Test Adaptation in a Shrinking World

A Presentation to
Shue Yan University, Hong Kong
Kurt F. Geisinger, Ph.D.

INTRODUCTION: 1/3 - Overview

The nature of this presentation

The main topics, which I hope to relate

1. Test Adaptation—a not-so-quick overview
2. Historical Perspectives
3. Some Examples

INTRODUCTION: 2/3 Why adapt?

- Globalization – the shrinking world
  - International comparisons / standardization
  - Shared issues and differentiated resources
- Psychological science is getting stronger
  - Our theories have now (sometimes) included culture
  - Differences between etic (culture-free/universal) and emic (culture-related)
- Fiscal and pragmatic reasons why we adapt tests (including multinational corporations)

INTRODUCTION: 3/3 Issues to Consider

What issues need to be considered with regard to the adaptation of an instrument to different countries, cultures, and people?

- **Different languages**: Most cross-cultural adaptations of assessment instruments involve the translation of an instrument from one language into another.
- **New target population/Cultural Differences:** The new target population differs appreciably from the original population with which the assessment device is used/normed, in terms of culture or cultural background, country, and language.
- **Same language, but different cultures**: In some instances, however, adaptations of assessment instruments are needed even when the language remains the same, because the culture or life experiences of those speaking the same language differ.

Psychological Assessment 1994

Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments.

Guidelines from the International Test Commission (ITC)

- ITC Guidelines for Translating and Adapting Tests - 15th July, 2005, Version 1.0 Final Version
- ITC Guidelines for Translating and Adapting Tests (2nd Edition) (2017)

- Both are freely available at: *https://www.intestcom.org/page/16*

ITC Test Adaptation Guidelines

- Context
- Test development and adaptation
- Administration
- Documentation / Score interpretation
- DIF studies

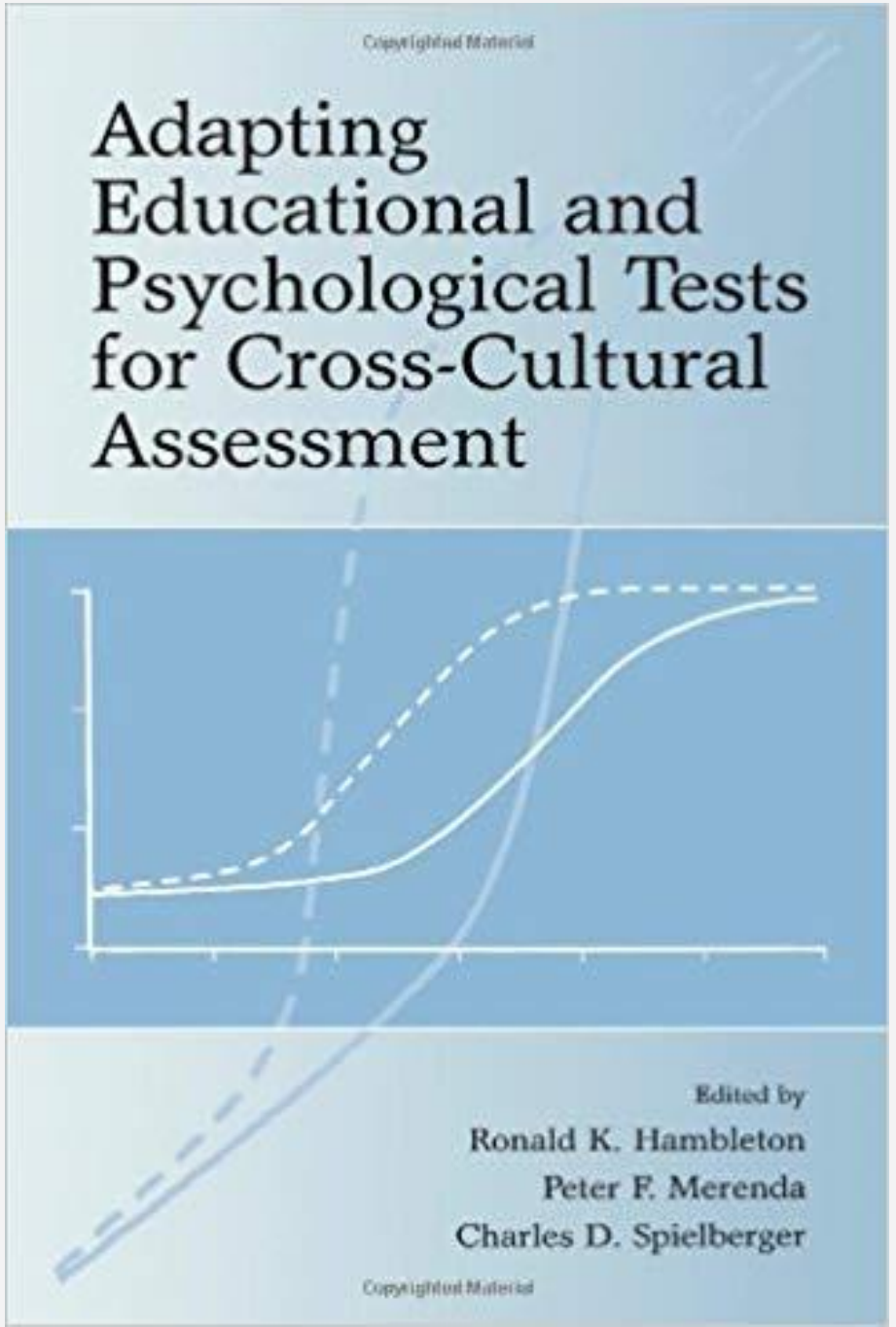
ITC – International Guidelines on Test Adaptation (2nd edition recent)

