of a translated test that should be checked during the adaptation process. Sample questions from the Hambleton and Zenisky (2010) include *"Is the language of the translated item of comparable difficulty and commonality with respect to the words in the item in the source language version?"* and *"Does the translation introduce changes in the text (omissions, substitutions, or additions) that might influence the difficulty of the test item in the two language versions?"*

- Use multiple translation designs if practically feasible. For example, a backward translation design can be used to double-check the target version created through double-translation and reconciliation by an expert panel.

- If a test is intended to be used cross-culturally, consider simultaneous / concurrent development of multiple language versions of the test from the start in order to avoid future problems with translating/adapting the source version. More information on concurrent test development can be found, for example, in Solano-Flores, Trumbull, and Nelson-Barber (2002). At the very least, design the source version that enables future translations and avoids potential problems as much as possible; specifically, avoiding cultural references, idiosyncratic item and response formats, etc.

- Considering the syntax differences across languages, using formats that rely on the rigid structure of sentences should be avoided in large-scale international assessments and probably with psychological tests, too, because of the translation problems that may arise.

TD-3 (6) Provide evidence that the test instructions and item content have similar meaning for all intended populations.

Explanation. The evidence demanded by the guideline can be gathered through a variety of strategies (see, for example, van de Vijver and Tanzer, 1997). These strategies include (1) use of reviewers native to local culture and language; (2) use of samples of bilingual respondents; (3)

use of local surveys to evaluate the test; and (4) use of non-standard test administrations to increase acceptability and validity.

Conducting a small try-out of the adapted version of the test is a good idea. The small try-out can employ not just test administration and data analysis, but also, and most importantly, interviews with the administrators and the examinees to obtain their criticisms of the test itself. Other designs using content experts from different language backgrounds, or bilingual content experts, are also possible. For example, bilingual content experts could be asked to rate the similarity of the difficulty of the item formats and content of the two tests. Cognitive interviewing is another method that is showing promise (Levin, et al., 2009).

Suggestions for practice. Several suggestions were offered above for addressing this guideline. For example,

- Use reviewers native to local culture and language to evaluate the test translation/adaptation.

- Use samples of bilingual respondents to provide some suggestions about the equivalence of the two versions of the test, both on test instructions and test items.

- Use local surveys to evaluate the test. These small-scale try-outs can be very valuable. Be sure to interview the administrator and the respondents following the test administration because often administrator and respondent comments are more valuable than the respondents' actual responses to the items in the test.

- Use adapted test administrations to increase acceptability and validity. Following similar test instructions makes no sense if they will be misunderstood by respondents in the second language/cultural group.

TD-4 (7) Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations.

Explanation. Item formats such as five-point rating scales or new item formats such as "drag and drop" or "answer all that are correct" or even "answer one and only one answer choice" can be confusing to respondents who have not seen the item formats before. Even item layouts, the use of graphics, or rapidly emerging computerized item formats can be confusing to candidates. There are many examples of these types of errors found in the United States with their initiative to move much of the standardized testing of children to the computer. Through practice exercises, the problems can be overcome for most children. These new item formats must be familiar to respondents or a source of testing bias is introduced that can distort any individual and group test results.

A newly emerging problem might be associated with computer-administered versions of a test. If respondents are not familiar with the computer-based test platform, a tutorial is needed to ensure that these respondents gain the familiarity they need for a computer-administered test to provide meaningful scores.

Suggestions for practice. Both qualitative and quantitative evidences have a role to play in assessing this guideline. There are several features of an adapted test that might be checked:

- Check that any practice exercises are sufficient to bring respondents up to the level required for them to provide honest and/or responses that reflect their level of mastery of the material.

- Ensure that respondents are familiar with any novel item formats or test administrations (such as a computer-administration) that have been incorporated into the testing process.

- Check that any test conventions (e.g., the placement of any exhibits, or the marking of answers on an answer sheet) will be clear to respondents.

- Again, the rating forms provided by Jeanrie and Bertrand (1999) and Hambleton and Zenisky (2010) are helpful. For example, Hambleton and Zenisky included questions such as "Is the item format, including physical layout, the same in the two language versions?", and "If a form of word or phrase emphasis (bold, italics, underline, etc.) was used in the source language item, was that emphasis used in the translated item?"

TD-5 (8) Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made.

Explanation. Prior to initiating any large-scale test scores reliability and validity studies and/or norming studies that may be time-consuming and expensive, it is important to have confirming evidences about the psychometric quality of the adapted test. There are many psychometric analyses that can be carried out to provide initial evidences of score reliability and validity. For example, at the test development stage, an item analysis using at least a modest sample size (e.g., 100) can provide much-needed data about the functioning of particular test items. Items, which are very easy or hard by comparison to other items, or showing low or even negative discriminating powers, can be reviewed for possible item flaws. With multiple-choice items, it would be appropriate to investigate the effectiveness of item distractors. Problems can be spotted and revisions made. Also, with the same data compiled for item analysis, coefficient alpha or coefficient omega (McDonald, 1999) provide the test developer with valuable information that could be used to support decisions about the appropriate length of the source and target language versions of the test.

In some cases, questions may still remain about certain aspects of the adaptation: Will the test instructions be fully understood? Should the instructions be different to effectively guide test takers in the new language and culture? Will the computer-administered test cause problems for selected respondents (e.g., low socioeconomic status respondents) in the population of interest for the adapted test? Are there too many questions being presented in the available time? All of these questions and many more could be answered with modest-sized validity studies. The goal would be to compile enough data that a decision can be made about whether or not to move forward with the adapted test. If the decision is to move forward, then a series of substantially more ambitious studies can be planned and carried out (e.g., studies of item level DIF, and studies to investigate test factorial structure).

