Psychological Assessment –
Science and Practice

Principles and Methods of Test Construction

Standards and Recent Advances

Karl Schweizer
Christine DiStefano
(Editors)

hogrefe

Principles and Methods of Test Construction

About the Editors

Karl Schweizer, PhD, is a Professor in the Department of Psychology at Goethe University Frankfurt, Germany. He received his PhD from the University of Freiburg and was a faculty member there and at the University of Tübingen. His research interests focus on assessment, cognitive-differential psychology, psychological statistics, and research methods. He has served on the Executive Committee of the European Association of Psychological Assessment and was editor of *European Journal of Psychological Assessment.*

Christine DiStefano, PhD, is an Associate Professor of Educational Measurement in the College of Education at the University of South Carolina. Her research interests are in validity, survey construction, structural equation modeling with categorical data, and classification.

Acknowledgments

We would like to acknowledge the support of people who were important for the success of this book. We are grateful to the editor of the book series, Dr. Anastasia Efklides of Thessaloniki University, and the evaluation committee, which included Dr. Tuulia Ortner of Salzburg University and Dr. Willibald Ruch of Zurich University. Furthermore, we would like to thank Rob Dimbleby, who served as the representative of the publisher. We would like to acknowledge the invaluable contribution of the authors of the chapters who invested their expertise, time, and effort in providing state-of-the-art manuscripts. Chris wishes to thank George, Nora, and Fred for their unending support.

Psychological Assessment – Science and Practice

Each volume in the series *Psychological Assessment – Science and Practice* presents the state-of-the-art of assessment in a particular domain of psychology, with regard to theory, research, and practical applications. Editors and contributors are leading authorities in their respective fields. Each volume discusses, in a reader-friendly manner, critical issues and developments in assessment, as well as well-known and novel assessment tools. The series is an ideal educational resource for researchers, teachers, and students of assessment, as well as practitioners.

Psychological Assessment – Science and Practice is edited with the support of the European Association of Psychological Assessment (EAPA).

Psychological Assessment – Science and Practice, Vol. 3

Principles and Methods of Test Construction

Standards and Recent Advances

Edited by
Karl Schweizer and Christine DiStefano

hogrefe

Library of Congress Cataloging in Publication information for the print version of this book is available via the Library of Congress Marc Database under the LC Control Number 2016930840

Library and Archives Canada Cataloguing in Publication

Principles and methods of test construction : standards and recent advances / edited by Karl Schweizer and Christine DiStefano.

(Psychological assessment--science and practice ; vol. 3)
Includes bibliographical references and index.
Issued in print and electronic formats.
ISBN 978-0-88937-449-2 (paperback).--ISBN 978-1-61676-449-4 (pdf).--ISBN 978-1-61334-449-1 (epub)

1. Psychological tests--Design and construction. 2. Psychometrics.
I. Schweizer, Karl, 1951-, author, editor II. DiStefano, Christine, 1969-, author, editor
III. Series: Psychological assessment--science and practice ; v. 3

BF176.P75 2016	150.28'7	C2016-900441-4 C2016-900442-2

PUBLISHING OFFICES

USA: Hogrefe Publishing Corporation, 38 Chauncy Street, Suite 1002, Boston, MA 02111
Phone (866) 823-4726, Fax (617) 354-6875; E-mail customerservice@hogrefe.com

EUROPE: Hogrefe Publishing GmbH, Merkelstr. 3, 37085 Göttingen, Germany
Phone +49 551 99950-0, Fax +49 551 99950-111; E-mail publishing@hogrefe.com

SALES & DISTRIBUTION

USA: Hogrefe Publishing, Customer Services Department, 30 Amberwood Parkway, Ashland, OH 44805
Phone (800) 228-3749, Fax (419) 281-6883; E-mail customerservice@hogrefe.com

UK: Hogrefe Publishing, c/o Marston Book Services Ltd., 160 Eastern Ave., Milton Park, Abingdon, OX14 4SB, UK
Phone +44 1235 465577, Fax +44 1235 465556; E-mail
direct.orders@marston.co.uk