[https://www.intestcom.org/page/16]

- **Intended audience:** psychological associations and organizations associated with testing.
- **Impetus:**
  - Promote good practice in test adaptation
  - Assure uniformity in the quality of tests being adapted for use in other cultures and languages
  - More national resources are invested in educational testing, especially in emerging and developing nations
  - Globalization of industry and internet accessibility have proliferated testing and selection practices that were previously associated with western cultures (North America, Australia, and Europe)

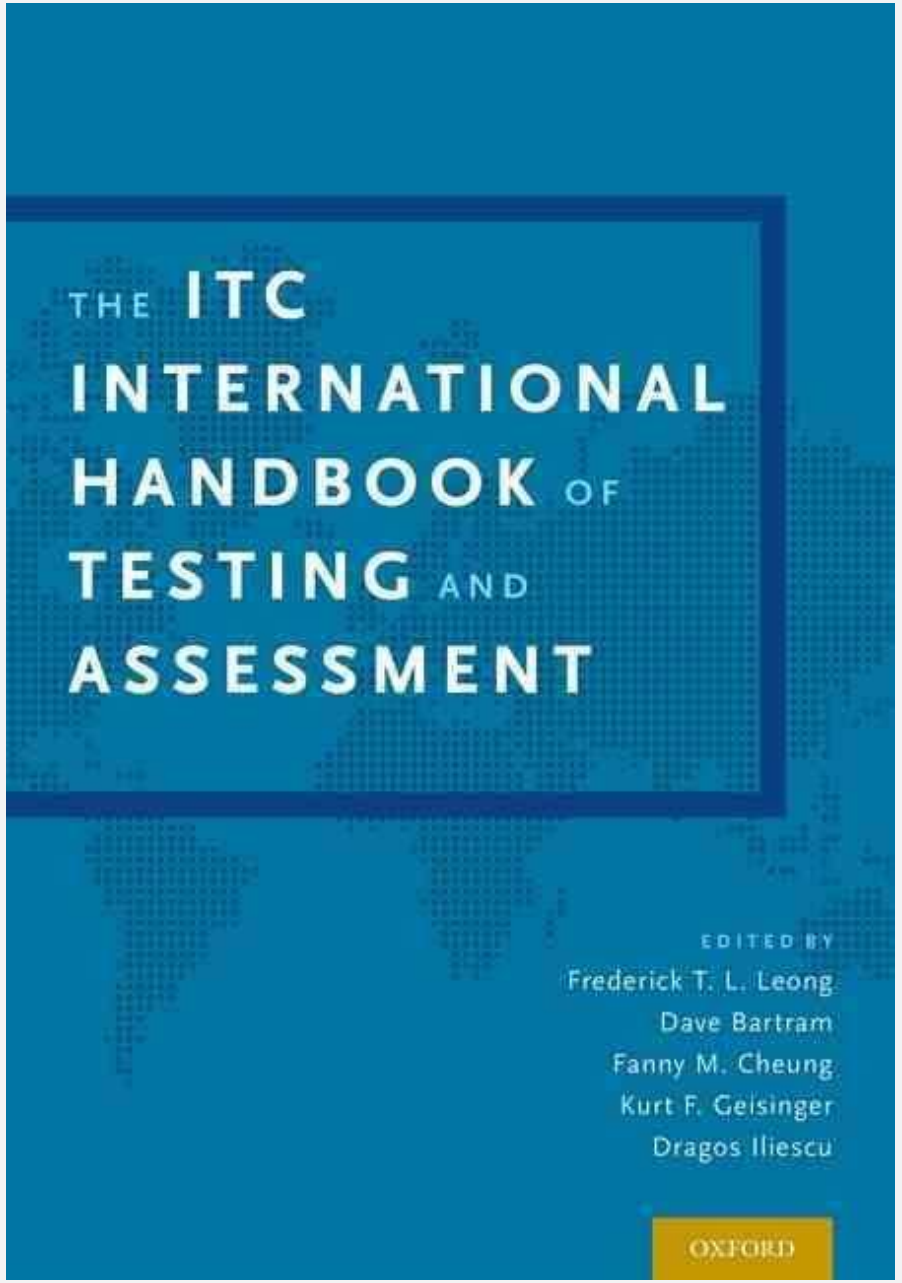
A Book Devoted to Test Adaptation

- An edited book (2005)
- Covers all aspects of educational and psychological testing
- Edited by Ron Hambleton, Peter Merenda, and Charles Spielberger



