



Universidad Nacional de Catamarca - Facultad de Humanidades
Instituto de Investigación en Teorías del Arte y Estética
Revista Rigel – ISSN 2525-1945.

An Analytical Examination of International Guidelines for Translating and Adapting Educational and Psychological Tests: Toward an Objective Reference Framework

Fadila Maatoug

Doctor of Psychology, Mohamed Boudiaf University of M'Sila, Algeria
Email: fadila.maatoug@univ-msila.dz

Ammar Souici

Associate Professor, Class A, Constantine 2 University, Algeria
Email: ammar.souici@univ-constantine2.dz

Received: 15/04/2026, Accepted: 15/07/2026, Published: 02/10/2026

Abstract

Numerous psychological measurement instruments originated and flourished predominantly within Western sociocultural environments. Consequently, their architectural design and psychometric standardization inherently mirrored the specific cultural, linguistic, and environmental realities of their native societies. In the process of cross cultural transfer and translation, researchers in psychology and psychometrics must critically account for environmental, cultural, civilizational, and social disparities that fundamentally distinguish one society from another. This research paper aims to identify the principal recommendations established in the guidelines for translating and adapting tests formulated by the International Test Commission. It systematically traces the mandatory methodological procedures required to adapt and translate assessment instruments systematically and rigorously, accounting for target population characteristics, temporal variables, and cultural bias. Accordingly, the authors provide an analytical reading of the eighteen international guidelines across their six functional categories: pre conditions, test development, confirmation through empirical validation, administration, scoring and interpretation, and technical documentation. This inquiry addresses contemporary developments in educational and psychological measurement to ensure valid, fair, and equitable cross cultural assessment.

Keywords: International guidelines, test translation, cross cultural adaptation, psychological tests, psychometrics, cultural equivalence.



1. Introduction

The discipline of psychological measurement has undergone continuous, progressive evolution since its inception, propelled by dedicated efforts of psychologists and psychometricians seeking to advance standardized, objective assessment tools. This scientific momentum extended beyond Western academia into the Arab world. Initially, Arab scholarship relied on direct literal translation of psychological tests designed in Western societies. Subsequently, pressing scientific necessities compelled researchers to investigate how to develop and calibrate measurement instruments tailored to Arab psychological and social realities. Arab scholars have since deployed sustained endeavors to translate and adapt diverse psychological tests encompassing cognitive, emotional, and personality dimensions.

Direct importation of Western instruments and their mechanical application in original formats without accounting for cultural specificity, idiomatic nuances, and temporal shifts within recipient societies generates severe methodological distortions. This realization underscores the necessity of cultural adaptation, particularly within the Algerian and wider Arab contexts.

In 2007, the Council of the International Test Commission appointed an expert committee of six psychometricians tasked with revising and updating the international guidelines for test adaptation. This initiative integrated emerging technological breakthroughs and accumulated empirical expertise, culminating in the second edition of the International Test Commission Guidelines for Translating and Adapting Tests, finalized between 2005 and 2015 and published in 2017. The revised guidelines comprise eighteen operational recommendations organized into six functional categories: pre conditions, test development, confirmation and empirical validation, test administration, scoring and score interpretation, and comprehensive documentation. Each guideline provides detailed operational rationale, practical implementation suggestions, and audit checklists to ensure methodological rigor.

Accordingly, this paper presents an analytical deconstruction of the second edition of the International Test Commission guidelines, delineating the structural phases essential for rigorous cross cultural adaptation of psychological and educational tests.

2. Problem Statement

Psychological testing faces a pervasive methodological challenge known as cultural bias, or cultural specificity. Cultural bias introduces systematic measurement error, distorting the true score estimations of individuals and groups. Consequently, cross cultural measurement requires rigorous empirical verification of linguistic structure, construct equivalence, and behavioral predictive validity. Cultural bias is directly amplified when instruments are translated into environments markedly divergent from the culture of origin, impairing construct validity.

Kaufman, 2000, observed that the vast majority of psychological instruments transferred across cultural frontiers exhibit pronounced cultural bias. Numerous easily administered instruments evaluate only a narrow fraction of the target psychological trait, disregarding the authentic manifestations of that trait within divergent cultural settings. Kaufman therefore called for constructing and adapting tests in harmony with contemporary theoretical models of human behavior, giving full consideration to cultural contexts and situational factors. Furthermore, modern psychometricians emphasize the necessity of replacing outdated literal translations of Anglo American tests with rigorously adapted instruments capable of accommodating rapid cognitive, technological, and multicultural transformations.