Suggestions for practice. There are a number of basic analyses that can be carried out:

- Carry out a classical item analysis study to obtain information about item level means and item discrimination indices, and, with multiple-choice items or similar selection items, carry out a distractor analysis, too.

- Carry out a reliability analysis (e.g., KR-20 with dichotomously scored items, or coefficient alpha or coefficient omega with polytomously scored items).

- As necessary, carry out a study or two to gain insight about the validity of the adapted test. For example, suppose the adapted test is to be administered via a computer. It may be desirable to carry out a study to evaluate the mode of test administration (i.e., paper-and-pencil versus computer-administered). Suppose that the instructions require respondents to answer all questions. It may be necessary to do some research to determine the best instructions for accomplishing this goal. Researchers have found that it is surprisingly difficult to get some respondents to answer every question, if guessing is being encouraged when respondents do not have the required information.

Confirmation Guidelines

The Confirmation Guidelines are those that are based on empirical analyses of full-scale validity studies.

C-1 (9) Select sample with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses.

Explanation. The data collection design refers to the way that the data are collected to establish norms (if needed) and equivalence among the language versions of a test, and to conduct validity and reliability studies, and DIF studies. A first requirement with respect to the data collection is that samples should be sufficiently large to allow for the availability of stable

statistical information. Though this requirement holds for any type of research, it is particularly relevant in the context of a test adaptation validation study because the statistical techniques needed to establish test and item equivalence (e.g., confirmatory factor analysis, IRT approaches to the identification of potentially biased test items) can most meaningfully be applied with samples large enough to reliably estimate model parameters (the recommended sample size will depend on model complexity and data quality).

Also, the sample for a full-scale validity study should be representative of the intended population for the test. We draw attention to the important paper by van de Vijver and Tanzer (1997), and the methodological contributions found in van de Vijver and Leung (1997), Hambleton, Merenda, and Spielberger (2005), Byrne (2008), and Byrne and van de Vijver (2014), to guide the selection of appropriate statistical designs and analyses. Sireci (1997) provided a discussion of the problems and issues in linking multi-language tests to a common scale.

Sometimes, in practice, the intended population for the target language version of a test may score much lower or higher, and/or be more or less homogeneous than the source language group. This creates major problems for certain methods of analyses, such as reliability and validity studies. One solution is to choose a subsample of the source language group to match the target language group sample. With matched samples, any differences in the results for the matched samples that may be due to differences in the shapes of the distributions in the two groups can be eliminated (see Sireci & Wells, 2010). For example, comparisons of test structure typically involve covariances, and these will vary as a function of the score distributions. By using matched samples, whatever role the distribution of scores might play in the results is matched in the two samples, and so the role of score distributions on the results can be ruled out as an explanation for any differences in the results.

Perhaps one more example might help to explain the problem of different score distributions in the source and target language groups. Suppose the test score reliability is .80 in the source language group, but only .60 in the target language group. The difference might appear worrisome and raise questions about the suitability of the target language version of the test. However, it is often overlooked that reliability is a joint characteristic of the test and the population (McDonald, 1999) – because it depends on both the true score variance (population characteristic) and error variance (test characteristic). Therefore, the same error variance can lead to a higher reliability simply due to the larger true score variance in the source language group. McDonald (1999) shows that the Standard Error of measurement (which is the square root of error variance) is in fact a more appropriate quantity to compare between samples, not reliability. Another alternative using reliability coefficients would be to draw a matched sample of candidates from the source language group and recalculate the test score reliability.

Modern approaches to testing measurement invariance using multiple-group Confirmatory Factor Analysis (CFA) allow for samples with different distributions of the latent traits to be assessed. In such models, while measurement parameters such as item factor loadings and intercepts are assumed equal across groups, the means, variances and covariances of the latent

traits are allowed to vary across groups. This allows for the use of full samples, and accommodates the more realistic scenario of different distributions of measured traits across different populations.

Suggestions for practice. In nearly all research, there are two suggestions that are made when describing the sample(s):

- Collect as large a sample as reasonable given that studies to identify potentially biased test items require a minimum 200 persons per version of the test (Mazor, Clauser & Hambleton, 1992; Subok, 2017). To undertake item response theory analyses and model fit investigations a sample of at least 500 respondents is required (Hulin, Lissak & Drasgow, 1982; Hambleton, Swaminathan & Rogers, 1991), while studies to investigate the factorial structure of a test require fairly large sample sizes, perhaps 300 or more respondents (Wolf, Harrington, Clark & Miller, 2013). Clearly, analyses with smaller samples are possible, too – but the first rule is to generate large participating samples whenever possible.

- Choose representative samples of respondents whenever possible. Generalizations of findings from non-representative samples of respondents are limited. To eliminate differences in the results due to methodological factors such as variations in score distributions, drawing a sample from the source language group to match the target language group is often a good idea. Comparisons of standard errors of measurement may be more appropriate.

C-2 (10) Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations.

Explanation. Establishing the construct equivalence of the source and target language versions of a test is important, but it is not the only important empirical analysis to carry out. Also, approaches for construct equivalence (PC-2) and method equivalence (PC-3) were addressed briefly earlier in the guidelines.

Researchers need to address the equivalence at the item level as well. Item equivalence is studied under the title, *"differential item functioning (DIF) analysis."* In general, DIF exists if two test- takers, from two different (cultural- linguistic) populations, have the same level of the measured trait but have a different response probability on a test item. Overall differences in test performance across the groups could possibly occur, but this does not present a problem by itself. Whereas, when the members of the populations are matched on the construct measured by the test (typically a total test score, or total test score minus the score for the item being studied), and performance differences exist on the item across the groups, DIF is present in the item. This type of analysis is performed for each item in the test. Later, an attempt is made to understand the reasons for the DIF in the items, and, based on this judgemental review, some items may be identified as flawed, and altered or removed completely from the test.