EUROPE: Hogrefe Publishing, Merkelstr. 3, 37085 Göttingen, Germany
Phone +49 551 99950-0, Fax +49 551 99950-111; E-mail
publishing@hogrefe.com

OTHER OFFICES
CANADA: Hogrefe Publishing, 660 Eglinton Ave. East, Suite 119-514, Toronto, Ontario, M4G 2K2
SWITZERLAND: Hogrefe Publishing, Länggass-Strasse 76, CH-3000 Bern 9

Format: EPUB
ISBN 978-0-88937-449-2 (print) • ISBN 978-1-61676-449-4 (PDF) • ISBN 978-1-61334-449-1 (EPUB)
http://doi.org/10.1027/00449-000

eBook-Herstellung und Auslieferung:
Brockhaus Commission, Kornwestheim
www.brocom.de

Table of Contents

[1] Part I

Introduction

[2] [3] # Chapter 1

Introduction

Karl Schweizer[1] and Christine DiStefano[2]
[1]Department of Psychology, Goethe University Frankfurt, Germany
[2]Department of Educational Studies, University of South Carolina, USA

During the past 10–20 years, the methodology used to develop and to administer tests has experienced a number of substantial advancements. However, many of these advancements are dispersed across numerous outlets, such as journal articles, conference papers, or presentation materials. A major motivation for undertaking this book project was to collect knowledge concerning advancements in test construction, to provide information about the current practices, and to disseminate information about recent advances. We hope that in this way we may equip researchers and students with sufficient knowledge to successfully execute test construction projects, rather than learning of advancements through unfavorable interactions with discussants, editors, or journal reviewers. So, to us (and hopefully also to the readers), it appears to be valuable to collect information about the state of the art in test construction. The selection of the chapters is the result of our perceptions regarding advancements in test construction as well as issues that may benefit from further elaboration.

The first section provides a platform to examine and strengthen the role of the underlying theory when designing tests. The standards that govern test construction are explored to provide readers with information about the history and evolution of the guidelines that regulate best practices. Also included are chapters that discuss a modern test theory approach toward designing new measures according to a theory base and the study of the

associated psychometric properties. Both the factor analytic and the item response theory (IRT) frameworks are provided.

The second section considers features related to item format and test presentation. A variety of item formats are examined to assist researchers with best practices for writing items for cognitive or affective measures. Discussion includes both formats that are more traditional (e.g., multiple choice) as well as newer formats that incorporate technological advances into items, producing a more interactive testing experience for examinees. Also, computerized and online assessments provide favorable preconditions for the increased utilization of adaptive testing. Online assessment has become more and more important for research as the Internet provides the opportunity of accessing large samples without a personal contact or a visit to a central location, such as a laboratory, career center, or testing site, needed. Thus, it is necessary to understand the possibilities as well as the potential pitfalls and shortcomings of this type of assessment. Moreover, adaptive testing shows a number of advantages that generally require fewer items to achieve a precise measurement of latent constructs with a shorter time commitment; such advantages need to be balanced against the challenges that online testing poses.

The third section discusses features related to model testing and selection, primarily from the structural equation modeling framework. Recent advancements have seen the rise of alternative [4] estimators to deal with issues often encountered in test construction, such as analysis of nonnormally distributed observed level data or analysis and/or ordered categorical data. The chapters included provide information regarding selection of an estimation technique that can accommodate the characteristics of the collected data. Further, model selection and reporting of model-data fit information has been a controversial topic for many years and, in a way, has created insecurity of what constitutes best practice.

Group-specific biases of psychological measures have become a concern because of public sensitivity and, therefore, demand an especially

exhaustive treatment. The fourth section provides information regarding statistical methods that enable the identification of group-specific bias. These chapters discuss differential item functioning, originating from the IRT framework, as well as multiple group testing from the structural modeling framework. The use of these methods can be helpful in evaluating the general appropriateness of the items selected or the differences in conceptualization of latent variables for relevant subgroups of a population.