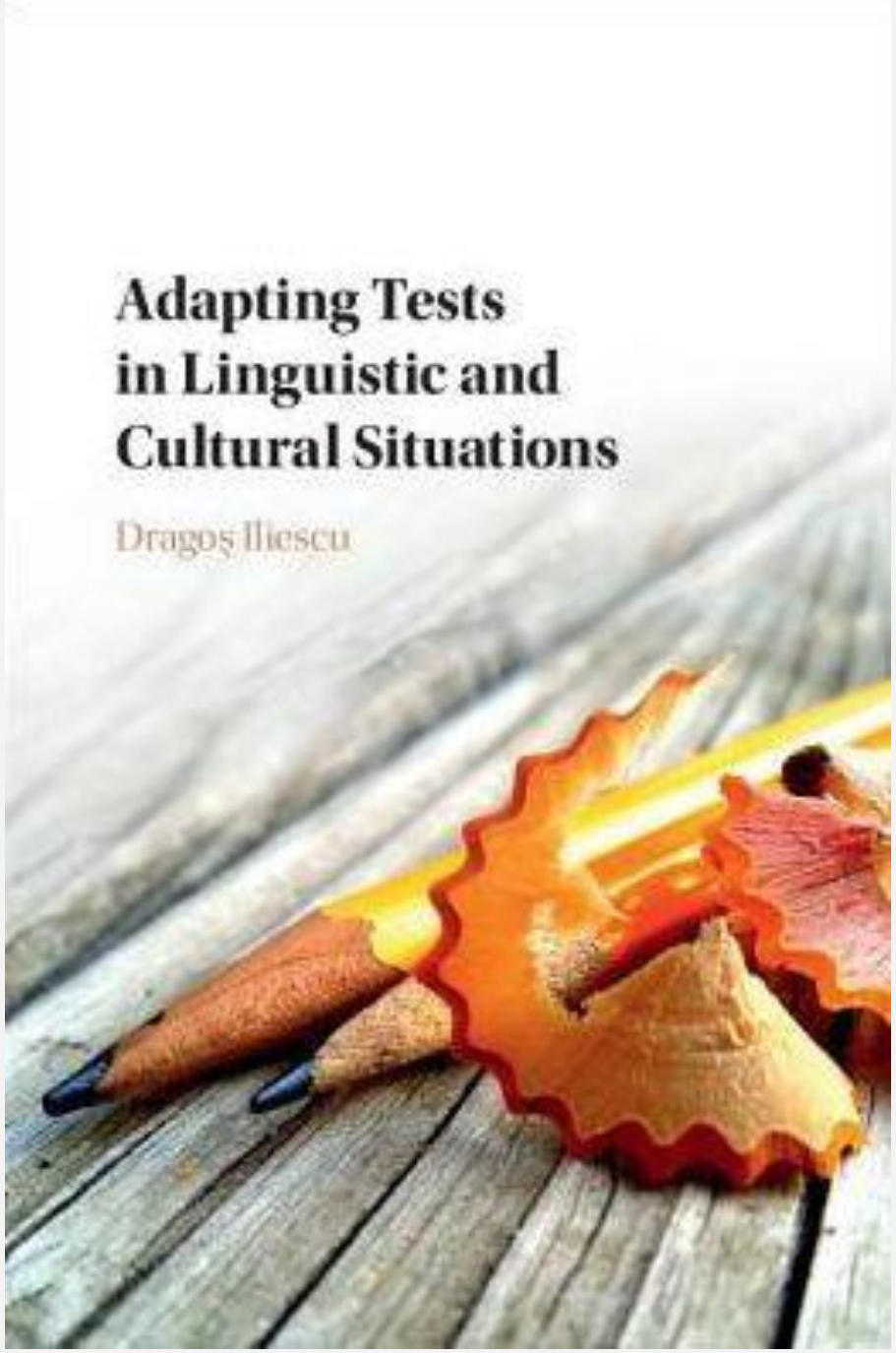
The ITC Handbook

- Chapter on "Test Adaptation" by Fons J. R. Van de Vijver
- Chapter on the ITC Guidelines and Standards by David Bartram and Ronald K. Hambleton



New Book by Dragos Iliescu (2017)

- This entire book, published by Cambridge University Press.
- One of the first book published as part of an ITC book series



Two Precursors

1. If a published measure → get permission from the copyright holder (see recent *Science* article) [http://www.sciencemag.org/news/2017/09/pay-or-retract-survey-creators-demands-money-rile-some-health-researchers]
2. If measure is not published or copyrighted → professional courtesy still suggests informing the original author

One must hypothesize whether the construct assessed by the measure is either an **etic** one or an **emic** that will transfer to this new culture (e.g., the cultures are similar)

Adopting a Test

- Intent to use a test (materials, protocols) in its current state and for its intended purpose
- Implications:
  - Research
  - Training / Certification
  - Commitment decisions (purpose, admin, use)
  - Psychometric properties
  - Validity
- What if the test is in a different language?

Research from Professors Elosua & Illiescu

- Identified the 10 Most Commonly Used Psychological Measures in Europe
- 8 Come from the United States and have been adapted
  - Examples: Wechsler Tests of Intelligence, MMPI, 16 PF, NEO-PI
- 2 That Emerged from Europe
  - Raven's Progressive Matrices
  - Rorschach

Why adapt an existing measure?

The reasons to support whether to adapt or not to adapt an assessment are numerous and very context dependent.

PROS	CONS
Established/recognized in other settings (comparisons)	Copyright issues & country membership requirements
It is more cost effective (time & resources) to adapt than to build	Will benefits justify efforts? Is there real need to assess/participate?
Globalization necessitates cross-culturally appropriate measures to fulfill the need to compare, evaluate, select, treat, etc.	Fairness and validity of scores for target population and use (e.g. must be normed for target demographic) issues are involved
Guidelines and best practice research offer more options to test users to make informed decisions to reduce negative outcomes.	Translated assessments, even with adaptation, still introduce additional negative psychometric and cross-cultural issues.

Why adapt an existing measure?