Two important potential sources of DIF to evaluate are translation problems and cultural differences. More specifically, DIF may be due to (1) translation non-equivalence that occurs from source to target language versions of the test such as familiarity with the vocabulary used, change in item difficulty, change in equivalence of the meaning, etc., and (2) cultural contextual differences (Scheuneman & Grima, 1997; van de Vijver & Tanzer, 1997; Ercikan, 1998, 2002; Allalouf, Hambleton, & Sireci, 1999; Sireci & Berberoğlu, 2000; Ercikan, et al., 2004; Li, Cohen, & Ibera, 2004; Park, Pearson & Reckase, 2005; and Ercikan, Simon, & Oliveri, 2013).

During translation, there is the possibility of using less common vocabulary in the target language. The meanings could be the same in the translated versions, but, in one culture, the word could be more common compared to the other. It is also possible to change the difficulty level of the item as a result of translation due to sentence length, sentence complexity, and use of easy or difficult vocabulary as well. Meaning may also change in the target language with deletion of some parts of the sentences, inaccurate translations, having more than one meaning in the vocabulary used in target language, non-equivalent impressions of the meanings of some words across the cultures, etc. Above all, cultural differences might cause the items to function differently across the languages. For example, words like "hamburger" or "cash register" may not be understood or have a different meaning in two cultures.

There are at least four groups of analyses to check if items are functioning differently across the language and/or cultural groups. These are (a) IRT-based procedures (see, e.g., Ellis, 1989; Thissen, Steinberg, & Wainer, 1988; 1993; Ellis & Kimmel, 1992), (b) Mantel-Haenszel (MH) procedure and extensions (see, e.g., Dorans & Holland, 1993; Hambleton, Clauser, Mazor, & Jones, 1993; Holland & Wainer, 1993; Sireci & Allalouf, 2003), (c) logistic regression (LR) procedures (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993), and (d) restricted factor analysis (RFA) procedure (Oort & Berberoğlu, 1992).

In the IRT-based approaches, test-takers across two languages are matched based on the latent trait scores. In MH and LR methodologies, the observed or estimated test score is used as the matching criterion prior to comparing item performance of respondents in the two groups. Although the sum score is most popular matching criterion in these procedures, other estimated scores, for instance from factor analysis can also be used. These scores are also iteratively "purified" by deleting the questionable items. The matching criterion should be valid and reliable enough to evaluate the DIF properly. In RFA, each item is regressed on the grouping variable (potential violator) as well as the latent trait. Each item loading is set free and the fit to the model is evaluated with reference to the null model where no item is loaded on the grouping variable (no DIF model). If the model provides significantly better fit, this flags the item as DIF.

When a test is dimensionally complex, finding an appropriate matching criterion is an issue (Clauser, Nungester, Mazor & Ripkey, 1996). Using multivariate matching criteria, such as different factor scores obtained as a result of factor analysis, might change the item level DIF

interpretations as well. Accordingly, this guideline suggests that, if the test is multidimensional, the researchers might use various criteria to flag the items as DIF, and evaluate the items which are consistently flagged as DIF with respect to various matching criteria. Multivariate matching can reduce the number of items exhibiting DIF across the language and cultural groups.

These methodologies could require different sample sizes. MH, LR, and RFA are models that may reliably and validly work for relatively small samples compared to IRT-based techniques, which require larger samples for valid parameter estimations. Another consideration is the type of item response data. MH, LR, and RFA can be applied to binary - scored data. Other approaches, such as the generalized MH, are needed with polytomous response data.

This guideline requires researchers to locate possible sources of method bias in the adapted test. Sources of method bias include (1) the different levels of test motivation of participants, (2) differential experience on the part of respondents with psychological tests, (3) more speediness of the test in one language group than the other, (4) differential familiarity with the response format across language groups, and (5) heterogeneity of response style, etc. Biases in responses have been, for example, a major concern in interpreting PISA results, and have received some research attention.

Finally, yet importantly, this guideline will require researchers to address construct equivalence. There are at least four statistical approaches for assessing construct equivalence across source and target language versions of a test: Exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling (MDS), and comparison of nomological networks (Sireci, Patsula, & Hambleton, 2005).

According to van de Vijver and Poortinga (1991), factor analysis (both EFA and CFA) are the most frequently used statistical technique to assess whether a construct in one culture is found in the same form and frequency in another culture. This statement from 1991 remains true today, though the statistical modelling approaches have advanced considerably (see, for example, Hambleton & Lee, 2013, Byrne, 2008). Since, with EFA, it is difficult to compare separate factor structures, and there are no commonly agreed-upon rules for deciding when the structures can be considered equivalent, statistical approaches such as CFA (see, for example, Byrne, 2001, 2003, 2006, 2008) and weighted multidimensional scaling (WMDS) are more desirable as they can simultaneously accommodate multiple groups (Sireci, Harter, Yang, & Bhola, 2003).

There have been many studies in which CFA was used to evaluate whether the factor structure of an original version of a test was consistent across its adapted versions (e.g., Byrne & van de Vijver, 2014). CFA is attractive for evaluating structural equivalence across adapted tests because it can handle multiple groups simultaneously, statistical tests of model fit are available, and descriptive indices of model fit are provided (Sireci, Patsula, & Hambleton, 2005). The capability to handle multiple groups is especially important as it is becoming common to adapt tests into many languages (e.g., some intelligence measures are now translated/adapted into

over one hundred languages, and, in TIMSS and OECD/PISA, tests are adapted into over 30 languages). As the strict requirement of zero cross-loadings in CFA, however, often does not fit well the data on complex multidimensional instruments, Exploratory Structural Equation Modelling (ESEM) is becoming more and more popular, especially with personality data or more complex and inter-related variables (Asparouhov & Muthén, 2009).