To mitigate these systematic errors, the International Test Commission formulated comprehensive guidelines governing cross cultural test adaptation. The central problem of this study is articulated through the following core question: What are the fundamental analytical dimensions and methodological principles embedded within the International Test Commission guidelines for translating and adapting educational and psychological tests?

3. Conceptual Foundations of Cross Cultural Assessment

3.1 The Psychological Test

Anastasi, 1976, defined the psychological test as an objective and standardized measure of a sample of behavior. Behavior in this context encompasses verbal proficiency, mechanical aptitude, fundamental personality dimensions such as extraversion and introversion, or precise psychomotor coordination evaluated through specialized diagnostic tasks.

3.2 Test Adaptation

Hambleton, as cited in Hambleton, Merenda, and Spielberger, 2006, page 22, defined adaptation as encompassing the complete continuum of activities involved in cross cultural assessment. This process begins with determining whether an instrument can validly assess the same psychological construct in another language and culture, selecting qualified bilingual and bicultural translators, deciding necessary procedural modifications, and executing empirical verifications to establish structural equivalence between the original and adapted versions.

Bousalem, 2015, page 22, characterized psychological test adaptation as all methodological procedures executed by a researcher from initial construct feasibility evaluation to securing linguistically, culturally, and psychologically equivalent expressions. Adaptation transcends literal linguistic translation, incorporating deliberate, progressive modifications supported by empirical psychometric evidence to ensure that the adapted instrument functions validly within the new cultural demographic.

3.3 Scientific Translation of Psychological Instruments

Scientific translation in psychometrics, as conceptualized by Hambleton, 1999, denotes establishing culturally, psychologically, and linguistically equivalent concepts and expressions between source and target languages. It requires matching verbal frequency across languages while accounting for internal regional dialects. Boukraa, 2020, page 185, identified five essential dimensions of measurement equivalence:

- Content Equivalence: Ensuring that each item content remains relevant, appropriate, and meaningful for assessing the target phenomenon across all surveyed cultures.
- Semantic Equivalence: Guaranteeing that the underlying meaning of each item remains identical following linguistic translation.
- Technical Equivalence: Preserving the assessment modality and administration format, such as written self report inventories, across linguistic groups.
- Normative Equivalence: Ensuring that score interpretations and standard score distributions reflect identical behavioral thresholds across populations.
- Conceptual Equivalence: Verifying that the measurement tool evaluates the exact same theoretical psychological construct across diverse cultural environments.

Hambleton et al., 2006, page 22, highlighted that international psychometricians prefer the term test adaptation over test translation. While translation denotes linguistic textual rendering, adaptation encapsulates the broader systemic adjustments necessary to ensure cultural, conceptual, and measurement equivalence across different languages and societies.

3.4 The Imperative of Cultural Adaptation

Abbas, 2017, page 136, delineated critical hazards resulting from the direct application of unadapted translated tests in Arab and Algerian contexts:

- Vulnerability to Response Faking: Poorly translated items often obscure subtle psychometric nuances, prompting respondents to provide socially desirable answers that conform to cultural expectations rather than authentic personal traits.
- Cultural Incongruence: Direct translations frequently contain metaphors, idioms, or scenarios alien to Algerian cultural experience, leading respondents to approach the test with detachment or confusion, thereby compromising measurement accuracy.
- Theoretical Ambiguity: Researchers sometimes utilize foreign tests without mastering the underlying psychological theories that guided original construct operationalization, undermining construct validity.
- Inappropriate Sample Extrapolation: Instruments standardized on specific Western demographic strata are frequently administered indiscriminately to culturally, educationally, or developmentally distinct groups.
- Distortions from Item Rephrasing: Altering item syntax across languages can shift item difficulty or provoke differential response patterns, altering the psychometric properties of the scale.