The fifth section of the book discusses topics of special relevance. For example, test construction assuming one latent source gives rise to the expectation of one underlying dimension. This preferred model has stimulated the construction of measures showing a high degree of homogeneity, but may impose a structure that is not appropriate for the construction of measures representing intermediate or even higher-order constructs. The elaboration of the bifactor model may provide a method for researchers to consider. Also, since the publication of the seminal work by Campbell and Fiske it is known that the true variance characterizing a measure may be inflated by a contribution of the observational method. More recent research suggests that the presence of method effects in survey data is more likely than its absence. Chapters describing modern techniques for conducting multitrait–multimethod research as well as examination of method effects due to position effects are included. Another challenge is the creation of abridged scales or screeners, as short forms of instruments are increasingly common owing to testing expenses in terms of time and cost. Methods for creating both abridged forms and screening instruments are provided in this section.

[5] # Part II

Major Approaches to Test Construction

[6] [7] # Chapter 2

The Use of Standards in Test Development

Fons J. R. van de Vijver
Department of Culture Studies, Tilburg University, The Netherlands

The present chapter deals with the question of the use of standards in test development. The best known example of such standards are the "Standards for Educational and Psychological Testing" published by the American Educational Research Association, the American Psychological Association, and the national Council on Measurement in Education (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). The latest version, published in 2014, has just been released. This version updated earlier editions of the standards. For example, the 1999 publication of the standards was an update of the 1985 edition (http://www.teststandards.org/history.htm), which mainly covered the groundwork of test development, grounded in classical statistics, such as classical test theory (Lord & Novick, 1968). At that time it was the most comprehensive overview of do's and don'ts in test development. The 1999 version was updated to recognize the following (American Psychological Association, 2013):

> Changes in federal law and measurement trends affecting validity; testing individuals with disabilities or different linguistic backgrounds; and new types of tests as well as new uses of existing tests. The Standards is written for the professional and for the educated layperson and addresses professional and technical issues

> of test development and use in education, psychology and employment.

Changes from 1999 to the current standards are discussed in this chapter. The Standards, as they are usually referred to, were originally meant for the American market of test users, test developers, and policy makers. However, since the Standards were so comprehensive and similar standards were not formulated in many other countries, the book became an authoritative source in the area of test development.

The aforementioned quotation reflects important characteristics of many standards. First, they are compiled on a voluntary basis. Also, they provide links with the recent developments and psychometrics so as to ensure their scientific soundness and up-to-date nature. Finally, standards are influenced by various developments in science and society. Psychology is not unique in its attempts to enhance the quality of its services by implementing standards. The ISO (International Organization for Standardization; http://www.iso.org/iso/home/about.htm) is the world's largest developer of voluntary international standards. In this organization, "a standard is a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose" (http://www.iso.org/iso/home/standards.htm). Since 1947, almost 20,000 standards [8] have been developed, covering almost all aspects of technology and business. ISO certification has become an important hallmark of quality. Psychology as a discipline does not have the highly formalized systems of service delivery and quality checks as implemented in the ISO Standards. Still, the same underlying reasoning of enhancing quality by agreeing on standardized procedures can be found in psychology.

Since the launch of the Standards for Educational and Psychological Testing in 1955, many more standards have been developed. The present chapter gives an overview of recent advances in the development of standards in the domain of psychological assessment. It is impossible to review all aspects

of the standards presented here. Therefore, I present the contents of standards in tables and deal with other aspects in the main text. I focus on various sets of standards that have been proposed in the last 20 years so as to accommodate new target groups and new modes of administration. Furthermore, I move outside of the realm of psychological and educational testing where the standards were originally developed. More specifically, I describe guidelines that were designed for international testing, notably dealing with translations and adaptations, standards for computer-based and Internet testing, standards for test use, and standards for quality control. Conclusions are drawn in the last section.