WMDS is another attractive approach for evaluating construct equivalence across different language versions of an assessment. Like EFA, WMDS analysis does not require specifying test structure a priori, and, like CFA, it allows for the analysis of multiple groups (e.g., Sireci, et al., 2003).

Van de Vijver and Tanzer (1997) have suggested that cross-cultural researchers should examine the reliability of each cultural version of the test of interest and search for both convergent and discriminant validity evidence in each cultural group. These studies may often be more practical than studies of test structure that require very substantial sample sizes.

It must be recognized, however, that comparison of test-taker performance across two language versions of a test is not always the goal of translating/adapting a test. Perhaps, for example, the goal is simply to be able to assess test-takers in a different language group on a construct. In this instance, careful examination of the validity of the test in the second language group is essential, but research to find evidence of the equivalence of the two forms is not so critical. The importance of this guideline will depend on the intended purpose or purposes of the test in the second language (i.e., target language group). Tests like those used in PISA or TIMSS require evidence of high content overlap because the results are used to compare the achievement of students in many countries. The use of a depression inventory translated from English into Chinese for researchers to study depression or for counsellors to assess depression of their clients would not require high overlap in content. Instead, validity to support the depression inventory in China would be needed.

This guideline can also be addressed with statistical methods after the test has been adapted. For example, if cultural groups are thought to differ on important variables irrelevant to the construct measured, comprehensive designs and statistical analyses can be used to control for these 'nuisance' variables. Analysis of covariance, randomized-block designs, and other statistical techniques (regression analysis, partial correlation, etc.) can be used to control the effects of unwanted sources of variation among the groups.

Suggestions for practice. This is a very important guideline and there are many analyses that might be carried out. For equivalence analyses, we offer the following suggestions for practice:

- If sample sizes are sufficient, carry out a comparative study of the construct equivalence of the source and target language versions of the test. There are lots of software packages to facilitate these analyses (see Byrne, 2006).

- Carry out exploratory (preferably rotating to a target structure – so-called "target rotation") or confirmatory factor analysis, and/or weighted multidimensional scaling analysis, to determine the level of agreement in the structure of the test of interest across language and/or cultural groups. The requirement of large sample sizes (10 persons per variable) makes these studies difficult to carry out in many cross-cultural studies. An excellent model for a study of this type is Byrne and van de Vijver (2014).

- Look for evidence of convergent and discriminant validity (essentially, look for correlational evidence among a set of constructs and check the stability of these correlations across language and/or cultural groups) (see van de Vijver & Tanzer, 1997).

For DIF analyses, some suggestions are identified below. For more sophisticated approaches, researchers are encouraged to read the professional literature on DIF:

- Carry out a DIF analysis using one of the standard procedures (if items are binary scored, the Mantel-Haenszel procedure may be the most straightforward; if items are polytomously scored, the generalized Mantel-Haenszel procedure is an option). Other more cumbersome solutions include IRT-based approaches. If sample sizes are more modest, a "delta plot" can reveal potentially flawed items. Conditional comparisons are another possibility (for a comparison of results with small sample methods, see, for example, Muñiz, Hambleton, & Xing, 2001).

C-3 (11) Provide evidence supporting the norms, reliability and validity of the adapted version of the test in the intended populations.

Explanation. The norms, validity evidence and reliability evidence of a test in its source language version do not automatically apply to other possible adaptations of the test into different cultures and languages. Therefore, empirical validity and reliability evidence of any new versions developed must also be presented. All kinds of empirical evidence supporting the inferences made from the test should be included in the test manual. Special attention should be paid to the five sources of validity evidence based on: test content, response processes, internal structure, relations to other variables, and consequences of testing (AERA, APA, NCME, 2014). Exploratory and confirmatory factor analysis, structural equation modelling, and multitrait-multimethod analyses are some of the statistical techniques that can be used to obtain and analyse data addressing validity evidence based on internal structure.

Suggestions for practice. The suggestions are the same as would be required for any test that is being considered for use:

- If the norms developed for the original version of the test are suggested to be used with the adapted version, evidence should be provided that this use is statistically appropriate and fair. If no evidence can be provided for such use of the original norms,

specific norms should be developed for the adapted version according to the standards for norm development.

- Compile a sufficient amount of reliability evidence to justify the use of the target language version of the test. The evidence might normally include an estimate of internal consistency (e.g., KR-20, or coefficients alpha or omega).

- Compile as much validity evidence as is needed to determine whether the target language version of the test should be used. The type of evidence compiled would depend on the intended use of the scores (e.g., content validity for achievement tests, predictive validity for aptitude tests etc.).

C-4 (12) Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test.

Explanation. When linking two language versions of a test to a single reporting scale, several options are possible. If a common set of items is used, the functioning of these common items across the two language groups should be evaluated, and if differential functioning is observed, their removal from the data used in establishing the link should be considered. Delta plots (Angoff & Modu, 1973) serve this purpose well, and Cook and Schmitt-Cascallar (2005) provided a good illustration of how to use delta plots to identify items that have a different meaning for the two groups of examinees. Not all item types have the same potential to link between language versions. Item difficulty and discrimination parameter estimates derived in the framework of item response theory for the common items can be plotted to help identify inappropriately performing common items (see Hambleton, Swaminathan, & Rogers, 1991).

But linking (i.e., "equating") scores across two language versions of a test will always be problematic because strong assumptions need to be made about the data. Sometimes, a highly problematic assumption is made that the different language versions of the test are equivalent, and then scores from the two versions of the test are used interchangeably. Such an assumption may have merit with mathematics tests because translation/adaptation is typically straightforward. It may have merit, too, if the two versions of the test have been carefully constructed, and so the assumption can be made that the source language version of the test functions with the source language population in an equivalent way to which the target language version of the test functions in the target language population. This assumption may have merit if all of the other evidence available suggests that the two language versions of the test are equivalent and there are no method biases influencing the scores in the target language version of the test.