May wish to make international, inter-cultural, or cross-lingual comparisons

- Note the various OECD comparisons of science, mathematics, critical thinking, engineering, and economics
- The Mexican model for indigenous peoples
- In South Africa, 11 official languages
- We live in a world with much immigration and diversity

Why NOT adapt an existing measure?

- There is reason to believe that the construct differs across cultures
  - Example: ***the Big 5*** in Western cultures is ***the Big 6*** in Eastern ones (Cheung)
- The original mode of measurement is not existent, familiar or common in the target country
  - Multiple-choice testing is very common in the United States but much less so in England
- Emic (within a culture) vs. Etic (across cultures) measures/characteristics

Van de Vijver's Three Equivalence Foci of Test Adaptation

There are both historical and methodological stages:

- Initially, test translation focused on **Language equivalence**
- **Cultural equivalence** became an issue that also needed consideration
- **Psychometric equivalence** is the latest approach and the final one engaged in during the test adaptation process

Test Adaptation

- Current professional term of preference is adaptation rather than translation, because cultural, geographic, and other considerations affect the content of a measure.
- Studies show that simply translating the words of any measure rarely works.
- Adaptation is a process of transforming an existing measure/test from one language to another with attention to cultural nuances and maintaining psychometric properties so that the scores are comparable and valid for the intended interpretations.

Types of Translation/Adaptation Processes

Note: translation only refers to language

- Simple translation/literal
- Adaptation or translation with checks
  - Back translation is most common
  - Can also ask experts to rate the quality of the adaptation
- Committee approach where necessary skills are shared
- Multiple Concurrent Development of Assessments (not really an adaptation process) as used by OECD—necessarily by committee

Skills Needed to Adapt a Measure

- Fluency in both languages
- Comprehensive understanding of the construct being assessed
- Thorough understanding of both cultures
- Ability to work on tests and measures—e.g., item writing and so on (writing instructions)

Translation (adaptation inherent) Goals

The goal is to maintain:

- Item difficulty within reason (Intelligence, educational test items)
- Content relevance and access (Caribbean SAT Example)
- Construct relevance and validity
- Formatting: appearance and comparable tasks

Concurrent Assessment Model

- This model is an alternative option to traditional adaptation methods
- It is appropriate at the design and development stages of an assessment, but NOT for pre-existing measures.

Why Back Translation is Problematic

- If translators are evaluated by the quality of the back translation
- Then they choose wording that leads to the best back translation
- Rather than the one that is most appropriate

An Example of a Back Translated Item

- An analogy item where the stem was:

  "Out of sight::Out of mind:

- The item was translated and came back:

  Blind::Insane

An Example Item Difficult to Translate

- Fanny Cheung's MMPI item (for depression)

  "I feel blue." (True/False)

- One really must avoid colloquial expressions in items, especially if they are being taken to other cultures

"International" Adaptation 1/3

Ercikan,1998 pp 543-553

- ***Quality Matters*:** The quality of an adaptation has direct influence and impact on the comparability and validity of scores
- ***Systematic Stability*:** When an instrument is adapted well, the psychometric properties of the instrument should not change significantly even if the populations are significantly different.

"International" Adaptation 2/3

- ***Personnel***: There are benefits/advantages associated with the inclusion of bilingual individuals in the item writing process (Canada example)
- ***Evidence:*** Significant differences between forms can be attributed to poor translation
  - When word choice/terms are not of similar profile or prevalence (level of difficulty or commonness)
  - When syntax or expression results in unequal length or complexity of sentences
  - Differential contextual meaning of vocabulary

"International" Adaptation 3/3

- ***Prevention/Oversight:*** Items should be screened to avoid/minimize poorly adapted items (systematic shortcomings)

- ***DIF*** is a marked threat to validity
  - Examination of item statistics and differential item functioning (DIF) methodologies can be used to detect/identify poorly translated items.
  - If DIF is detected and cause is associated with the adaptation process due to culture or curricular issues, it means that the item is not measuring what was intended to measure.