Two caveats are needed on terminology. The first is the distinction between educational and psychological testing. This distinction is made more in the American literature than in the European literature, in which the two types of assessment are often considered together. I follow here the European tradition and refer to testing and assessment as involving both educational and psychological instruments. Second, the literature uses two related concepts to refer to desirable features of psychological assessment: standards and guidelines. There is a subtle, yet essential, difference between the two. Standards typically have a prescriptive meaning. Standards describe prerequisites of instruments and their administration needed to ensure valid outcomes of the assessment process. Guidelines, on the other hand, are typically less prescriptive and are formulated as aspired or best practices. The distinction between these aspects seems to be easy to make. In practice, the distinction can be fuzzy as the terms are not always used from the perspective of this difference. Some guidelines are prescriptions, while some standards describe recommendable practices.

The Standards for Educational and Psychological Testing

The Standards for Educational and Psychological Testing are an initiative of the American Educational Research Association (AERA), the American

Psychological Association (APA), and the National Council on Measurement in Education (NCME). The Standards for Educational and Psychological Testing have been very influential in psychology and education; the latest version, the fifth revision, was launched in 2014 (a description of the changes in this version was made by Plake & Wise, 2014). The history of the standards has clearly shown that defining norms regarding development, administration, and interpretation of tests helps to advance the quality of the field of assessment. References in the literature to the Standards for Educational and Psychological Testing are numerous (see, e.g., http://teststandards.org/files/Standards_citations_Jan_2010.pdf) and to the best of my knowledge, their reception has not been controversial. The standards are meant to provide criteria for the evaluation of tests, testing practices, and the effects of test use (AERA, APA, & NCME, 2014). The standards are not meant to influence policy, but they can provide recommendations on how psychometrics can be used to underline policy decisions. For instance, rather than prescribing which minimum cutoff score should be established for an admission test, the standards can help to identify conditions that are critical for determining cutoff scores.

[9] The Standards for Educational and Psychological Testing cover three domains (see Table 2.1). The description of each domain starts with a general presentation of the context. Important concepts are defined and an overview of the main issues in the domain is presented.

Table 2.1. Overview of topics covered in 1999 Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014)

(a) Aim and domains covered	
Aim	To promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices
Domains	1. Foundations

covered

2. Operations

3. Testing Applications

(b) Guidelines

Part I. Foundations

1. Validity
2. Reliability/precision and errors of measurement
3. Fairness in testing

Part II. Operations

1. Test design and development
2. Scores, scales, norms, score linking, and cut scores
3. Test administration, scoring, reporting, and interpretation
4. Supporting documentation for tests
5. The rights and responsibilities of test takers
6. The rights and responsibilities of test users

Part III. Testing Applications

1. Psychological testing and assessment
2. Workplace testing and credentialing
3. Educational testing and assessment
4. Uses of tests for program evaluation, policy studies and accountability

Source: http://www.apa.org/science/programs/testing/standards.aspx

The first part of the Standards for Educational and Psychological Testing, called Foundations, refers to the core business of psychometrics: test construction, evaluation, and documentation. Validity, viewed as pivotal in

the psychological assessment process, refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (AERA, APA, & NCME, 2014). Thus, the standards describe how validity evidence can be obtained. The standards emphasize the need for finding validity evidence, specifying intended test use, and for clearly stating the recommended interpretation and use of test scores. The common theme of norms in the validity section of the standards is that it is incumbent on the user to provide validity evidence and to refrain from making untested assumptions about test use. The chapter on reliability, referring to consistency when a testing procedure is repeated, emphasizes the need for standardized administration procedures so as to increase reliability. The chapter describes various approaches to reliability, such as classical test theory with its internal consistency coefficients, standard errors, and interrater agreement (e.g., Lord & Novick, 1968), item response theory (e.g., Van der Linden & Hambleton, 1997) with its test information functions, and generalizability theory (e.g., Webb, 1991) with its multiple ways of computing reliability. The remaining chapters of the first section deal with test development and revision, scales, norms, administration scoring, and documentation. The standards mentioned there provide an excellent overview of commendable practices in these domains.