Two other solutions exist, but neither is perfect. First, the linking could be done with a subsample of the items that are deemed to be essentially equivalent in the two language versions of the test. For example, the items may be the ones that were judged as very easy to translate/adapt. In principle, the solution could work, but requires the linking items and the

remainder of the test items to be measuring the same construct. A second solution involves linking through a sample of test-takers who are bilingual. With this sample taking both versions of the test, it would be possible to establish a score conversion table. The sample could not be too small, and in the design, the order of presentation of the forms of the test would be counterbalanced. The big assumption in this approach is that the candidates are truly bilingual, and so, apart from the relative difficulties of the forms, the candidates should do equally well on both forms. Any difference is used to adjust the scores in converting scores from one version of the test to the other.

Suggestions for practice. Linking scores across adapted versions of a test is going to be problematic at the best of times because all of the equating designs have at least one major shortcoming. Probably the best strategy is to completely address all of the steps for establishing score equivalence. If the evidence addressing the three questions below is strong, even the scores from the two versions of the test can be treated interchangeably:

- Is there evidence that the same construct is being measured in the source and target language versions of the test? Does the construct have the same relationship with other external variables in the new culture?

- Is there strong evidence that sources of method bias have been eliminated (e.g., no time issues, formats used in the test are equally familiar to candidates, no confusion about the instructions, no systematic misrepresentation in one group or the other, standardized instructions, absence of response styles (extreme ratings, differential motivation…)?

- Is the test free of potentially biased test items? Here, a plot of p values or, better, delta values, from items in the two versions of the test can be very helpful. Points not falling along the linear equating line should be studied to determine if the associated items are equally suitable in both languages. DIF analyses provide even stronger evidence about item equivalence across language and cultural groups.

- If linking of scores is attempted, then an appropriate linking design needs to be chosen and implemented. Evidence for the validity of the design should be provided.

Administration Guidelines

A-1 (13) Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores.

Explanation. Implementing the administration guidelines should start from an analysis of all factors that can threaten the validity of test scores in a specific cultural and linguistic context. Experience with the administration of an instrument in a monolingual or monocultural context may already be helpful in anticipating problems that can be expected in a multilingual or

multicultural context. For example, experienced test administrators often know which aspects of instruction may be difficult for respondents. These aspects may remain difficult after translation or adaptation. Applications of instruments in a new linguistic or cultural context could also find issues, not previously found in monocultural applications.

Suggestions for practice. It is important with this guideline to anticipate potential factors that might create problems in test administration. Some of those factors that need to be studied to ensure fairness in test administration are the following:

- Clarity of test instructions (including translation of those instructions), the answering mechanism (e.g., the answer sheet), the allowable time (one common source of error is the failure to allow sufficient time for test-takers to finish), motivation for candidates to complete the test, knowledge about the purpose of the test, and how it will be scored.

A-2 (14) Specify testing conditions that should be followed closely in all populations of interest.

Explanation. The goal of this guideline is to encourage test developers to establish testing instructions and related procedures (e.g., testing conditions, time limits, etc.) that can be followed closely in all populations of interest. This guideline is primarily meant to encourage test administrators to stick to standardized instructions. At the same time, accommodations might be specified to address special subgroups of individuals within each population who may need testing accommodations such as additional time, larger print, extra quiet test administration conditions, and so on. In the testing field today, these are known as "test accommodations." The goal of these accommodations is not to inflate test-taker scores, but rather to create a testing environment for these candidates so that they can show what they may feel, or know and can do.

Variations from the standardized testing conditions should be noted, so that, later in the process, these variations and their impact on generalizations and interpretations can be considered.

Suggestions for practice. This guideline may in part overlap with A-1 (13), but it is restated here to highlight the importance of candidates taking the test under as similar conditions as possible. This is essential if the scores from the two language versions are going to be used interchangeably. Here are some suggestions:

- Testing instructions and related procedures should be adapted and re-written in a standardized way, which is suitable to the new language and culture.

- If testing instructions and related procedures are changed to the new cultures, administrators should be trained on the new procedures; they should be informed to respect to these procedures and not to the original ones.

Score Scales and Interpretation Guidelines

SSI-1 (15) Interpret any group score differences with reference to all relevant available information.

Explanation. Even if a test has been adapted through technically sound procedures, and validity of the test scores has already been established to some extent, it should be kept in mind that the meaning of inter-group differences can be interpreted in many ways because of cultural or other differences across the participating countries and/or cultures. Sireci (2005) reviewed the approach for evaluating the equivalence of two different language versions of a test by administering the separate language versions of the test to a group of test-takers who are proficient in both languages (bilingual) and who come from the same cultural or language group. He outlined some research design options for equivalence studies using bilingual respondents, listed the possible confounding variables needing to be controlled, and offered some valuable suggestions for interpreting findings.

Suggestions for practice. One suggestion for improving practice follows:

- Depending on the research question (or context for which group comparisons are made), a number of possible interpretations may be considered, before finally setting on one. For example, it is important to rule out differential motivation to perform well on the test prior to inferring that one group performed better on the test than another. There may be context effects, too that significantly impacted test performance. For example, one group of persons may simply be part of a less effective education system, and this would have a significant impact on test performance.

SSI-2 (16) Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported.

Explanation. When comparative studies across language and cultural groups are the central focus of the translation and adaptation initiative, the multi-language versions of a test need to be placed on a common reporting scale, and this is carried out through a process called "linking" or "equating." This requires substantial sample sizes, and evidence that construct, method, and item bias are not present in the adapted version of the test.