The Critical Role of Culture

- How does culture influence adaptation? (unseen relative to language)
- How does this affect using the same tests in different countries that use the same language?
- Will have some examples later

10 Steps in Translating/Adapting a Measure

1. Translate and adapt measure
2. Review the translated measure
3. Revise the measure based upon comments or suggestions provided in the review
4. Pilot the instrument
5. Field test the instrument
6. Standardize the scores

10 Steps in Translating/Adapting a Measure

(continued)

7. Perform validation research as appropriate
8. Develop a manual and other documents for users of the assessment
9. Train users
10. Collect reactions from users

- Ercikan & Lyons-Thomas have another set of steps, as do some others

More Steps from Others

- Hambleton & Patsula
  - Hire appropriate translators (may need to hire others as well)
  - Insure construct equivalence (a first step)
  - Decide whether to adapt or build anew
  - Link scores across (if comparisons are appropriate)

- My next question – *"What about scoring issues?"*

Documentation (Test Manuals)

- A test is adapted to a new culture and country
- What kinds of documentation is now needed?

A manual, a website, articles,…

Document the Adaptation Process

- Standard 3.12 of the *Standards for Educational and Psychological Testing* (2014):

"When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretation for intended use."

Pre-Use Research (Qualitative)

- Reviews of the assessment for usability
- Reviews of the instrument for comparability
- Pre-tests with relevant individuals
- Timing of taking the measure
- Suitability of instructions (some cultural differences in speed)
- Questions about appropriateness

What kind of research is needed after adaptation?

- **Preliminary**
  - Reliability (internal consistency; test-retest; alternate forms, if possible)
  - Item Analysis
  - Factor analyses of items(?)
- **Secondary**
  - Validation of scores and inferences (often SEM)
  - Fairness analysis…dif—not always?
  - Norms (generally need to be separate)
  - Linking, possibly

Beyond Validity—Usefulness

- How does this issue differ from construct equivalence? (Example of Canadian SAT)

- Is there any utility to using the measure in the target culture?

Defining Test Translation Error

Translation Error is *"the lack of equivalence between the source language version and the target language version of test items."* [1] (Solano-Flores, et.al., 2009, p80)

- Due to the nature of languages, it is possible that an adapted form of an assessment does not capture or transfer nuances.
- Psychometric error consequence: the adapted version potentially tests different constructs from the original form.

How to Establish Measurement Equivalence

Equivalence of measurement includes:

(a) equivalence of constructs,

(b) equivalence of tests, and

(c) equivalence of testing conditions

Psychometric Consequences

- Grisay (2003) noted that adaptation errors are the most prevalent source of DIF in International Educational Assessments:
  1. Curricular Differences
  2. Cultural Biases
  3. Translation Errors

- Sireci and Swaminathan have questioned the use of DIF procedures for adapted items

Fairness Analyses

- Underserved groups often vary by cultures
- Not sure it makes sense to compare the same racial, ethnic, or religious groups.
- Gender group fairness comparisons may make more sense

Adaptation Issues

The AHELO Project

- Translating a ***Collegiate Learning Assessment*** (CLA)
- What is the CLA? (explain); used at more than 1,000 college and universities as part of outcomes assessment
- Buros was hired to perform adaptations to South Korea, Slovakia, Egypt, and Columbia
- We also reviewed results from some other countries (Kuwait…)

The CLA is an "Essay Performance Assessment"

- Three-four pages on a problem
  - Want to generate water power, two lakes connected by a river
  - An endangered fish is endangered further by the water power facility
  - Motive is profit generation

- Students must write an essay describing what they would do as a consultant to the power company

The AHELO Adaptation Process

- We worked with teams of educators in different countries
- We needed to approve the translation/ adaptation and review data on its pre-testing
- We needed to evaluate the translation/adaptation to the extent possible

Adaptation Issues: Slovakia

- Few problems
- Reasonably easy translation/adaptation
- They are a quasi-Western nation; a NATO country
- Recent country that is a split of Czechoslovakia
- They have rivers and hydroelectric power