[10] The next section is called Operations. The section is a major revision of the previous (1999) version. An important aspect of the section is fairness. Issues of fairness are salient aspects of assessment in all multicultural societies. Where in the 1999 version there was an emphasis on fairness for various subgroups in society that could be adversely affected by the use of psychological tests, such as women, members of specific ethnic groups, or people from the LGBT community, the 2015 version takes a broader perspective.

A measure is fair if it is free from bias. The conceptualization that is used in the standards is borrowed from the literature on item bias (usually labeled *differential item functioning*; Van de Vijver & Leung, 1997). An item is said to be biased if individuals from different groups (e.g., different genders, age

groups, or ethnic groups) with the same standing on the latent trait that is being assessed do not have the same expected scores on the item. A strong aspect of this definition is that it is supported by many statistical procedures to identify this bias, such as analysis of variance, regression analysis, exploratory and confirmatory factor analysis, and contingency table analyses (see Van de Vijver & Leung, 1997; Zumbo, 2007). The weak spot of this definition is its emphasis on item-related sources of bias. In my experience, important sources of cross-cultural differences in scores that are not related to the target construct that is measured are usually not item based but instrument based. For example, there are large cross-cultural differences in response styles, such as social desirability (Van Hemert, Van de Vijver, Poortinga, & Georgas, 2002). Vieluf, Kuenther, and Van de Vijver (2013) analyzed the 2008 TALIS data on teacher self-efficacy. At country level, significant positive correlations were found between self-efficacy and job satisfaction; in addition, teacher self-efficacy was related to collectivism, modesty, and extremity scoring. It was concluded that mean score differences between 23 countries were strongly influenced by extremity scoring. Such response styles challenge the validity of cross-cultural comparisons in personality and attitude assessments, among many other domains. However, statistical procedures to identify item bias will typically not pick up cross-cultural differences in response styles, as the latter tend to have a global rather than item-specific influence on the assessment process.

Another problem with the chapter on operations is its focus on instrument characteristics. There are useful standards describing how to validate measures in each group and how to examine the identity of the meaning of test scores; however, there is no description of which group characteristics could impact bias. Examples are previous test exposure, response styles, education, and various other background variables that tend to differ across target groups, notably in the assessment of ethnic groups. Similarly, in the chapter on the assessment of individuals of diverse linguistic backgrounds, the description of the problem and the Standards for Educational and

Psychological Testing do not refer to specific groups or acculturation issues, but only to recommendations to be cautious and to present evidence about the validity of the measure. Apart from these qualms, the chapter on fairness describes many valuable procedures to achieve equitable treatment among all test takers.

The third part of the Standards for Educational and Psychological Testing, called Testing Applications, describes issues in testing applications, such as work place testing and credentialing. The Standards in this part are based on a rich experience of psychological assessment in many different domains and a keen awareness of the legal issues accompanying psychological assessment. There is also a chapter on specific issues in educational assessment.

The Standards for Educational and Psychological Testing are the most elaborate standards available in the field of psychology and education. It is a major strength of these standards that many experts have been involved in the process of writing standards and providing feedback on earlier versions. As a consequence, the standards integrate theoretical and practical insights [11] in the assessment process. The quality of the standards is so high that it is easy to appreciate why they have become so influential. In their quest for quality, the authors have attempted to be inclusive and exhaustive in many ways. For example, in the chapter on reliability, various theoretical perspectives on the concepts are presented, emphasizing common themes rather than highlighting differences between approaches. The quest has also been beneficial from another perspective. The Standards for Educational and Psychological Testing have influenced testing practices and have served as a template in many countries. Many standards that have been formulated are relevant in various countries. Notably the first part, dealing with test construction, evaluation, and documentation, has many standards that are widely applicable. The part on fairness also has a broad applicability, even though particular issues related to fairness may be country specific as the diversity of countries differs in nature. The part on testing applications is also widely applicable, although there are aspects such as credentialing that

are relatively more important in a country with much high-stakes testing, such as the US, than in other parts of the world.