Van de Vijver and Poortinga (2005) delineated several levels of test equivalence across language and cultural groups and their work is especially helpful in understanding this concept; in fact, the original concept was introduced by these authors. For example, they pointed out that measurement unit equivalence requires that reporting scales in each group have the same metric, thus ensuring differences between people within the groups have the same meaning. (For example, differences between males and females in a Chinese sample can be compared to a

French sample). However, valid direct score comparisons can only be done when scores show the highest level of equivalence, called scalar equivalence or full score equivalence, which requires scales in each group have the same measurement unit and the same origin across groups.

Numerous methods (both in the framework of classical test theory and item response theory) have been put forward for linking or equating scores from two groups (or language versions of a test). Interested readers can refer to Angoff (1984) and Kolen and Brennan (2004) to gain a deeper understanding of this topic. Cook and Schmitt-Cascallar (2005) suggest a basis for understanding statistical methods that are currently available for equating and scaling educational and psychological tests. The authors describe and critique specific scale linking procedures used in test adaptation studies, and illustrate selected linking procedures and issues by describing and critiquing three studies that have been carried out over the past twenty years to link scores from the *Scholastic Assessment Test* to the *Prueba te Aptitude Academica*.

Suggestions for practice. The key point here is that the test scores should not be over-interpreted:

- Interpret the results based on the level of validity evidence that is available. For example, do not make comparative statements about the levels of respondent performance in the two language groups unless measurement invariance has been established for test scores being compared..

Documentation Guidelines

Doc-1 (17) Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population.

Explanation. The importance of this guideline has been realized and emphasized by many researchers (see, for example, Grisay, 2003). TIMSS and PISA have been very successful in observing this guideline by carefully documenting the changes throughout the adaptation work. With this information, there can be focus on the suitability of the changes that were made.

The technical documentation should also contain sufficient detail of the methodology for future researchers to replicate the procedures used on the same or other populations. It should contain sufficient information from the evidence of construct equivalence and scaling equivalence (if carried out) to support the use of the instrument in the new population. Where inter-population comparisons are to be made, the documentation should report the evidence used to determine the equating of scores between populations.

Sometimes, the question arises about the intended audience for the technical documentation. The documentation should be written for the technical expert and for persons who will be required to evaluate the utility of the test for use in the new or other populations. (A brief supplementary document could be added for the benefit of a non-expert.)

Suggestions for practice. Adapted tests should have a technical manual that documents all the qualitative and quantitative evidence associated with the adaptation process. It is especially helpful to document any changes that were made to accommodate the test in a second language and culture. Basically, technical persons and journal editors will want documentation on the process which was completed to produce and validate the target language version of the test. Of course, too, they will want to see the results from all of the analyses. Here are the types of questions that need to be addressed:

- What evidence is available to support the utility of the construct and the adapted test in the new population?
- What item data were collected and from what samples?
- What other data were obtained to assess content, criterion-related and construct validity?
- How were the various data sets analysed?
- What were the results?

Doc-2 (18) Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population.

Explanation. The documentation should be written for people who will be using the test in practical assessment settings. It should be consistent with the good practice defined by the International Test Commission Guidelines on Test Use (see www.InTestCom.org).

Suggestions for practice. The test developer should provide specific information on the ways in which socio-cultural and ecological contexts of the populations might affect performance on the test. The user manual should:

- Describe the construct(s) measured by the test and summarize the information; describe the adaptation process.
- Summarize the evidence supporting the adaptation, including evidence for cultural suitability of the item content, test instructions, response format, etc.

- Define the suitability of using the test with various subgroups within the population and any other restrictions on use.

- Explain any issues that need to be considered in relation to good practice in test administration.

- Explain if and how inter-population comparisons can be made.

- Provide the information necessary for scoring and norming (e.g., relevant norm look-up tables) or describe how users can access scoring procedures (e.g., where these are computer-based).

- Provide guidelines for the interpretation of results, including information on the implications of validity and reliability data on the inferences that may be drawn from test scores.

FINAL WORDS

We have done our best to deliver a set of guidelines to help test developers and test users in their work. However, for the guidelines and other efforts to change poor practices to have an effect, there must be good dissemination mechanisms in place. A recent systematic review by Rios and Sireci (2014) demonstrated that the majority of test adaptation projects in the published literature did not, in fact, followed the ITC Guidelines that have been available for around 20 years now. We thereby encourage the readers to make every effort to increase awareness among colleagues of this Second Edition as a primary source of best practices to which so many professionals around the world have contributed.

At the same time, we know that just as the first edition of these guidelines is now being replaced, so in turn will these second edition guidelines. The well-known AERA, APA, and NCME test standards are now in their sixth edition (AERA, APA, & NCME, 2014). We expect the ITC Guidelines for Adapting Tests to undergo another revision too in the coming years. If you know of new studies that should be cited, or influence the third edition, or you want to offer new guidelines or revisions to the 18 guidelines presented here please let the ITC know. You can contact the current chair of the Research and Guidelines committee that produced the second edition and/or the secretary of the ITC at the email address found on www.InTestCom.org.

REFERENCES

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Modu, C. C. (1973). Equating the scales of the Prueba de Apititud Academica and the Scholastic Aptitude Test (Research Rep No. 3). New York: College Entrance Examination Board.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural modeling. *Structural Equation Modeling, 16*, 397-438.

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 Publications.

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing, 1*, 55-86.

Byrne, B. (2003). Measuring self-concept measurement across culture: Issues, caveats, and application. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), *International advances in self research*. Greenwich, CT: Information Age Publishing.

Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Publishers.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*, 872-882.

Byrne, B. M., & van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10,* 107-132.