4. Analytical Deconstruction of the International Test Commission Guidelines

The International Test Commission Guidelines for Translating and Adapting Tests, Second Edition, 2017, encompass eighteen operational standards structured into six core methodological categories:

4.1 Pre condition Guidelines

The pre condition category establishes the ethical, legal, and conceptual groundwork required before commencing test adaptation:

Pre	condition	Guideline	1
<i>Guideline PC 1: Obtain explicit authorization and formal copyright licensing from the intellectual property holder of the test prior to initiating any adaptation workflows.</i>			

Analytical Rationale: Intellectual property rights protect the moral and economic interests of test authors and publishers. As defined by the World Intellectual Property Organization, copyright encompasses the organizational architecture, item structure, scoring protocols, and diagnostic materials of tests. Replicating test architecture or diagnostic algorithms without authorized licensing constitutes intellectual property infringement.

Pre	condition	Guideline	2
<i>Guideline PC 2: Evaluate the degree of congruence between the construct definition and item content across source and target linguistic and cultural groups.</i>			

Analytical Rationale: Cross cultural equivalence requires that the assessed psychological construct possesses identical conceptual meaning across populations. Researchers must gather preliminary empirical evidence, examine analogous literature, and verify that the target construct is not culturally bound, ensuring measurement stability across language editions.

Pre	condition	Guideline	3
<i>Guideline PC 3: Systematically identify and minimize construct irrelevant cultural and linguistic variance that may impede intended test usage.</i>			

Analytical Rationale: Cultural distance between source and target societies encompasses linguistic divergence, family organization, religious norms, and values. Test developers must evaluate whether item delivery modes, such as computer interfaces or graphical illustrations, introduce cultural bias. Translators must be native speakers of the target language and culture, as illustrated by the comparative findings of Hambleton, Yu, and Slater, 1999.

4.2 Test Development Guidelines

Test development guidelines govern the translation process, item adaptation, and initial psychometric calibration:

Test	Development	Guideline	1
------	-------------	-----------	---

Guideline TD 1: Ensure translation and adaptation procedures account for linguistic, psychological, and cultural characteristics through expert translator teams.

Analytical Rationale: Effective psychometric translation requires multidisciplinary teams possessing linguistic fluency, psychological expertise, and cultural mastery. Contemporary standards reject reliance on isolated translators, demanding collaborative review panels to guarantee conceptual rather than literal equivalence.

Test	Development	Guideline	2
<i>Guideline TD 2: Employ rigorous translation designs, including forward and backward translation protocols, to optimize test appropriateness.</i>			

Analytical Rationale: Translators must strive for functional linguistic equivalence rather than mechanical literal replication. Methodological frameworks analyzed by Brislin, 1986, and Hambleton and Patsula, 1999, demonstrate that backward translation designs require dual review panels to prevent awkward syntax that facilitates reverse translation while distorting psychological meaning.

Test	Development	Guideline	3
<i>Guideline TD 3: Provide robust empirical evidence that test instructions and item content carry equivalent meaning across all target groups.</i>			

Analytical Rationale: As demonstrated by Van de Vijver and Tanzer, 1997, evidence of semantic equivalence is established through native expert panels, bilingual cognitive testing, localized pilot investigations, and cognitive debriefing interviews with test administrators and examinees.

Test	Development	Guideline	4
<i>Guideline TD 4: Verify that item formats, rating scales, scoring categories, and testing conventions are familiar and appropriate for the target population.</i>			

Analytical Rationale: Novel item formats, such as five point Likert scales, computerized drag and drop interactions, or timed multiple choice questions, can confuse examinees unaccustomed to these testing conventions. Pilot orientation exercises must be incorporated to neutralize format related bias.

Test	Development	Guideline	5
<i>Guideline TD 5: Conduct small scale pilot studies to execute item analysis, estimate preliminary reliability, and refine item wording.</i>			

Analytical Rationale: Prior to expensive national standardization, pilot studies on modest samples of approximately one hundred examinees yield essential data on item difficulty, discrimination parameters, and distractor performance. Calculating Cronbach Alpha and McDonald Omega reliability coefficients guides decisions regarding optimal scale length and item retention.

4.3 Confirmation and Empirical Validation Guidelines

Confirmation guidelines govern statistical verification of construct comparability, measurement invariance, and scoring linkage:

Confirmation	Guideline	1
	<i>Guideline C 1: Select representative sample sizes of adequate statistical power to conduct robust psychometric modeling.</i>	

Analytical Rationale: Advanced psychometric techniques, including confirmatory factor analysis and item response theory, require substantial sample sizes to yield stable parameter estimates. Sample sizes must reflect the statistical complexity of the structural model.