Adaptation Issues: Kuwait

- The year before Buros started, the CLA was translated into Arabic for Kuwait
- There are no rivers and no lakes in Kuwait
- Water power is unknown by students in Kuwait
- The item was changed to Ocean power and a sea going fish that was endangered

Adaptation Issues: Columbia

- My GRE experience in trying to build a Spanish version of the GRE: we would need at least three versions
- OECD/CAE (Organization for Economic Cooperation and Development/Council for Aid to Education) had already adapted a test in Spanish for Mexico
- Needed a different version for Columbia

Adaptation Issues: South Korea 1/2

- Pre-test Statistics were awful
- *Yet*…the Korean Team was pleased with the translation result
  - I had a graduate student from Korea look it over
  - She told me that it did not make any sense; Korea does not have power companies; the government owns the power companies and there is no profit motive

Adaptation Issues: South Korea (2/2)

- Re-wrote the scenario to focus on the power being supplied by the government
- The test taker/writer is to be a consultant to the government rather than a power company
- A few other more minor issues were changed and the results worked out well

Adaptation Issues: Egypt 1/3

- I ran the meetings in Egypt in the middle of the Egyptian revolution (a few asides)
- We had the revised Arabic measure for Egypt
- There are differences between Arabic in Kuwait (Gulf Arabic) and Egypt (Modern Standard Arabic (MSA))

Adaptation Issues: Egypt 2/3

- They knew rivers and lakes: near to the Nile
- We observed the testing of two individuals who described their testing experience
- Big problem…
  - No governmental agency in Egypt would ever hire a consultant
  - Instead, they believe the government believes that *it knows the solutions*

Adaptation Issues: Egypt 3/3

- They believed that they knew a solution
- Their solution was to make the situation work for them
- Instead they had students being hired by a US power company as a consultant

Our Experience Looking at Adapted Measures

- Many measures presume that the original scale's validation research in the original language and culture generalize to the new culture
- Some scales even apply the original norms to the target measure
- Where norms are provided for the new culture, they are often much smaller and less representative samples

- **Structural & Linguistic balance across forms** – to minimize differences in item difficulty, text length and psychometric quality (DIF, Fit, Discrimination).

Lessons Learned from OECD/PISA

- **National Expert Committee-** to review the content & appropriateness of translated material considering specific nuances to their country's culture(s), social structures, traditions, and geography.
- **Double Translation with extensive cross-checks:** Two independent translations of the same form to the target language with extensive cross-checks with another form.
- **Double Translation & Reconciliation-** while paying attention to cultural and semantic nuances, two independent translations (each from a different language base to the target language) and a third party reconciles the translations.

Fidelity lessons from AHELO

Issues found to impact fidelity of translations on a study completed on the Assessment of Higher Education Learning Outcomes (AHELO) international assessment:

- **Variance** in protocols
- **Cultures**
- **Training** - Specially trained translators vs bilingual academics
- **Economic climate** (funding compensation exchange)
- **Educational context** (systems, demographics, logistics)
- **Schema** - Inherent structural and stylistic differences

Other lessons from AHELO

- **Context matters**. Institutions of higher learning have different admissions criteria, requirements, and missions – it is possible that outcome differences are reflective of selection and input.
- The same issues must be considered with other international assessments due to heterogeneous educational systems, socio-economic factors, cultures, perceptions, linguistic structures and styles, etc.

Perhaps My Theme

- Adapted measures have huge appeal and potential
- We need to conduct more and better research on adapted measures
- They need to be thoroughly reviewed.

Issues in International Test Adaptation

1. The nature and complexity of the process of adapting a measure from one language to another.
2. Basic processes that help insure the cultural validity of a measure that is taken from one language and culture to another.
3. Basic psychometric concerns to ensure that an adapted measure is valid in the target language and culture.
4. Cautions in the use of translated measures, especially with immigrant populations or with international clients

Thank You!

Kurt F. Geisinger, Ph.D.

Meierhenry Distinguished University Professor and Director, Buros Center for Testing
Editor, *Applied Measurement in Education* (2006-Present)
21 Teachers College Hall
The University of Nebraska-Lincoln
Lincoln, NE 68588-0353