Byrne, B. M., & van de Vijver, F.J.R. (2014). Factorial structure of the Family Values Scale from a multilevel-multicultural perspective. *International Journal of Testing, 14*, 168-192.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripley, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214.

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139-170).

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and Practice* (pp. 137-166).

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

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77*, 177-184.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543-533.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3), 199-215.

Ercikan, K., Gierl, J. J., McCreith, T., 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*(3), 301-321.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *An international handbook for large-scale assessments* (pp. 110-124). New York:

Grégoire, J., & Hambleton, R. K. (Eds.). (2009). Advances in test adaptation research [Special Issue]. *International Journal of Testing, 9*(2), 73-166.

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

Hambleton, R. K. (2002). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127-240.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1-16.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Hambleton, R. K., & Lee, M. (2013). Methods of translating and adapting tests to increase cross-language validity. In D. Saklofske, C. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child assessment* (pp. 172-181). New York: Oxford University Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., Yu, L., & Slater, S. C. (1999). Field-test of ITC guidelines for adapting psychological tests. *European Journal of Psychological Assessment, 15* (3), 270-276.

Hambleton, R. K., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 46-74). New York, NY; Cambridge University Press.

Harkness, J. (Ed.). (1998). *Cross-cultural survey equivalence*.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Javaras, K. N., & Ripley, B. D. (2007). An 'unfolding' latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*, 454-463.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission Guidelines: Keeping validity in mind. *European Journal of Psychological Assessment, 15*(3), 277-283.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika, 68*, 563-583.

Kolen, M. J., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Stapleton Kudela, M., Stark, D., & Thompson, F. E. (2009). Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods, 3*(1), 13-25.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115-135.

Mazor, K.H., Clauser, B.E., & Hambleton, R.K. (1992). The effect of simple size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451.

Muniz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traduccion y adaptacion de los tests: segunda edicion. *Psicothema, 25(2),* 149-155.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115-135.

Oort, F. J., & Berberoğlu, G. (1992). Using restricted factor analysis with binary data for item bias detection and item analysis. In T. J. Plomp, J. M. Pieters, & A. Feteris (Eds.), *European Conference on Educational Research: Book of Summaries* (pp. 708-710). Twente, the Netherlands: University of Twente, Department of Education.

Park, H., Pearson, P. D., & Reckase, M. D. (2005). Assessing the effect of cohort, gender, and race on DIF in an adaptive test designed for multi-age groups. *Reading Psychology, 26*, 81-101.

Rios, J., & Sireci, S. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14*(4), 289-312.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Rotter, J.B. & Rafferty, J.E. (1950). *Manual: The Rotter Incomplete Sentences Blank: College Form*. New York: Psychological Corporation.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and Black examinees. *Applied Measurement in Education, 10*(4), 299-319.

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

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Mahwah, NJ: Lawrence Erlbaum Publishers.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148-166.

Sireci, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*(3), 229-248.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. Merenda, & C. Spielberger, C. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-116). Mahwah, NJ: Lawrence Erlbaum Publishers.

Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing, 3*(2), 129-150.

Sireci, S. G., & Wells. C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33-68). Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-129.

Subok, L. (2017). Detecting differential item functioning using the logistic regression procedure in small samples. *Applied Psychological Measurement, 41(1)*, 30-43.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for Test Adaptation. *European Journal of Psychological Assessment, 15*, 258-269.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Mahwah, NJ: Lawrence Erlbaum Publishers.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67-113). Mahwah, NJ: Lawrence Erlbaum Publishers.

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

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33-51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-308). Dordrecht, the Netherlands: Kluwer Academic Publishers.

van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17-24.

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

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodical issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and*

psychological tests for cross-cultural assessment (pp. 39-64). Mahwah, NJ: Lawrence Erlbaum Publishers.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263-279.

Wolf, E.J., Harrington, K.M., Clark, S.L., & Miller, M.W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73(6),* 913–934.

APPENDIX A. ITC GUIDELINES FOR TRANSLATING AND ADAPTING TESTS CHECKLIST

Here is a checklist to remind you of the eighteen ITC Guidelines. We recommend that you check those that you feel you have dealt with satisfactorily in your test translation/adaptation project, and then attend to those that remain unaddressed.

Pre-Condition Guidelines

[ ] **PC-1 (1) Obtain the necessary permissions from the holder of the intellectual property rights relating to the test before carrying out any adaptation.**

[ ] **PC-2 (2) Evaluate that the amount of overlap in the definition and content of the construct measured by the test in the populations of interest is sufficient for the intended use (or uses) of the scores.**

[ ] **PC-3 (3) Minimize the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest.**

Test Development Guidelines

[ ] **TD-1 (4) Ensure that the adaptation process considers linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise.**

[ ] **TD-2 (5) Use appropriate translation designs and procedures to maximize the suitability of the test adaptation in the intended populations.**

[ ] **TD-3 (6) Provide evidence that the test instructions and item content have similar meaning for all intended populations.**

[ ] **TD-4 (7) Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations.**

[ ] **TD-5 (8) Collect pilot data on the adapted test to enable item analysis, reliability assessment and other small-scale validity studies, so that any necessary revisions to the adapted test can be made.**

Confirmation Guidelines

C-1 (9) Select sample with characteristics that are relevant for the intended

[ ] **use of the test and of sufficient size and relevance for the empirical analyses.**

[ ] **C-2 (10) Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations.**

[ ] **C-3 (11) Provide evidence supporting the norms, reliability and validity of the adapted version of the test in the intended populations.**

[ ] **C-4 (12) Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test.**

Administration Guidelines

[ ] **A-1 (13) Prepare administration materials and instructions to minimize any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores.**

[ ] **A-2 (14) Specify testing conditions that should be followed closely in all populations of interest.**

Score Scales and Interpretation Guidelines

[ ] **SSI-1 (15) Interpret any group score differences with reference to all relevant available information.**

[ ] **SSI-2 (16) Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported.**

Documentation Guidelines

[ ] **Doc-1 (17) Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population.**

[ ] **Doc-2 (18) Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population.**

APPENDIX B. GLOSSARY OF TERMS

Alpha (or sometimes called "Coefficient Alpha" or "Cronbach's Alpha"). The reliability coefficient of a test whose items are assumed to measure one attribute in common and have equal discriminations (therefore is a special case of Omega – see below). In more general conditions, it is a lower bound to reliability.