Confirmation	Guideline	2
	<i>Guideline C 2: Provide statistical evidence establishing construct, method, and item equivalence, including Differential Item Functioning analysis.</i>	

Analytical Rationale: Evaluating equivalence requires identifying Differential Item Functioning. A test item exhibits Differential Item Functioning when examinees from different cultural groups who possess identical levels of the underlying latent trait exhibit significantly different probabilities of answering the item correctly. Items displaying significant differential performance must be refined or eliminated.

Confirmation	Guideline	3
	<i>Guideline C 3: Validate score interpretations through the five sources of validity evidence established by professional testing standards.</i>	

Analytical Rationale: Validity evidence established in the source culture cannot be assumed in the target culture. Researchers must substantiate target score interpretations across the five validity sources codified by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 2014: test content, response processes, internal structure, relations to external variables, and testing consequences, utilizing confirmatory factor analysis and structural equation modeling.

Confirmation	Guideline	4
	<i>Guideline C 4: Utilize appropriate statistical equating and linking methods to compare scores across different language versions.</i>	

Analytical Rationale: Linking multiple language versions to a common scale requires anchor item designs or concurrent calibration. Angoff and Modu, 1973, and Cook and Schmitt Cascallar, 2005, demonstrated that delta plots and item response theory difficulty parameters effectively detect malfunctioning common anchor items.

4.4 Test Administration Guidelines

Administration guidelines ensure standardized, equitable testing conditions across diverse cultural contexts:

Administration	Guideline	1
<i>Guideline A 1: Design administration protocols that minimize cultural and linguistic artifacts capable of compromising score interpretations.</i>		

Analytical Rationale: Test administrators must anticipate situational factors that threaten test validity, drawing upon cross cultural testing experience to ensure examinees clearly comprehend test instructions without unintended administrative confusion.

Administration	Guideline	2
<i>Guideline A 2: Standardize physical testing environments while articulating legitimate accommodations for specific test taker needs.</i>		

Analytical Rationale: Standardized environmental conditions, including lighting, acoustic control, and pacing, must be maintained across testing sessions. Simultaneously, legitimate accommodations, such as extended time or enlarged print for individuals with disabilities, must be documented to contextualize score interpretations without inflating candidate performance.

4.5 Scoring and Interpretation Guidelines

These guidelines govern cross group score comparability and equitable diagnostic interpretation:

Scoring	and	Interpretation	Guideline	1
<i>Guideline SI 1: Interpret cross group score differences with full consideration of contextual cultural and linguistic factors.</i>				

Analytical Rationale: Observed score discrepancies between demographic cohorts may reflect cultural, educational, or socioeconomic variables rather than genuine disparities in the underlying psychological trait. Sireci, 2005, highlighted protocols for evaluating bilingual cohorts while controlling for confounding demographic variables.

Scoring	and	Interpretation	Guideline	2
<i>Guideline SI 2: Restrict cross group score comparisons until measurement invariance has been rigorously established across scales.</i>				

Analytical Rationale: Direct score comparisons across cultural groups are psychometrically invalid unless measurement invariance is confirmed across configural, metric, and scalar tiers, ensuring scales are free from construct, method, and item bias.

4.6 Technical Documentation Guidelines

Comprehensive technical documentation ensures transparency, scientific replicability, and ethical practitioner usage:

Documentation	Guideline	1
	<i>Guideline DOC 1: Provide comprehensive technical manuals detailing all procedural adaptations and empirical equivalence evidence.</i>	

Analytical Rationale: Large scale cross cultural programs, such as the Trends in International Mathematics and Science Study and the Programme for International Student Assessment, exemplify rigorous documentation. Technical manuals must detail translation methodologies, sample demographics, and invariance proofs to enable scientific replication.

Documentation	Guideline	2
	<i>Guideline DOC 2: Formulate operational user guidelines that guide ethical and competent test administration by practitioners.</i>	

Analytical Rationale: Operational documentation must provide clear guidance for applied practitioners, ensuring adherence to the ethical standards codified in the International Test Commission Guidelines on Test Use, 2017.

5. Conclusion and Strategic Recommendations

This analytical examination demonstrates that translating and adapting psychological and educational tests is a complex, multidisciplinary undertaking governed by rigorous psychometric standards. Adapting an instrument across cultural boundaries requires far more than linguistic translation; it demands establishing structural, conceptual, and measurement equivalence to ensure equitable assessment across diverse populations.