Plake and Wise (2014) warn against possible misuse and misinterpretation of the standards. Their description is interesting as their recommendations go beyond the standards. The first aspect they mention is that the standards are meant to provide professional guidelines and are not meant to be applied in a literal fashion. Professional judgment, based on solid scientific insights, should undergird any decision about the application of the standards. Furthermore, the authors emphasize that there is no authority to enforce or guard applications of the standards, which implies that any claim about compliance with the standards should be checked. Finally, the standards cover a rapidly evolving field; as a consequence, older versions may no longer apply and elements of the current version may also need modification in the near or distant future. In short, the standards should be used judiciously and should not be used as a detailed guide of what (not) to do.

Guidelines for International Testing

In 1992 the International Test Commission (ITC; http://www.intestcom.org) took the initiative to set up a project to develop guidelines for international assessment; an updated version was published in 2010. Various international psychological associations participated in the project: European Association of Psychological Assessment, European Test Publishers Group, International Association for Cross-Cultural Psychology, International Association of Applied Psychology, International Association for the Evaluation of Educational Achievement, International Language Testing Association, and International Union of Psychological Science. The idea behind development of the Guidelines for International Testing was the perceived need to attend to issues of quality during the process of translating and adapting tests. In those days there was a continually growing body of international studies and there was no agreement as to the criteria

for evaluating quality standards regarding reliability, validity, sampling procedures, and translation procedures that apply to such studies (Hambleton, 1994, 2001; Gregoire & Hambleton, 2009; Hambleton, Yu, & Slater, 1999; Van de Vijver & Hambleton, 1996). The criteria that were taken to apply to these international tests adhered to the standard psychometric practice, as described earlier, as well as implementing checks to assess the quality of the translations. These criteria were greatly expanded by the ITC Guidelines.

The most common translation check was the use of the so-called back-translation procedure (Brislin, 1970). Such a procedure consists of three steps. In the first, an instrument is translated from a source language to a target language, followed in the second step by an independent [12] back translation. In the final step, the source and back-translated versions are compared. If the two instruments are identical or do not deviate in any major respect, the translation is taken to be adequate. If the two versions are not identical, some adjudication is needed, which usually takes place through interactions between the researcher and one or more translators. The main advantage of this widely applicable procedure is that the researcher does not need to have knowledge of the target language. However, in the 1990s it had already become clear that the procedure also has some disadvantages. For example, the most important quality criterion is the correspondence between the original and back-translated version. This favors translations that stay very close to the original source (literal translations). Such translations often do not have the natural flow and clarity of the original version. Notably if the original text includes metaphorical expressions (e.g., “I feel blue”), close translations are near impossible and back translations are almost never identical to the original text. Various new procedures have been proposed for translating instruments (Harkness, 2003), such as the comparison of multiple, independent forward translations, followed by an adjudication procedure to select the best translation. Also, a committee approach has been advocated. A group of experts, usually combining linguistic, cultural, and

psychological knowledge of the target construct, jointly prepare a translation. The main advantage of such an approach is the balanced treatment of various perspectives that are relevant in the translation process. Whereas translations in the past were mainly viewed as involving linguistic aspects, the Zeitgeist of the 1990s brought forth the idea that translating requires multiple types of expertise and that a good translation must try to do justice to psychological, linguistic, and cultural considerations. One of the most significant changes created by this new way of thinking has been the introduction of the term *adaptation* (Hambleton, Merenda, & Spielberger, 2005); subsequently, the term *translation* is used less frequently nowadays. Compared with the old procedures, adaptations tend to be more tailored to the specific cultural context in which the instrument will be applied.

These ITC Guidelines for International Testing are not the only ones that have been developed in the domain of test adaptations. Another example is Comparative Survey Design and Implementation (http://ccsg.isr.umich.edu/archive/pdf/fullguide061108.pdf). The group behind this initiative has developed an extensive set of guidelines concerning how to develop and implement cross-cultural surveys. There are various topics in the guidelines about Comparative Survey Design and Implementation that are minimally covered or not covered under ITC guidelines, such as costs, ethics considerations, sample design, and harmonizing data (e.g., converting socioeconomic status data based on country-specific indicators to a common metric). These Comparative Survey Design and Implementation Guidelines have been written from the perspective of large-scale international reviews, such as the International Social Survey Programme (http://www.issp.org). Where psychological and educational guidelines are often somewhat more focused on statistical procedures to ascertain equivalence, these survey guidelines focus more on design and implementation issues. Therefore, these are a valuable addition to psychological and educational guidelines.