Backward Translation Design. With this design, a test is translated from the source language version to the target language version of a test by one group of translators, and then the target language version is back translated to the source language, by a second translator or group of translators. The original source and back-translated source versions are compared, and a judgement is made about the suitability of the source language version of the test. If the two source language versions are very close, the assumption is made that the target language version of the test is acceptable.

Confirmatory Factor Analysis. A hypothesis about the structure of a test is made, and then analyses are carried out to obtain the test structure from the correlation matrix of items in the test. A statistical test is carried out to see if the hypothesized and estimated test structure are close enough that the null hypothesis that the two structures are equal cannot be rejected.

Delta Values. Delta values are simply non-linearly transformed *p* values and applied to binary-scored items. An item delta value is the normal deviate corresponding to the area under a normal distribution (mean=0.0, SD=1.0) where the area under the normal distribution is equal to the proportion of candidates answering the item correctly. So, if p=.84, then the delta value for the item would be −1.0. This transformation is made with the belief that delta values are more likely to be on an equal interval scale than p values.

Differential Item Functioning. There is a class of statistical procedures that can determine if an item is functioning more or less the same in two different groups. Comparisons of performance are made by first matching examinees on the trait measured by the test. When differences are observed, it is said that the item is potentially biased. Effort is made to explain the conditional differences in performance for candidates in the two groups matched on the trait measured by the item.

Double-Translation and Reconciliation. In this translation design, an independent translator or an expert panel identifies and resolves any discrepancies between alternative forward translations, and reconciles them into a single version.

Examinees. Used interchangeably in the testing field with "test-takers", "candidates", "respondents", and "students" (if achievement tests are involved).

Exploratory Factor Analysis. Factor analysis is a statistical procedure that is applied, for example, with the correlation matrix produced by the inter-correlations among a set of items in

a test (or a set of tests). The goal is to try and explain the inter-correlations among the test items (or tests) in terms of a small number of factors that are believed to be measured by the test (or tests). For example, with a mathematics test, a factor analysis might identify the fact that the items fall into three clusters - computation items, concepts, and problem-solving. It might be said, then, that the mathematics test is measuring three factors - computation, math concepts, and math problem-solving.

Forward Translation Design. With this design, a test is adapted into the target language of interest by a translator or, more often, a group of translators, and then a different translator or group of translators judge the equivalence of the source and target language versions of the test.

Item Response Theory. A class of statistical models for linking item responses to a trait or set of traits that are being measured by the items in the test. Specific IRT models can handle both binary and polytomous response data. Binary data might come from scoring multiple-choice items or true-false items on a personality scale. Polytomous response data might come from the scoring of performance tasks or essays on an achievement test, or from rating scales such as "Likert."

Kuder-Richardson Formula 20. (Or, sometimes simply called "KR-20."). The reliability coefficient of a test formed from binary items, which are assumed to measure one attribute in common and have equal discriminations.

Localization. This is a popular term in the testing field that is used to describe the process for making a test prepared in one language and culture acceptable for use in another. An equivalent term would be translation/adaptation.

Logistic Regression Procedure for Identifying Differential Item Functioning. This statistical procedure is one more way to carry out DIF analyses. A logistic curve is fit to the performance data of each group and then the two logistic curves, one for each language group, are compared statistically.

Mantel-Haenszel Procedure for Identifying Differential Item Functioning. A statistical procedure for comparing the performance of two groups of examinees on a test item. The comparisons are made for examinees in each group who are matched on the trait or construct measured by the test.

Omega (or sometimes called "Coefficient Omega" or "McDonald's Omega"). The reliability coefficient of a test whose items are assumed to measure one attribute in common (fit the general factor model). More generally applicable than coefficient Alpha.

PISA. Stands for "Programme for International Student Achievement." This is the international assessment of achievement that is sponsored by the Organization for Economic Cooperation and Development (OECD) with more than 40 participating countries.

Simultaneous Test Development. Development of source and target language questionnaires simultaneously, using standardised translation quality control procedures. Large-scale international projects increasingly use simultaneous development in order to avoid the problem that the version developed in one language cannot be translated/adapted to all the languages of the study.

Source Language Version. The language in which a test is originally written.

Structural Equation Modelling. A set of complex statistical models that are used to identify the underlying structure of a test or a set of tests. Often these models are used to investigate causal inferences about the relationships among a set of variables.

Target Language Version. The language to which a test is translated/adapted. So, for example, if a test is translated from English to Spanish, the English version is often called the "source language version" and the Spanish version is called the "target language version."

Test Dimensionality. This refers to the number of dimensions or factors that a test is measuring. Often, this analysis is carried out statistically using one of many procedures, including eigenvalue plots or structural equation modelling.

Test Score Equating. A statistical procedure for linking scores on two tests that measure the same construct but where the tests are not strictly parallel.

TIMSS. Stands for "Trends in International Mathematics and Science Studies" and is an international assessment of grades 4, 8, and 12 students in countries in the areas of mathematics and science and sponsored by IEA.

WDMS. This stands for "Weighted Multidimensional Scaling." It is another statistical procedure for addressing test dimensionality.