The International Test Commission Guidelines, Second Edition, provide an exhaustive methodological architecture that guides researchers through every stage of cross cultural assessment. Adhering to these eighteen standards prevents cultural bias, preserves construct integrity, and guarantees sound diagnostic decisions.

5.1 Applied Recommendations for Arab and Algerian Psychometrics

- Institutionalizing Methodological Training: Integrating the International Test Commission guidelines into graduate psychology curricula to elevate psychometric competencies.
- Mandating Multidisciplinary Adaptation Teams: Establishing adaptation panels uniting bilingual linguists, psychometric specialists, and cultural experts.
- Systematic Reporting of Invariance Evidence: Academic journals should require empirical proofs of measurement invariance, factor structure stability, and Differential Item Functioning analysis prior to publishing studies utilizing translated instruments.
- Respecting Intellectual Property: Ensuring researchers secure formal copyright permissions from original authors and publishers before undertaking adaptations.

- Constructing Indigenous Psychometric Instruments: Encouraging the development of native diagnostic tools grounded in local sociocultural realities, reducing reliance on foreign imports.

References

- Abbas, A. R., 2017. Cultural bias of psychological tests of Western origin in the Algerian environment. *Al-Siraj Journal in Education and Community Issues*, Issue 1, pages 130 to 148.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, District of Columbia.
- Anastasi, A., 1976. *Psychological Testing*, 4th edition. Macmillan Publishing, New York.
- Angoff, W. H., and Modu, C. C., 1973. Paired comparison analysis of item difficulty across cultural groups. *Educational and Psychological Measurement*, Volume 33, Issue 2, pages 345 to 358.
- Boukraa, I., 2020. Constructing and adapting psychological and educational scales: Basic concepts and methodological steps. *Journal of Human Sciences*, University of Constantine 2, Volume 31, Issue 2, pages 177 to 190.
- Bousalem, A. A., 2015. Psychological tests applied in Algerian academic studies and the necessity of adaptation for validity. *Academic Journal of Social and Human Studies*, Issue 14, pages 20 to 26.
- Brislin, R. W., 1986. The wording and translation of research instruments. In W. J. Lonner and J. W. Berry, Editors, *Field Methods in Cross Cultural Research*, pages 137 to 164. Sage Publications, Beverly Hills, California.
- Cook, L. L., and Schmitt Cascallar, M. A., 2005. Establishing score comparability for tests administered in different languages. In R. K. Hambleton, P. F. Merenda, and C. D. Spielberger, Editors, *Adapting Educational and Psychological Tests for Cross Cultural Assessment*, pages 167 to 194. Lawrence Erlbaum Associates, Mahwah, New Jersey.
- Hambleton, R. K., 1999. Issues, designs, and methods for adapting tests into multiple languages and cultures. In R. K. Hambleton and P. F. Merenda, Editors, *Advances in Test Adaptation*, pages 1 to 32. Lawrence Erlbaum Associates, Mahwah, New Jersey.
- Hambleton, R. K., Merenda, P. F., and Spielberger, C. D., Editors, 2006. *Adapting Educational and Psychological Tests for Cross Cultural Assessment*, translated into Arabic by H. Barmada, 1st edition. Obeikan Bookstore, Riyadh, Saudi Arabia.

- Hambleton, R. K., Yu, J., and Slater, S. C., 1999. Field test of the ITC guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment*, Volume 15, Issue 3, pages 270 to 276.
- Hambleton, R. K., and Patsula, L., 1999. Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, Volume 1, Issue 1, pages 1 to 30.
- International Test Commission, 2017. *The ITC Guidelines for Translating and Adapting Tests*, 2nd edition. International Test Commission, www.intestcom.org.
- Kaufman, A. L., 2000. *Tests of Intelligence*. Cambridge University Press, Cambridge, United Kingdom.
- Sireci, S. G., 2005. Using bilinguals to evaluate test score comparability. In R. K. Hambleton, P. F. Merenda, and C. D. Spielberger, Editors, *Adapting Educational and Psychological Tests for Cross Cultural Assessment*, pages 117 to 138. Lawrence Erlbaum Associates, Mahwah, New Jersey.
- Van de Vijver, F. J. R., and Tanzer, N. K., 1997. Bias and equivalence in cross cultural assessment: An overview. *European Review of Applied Psychology*, Volume 47, Issue 4, pages 263 to 279.