The Guidelines for International Testing are presented in Table 2.2. It is significant that the guidelines start with recommendations regarding the context. Thus, rather than opening with specific recommendations, the guidelines start with the notion that is considered to be crucial in developing adaptations: It is important to study the context in which this study will take place and try to minimize the relevant yet confounding cross-cultural differences in the background variables as much as possible. This recommendation does not primarily refer to psychometric concentrations or to procedures to prepare translations, but emphasizes the need to take the linguistic and cultural context seriously. The second recommendation that deals with a description of the context of the study in general argues that we cannot simply assume that [13] constructs or instruments work the same way in all cultures, and that this should be empirically demonstrated.

Table 2.2. International Test Commission Guidelines for Translating and Adapting Tests (version January 2010)

(a) Aim and domains covered	
Aim	"The objective was to produce a detailed set of guidelines for adapting psychological and educational tests for use in various different linguistic and cultural contexts" (ITC, nd)
Domains covered	1. Cultural context 2. Technicalities of instrument development and adaptation 3. Test administration 4. Documentation and interpretation
(b) Guidelines	
Context	
C.1 Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.	
C.2 The amount of overlap in the construct measured by the test or instrument in the	

populations of interest should be assessed.

Test Development
and Adaptation

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

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

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

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

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

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

D.7 Test developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the test or instrument, and (2) identify problematic components or aspects of the test or instrument which may be inadequate to one or more of the intended populations.

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

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

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

Administration

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

the preparation of appropriate materials and instructions.

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

A.3 Those aspects of the environment that influence the administration of a test or instrument should be made as similar as possible across populations of interest.

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

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

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

Documentation/Score Interpretations

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

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

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

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

Source: International Test Commission at http://www.intestcom.org/upload/sitefiles/40.pdf

These context guidelines describe the framework from which the other guidelines have been derived. This second type of guideline refers to test development and adaptation. These form the core of the adaptation guidelines, as they describe the do's and don'ts in designing new instruments. There are some recurring themes in these guidelines. The first is the need to take full cognizance of the cultural and linguistic context of

the study. The second is the need to combine adequate instrument design with appropriate statistical analysis. A good test adaptation starts from a conceptual analysis of the underlying construct(s), including an analysis of the applicability of the construct and its measure in the new cultural context. These theoretical considerations, which may result in smaller or larger changes of the stimuli so as to increase their cultural fit, should be complemented by cognitive interviews (Miller, 2003), pilot studies, or field trials in which the appropriateness of the new instrument is tested. In the next stage, statistical evidence should be accumulated to demonstrate the adequacy of the instrument in the new cultural context. If a comparative study is conducted in the quantitative stage, evidence should be collected to test that the instrument measures the same construct in each culture by demonstrating invariance (Van de Vijver & Leung, 1997; Vandenberg & Lance, 2000).

Administration Guidelines deal with issues arising in implementing the instrument in the new cultural context, including the need to keep the ambient conditions of testing as similar as possible across cultures, the need to standardize test instructions and administrations, and the need to minimize the influence of the test administrator on the test outcome. Some of these aspects tend not to be influential in applications of an instrument within a single cultural group, but experience shows that these factors can contribute to unwanted score differences in cross-cultural applications.

[15] The last set of Guidelines for International Testing deal with documentation and score interpretations. The guidelines in this section refer to the need to clearly document all the adaptations that were implemented, the reasons for the adaptations, as well as a description of the potential influence of these adaptations on scores. In short, these guidelines express the frequently observed need to describe and document the changes of the original instrument in the adaptation process and to provide evidence for the validity of these procedures.