This is to certify that the
dissertation entitled

COMPARABILITY OF MULTILINGUAL ASSESSMENTS:
AN EXTENSION OF META-ANALYTIC METHODOLOGY TO
INSTRUMENT VALIDATION

presented by

KEVIN B. JOLDERSMA

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Counseling, Educational Psychology and
Special Education

Mark W. Roehrle
Major Professor's Signature

May 25, 2006
Date

MSU is an Affirmative Action/Equal Opportunity Institution

COMPARABILITY OF MULTILINGUAL ASSESSMENTS: AN EXTENSION OF META-ANALYTIC METHODOLOGY TO INSTRUMENT VALIDATION

By

Kevin B. Joldersma

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Counseling, Education, Psychology and Special Education

2006

ABSTRACT

COMPARABILITY OF MULTILINGUAL ASSESSMENTS:
AN EXTENSION OF META-ANALYTIC METHODOLOGY TO INSTRUMENT VALIDATION

By

Kevin B. Joldersma

The translation and adaptation of multilingual instruments is of ever increasing importance due to the use of international high stakes assessments. Educational policy is shaped by the findings of these instruments. Test developers of these multilingual assessments have traditionally relied upon expert-dependent or psychometric methods to create comparable instruments across languages or cultures. However, expert-dependent methods are subjective in nature and while psychometric tests remove subjectivity, they also remove the valuable insights of experts that account for the multi-faceted problem of multilingual instrument comparability. This dissertation seeks to create a parsimonious situation where the validity of a multilingual instrument's inferences can be stabilized across its language versions. This will be done, in part, by assessing the efficacy of meta-analysis as a means for synthesizing expert-dependent and psychometric findings in order to improve the comparability of multilingual assessments.

KEYWORDS: Test translation, Meta-analysis, Differential Item Functioning

DEDICATION:

To K, O, H and my extended family. Thanks for your support through the years and making this possible.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: MULTILINGUAL INSTRUMENTS AND THEIR USES

The issues surrounding test translation or adaptation are of great interest to a growing number of test designers, users and policy makers. Policy makers seek tests that can evaluate the same or a similar construct in a growing number of multilingual contexts ranging from international assessments such as PISA and TIMSS[1] to the Texas Assessment of Academic Skills (TASS)[2] to small-scale university and school level placement of Limited English Proficiency (LEP) students. Test-users and administrators, thus seek tests that can accomplish these multilingual measurement mandates.

The problem is two-fold. Test-users and administrators may be subject to the following: a) dissatisfaction with the currently available instruments or b) ignorance of the measurement difficulties or possibly invalid inference made from instruments that have not been adapted appropriately for a multilingual environment. Hence, the task of test-developers is also two-fold: a) to educate or inform test users and policy makers of the problems of multilingual instruments and b) to better develop and evaluate tests with multilingual mandates (as per the APA, AERA & NCME Standards, 1999).

The purposes of this dissertation are as follows: 1) to review current usage and techniques of translation and adaptation of multilingual instruments, 2) to evaluate the relative success of a selected number of these techniques, and 3) to provide recommendations and possibly new techniques to aid test designers with the unique problems of multilingual instruments. As a consequence, it may be possible to create

[1] PISA (Program for International Student Assessment) is a new system of international assessments that focus on 15-year-olds' capabilities in reading literacy, mathematics literacy, and science literacy. It is organized by the Organization for Economic Cooperation and Development (OECD), an intergovernmental organization of industrialized countries, and was administered for the first time in 2000 with 32 countries participated (PISA, 2000). TIMSS (Trends in International Math and Science Study) tracks math and science internationally (TIMSS, 1999).

more effective multilingual instruments, from which, more comparable and valid inferences may be made.

Current methods need to be evaluated, and this will be done using their identified strengths and weaknesses as a starting point. Following, the methods will be evaluated in two manners; both using recognized expert judgment protocols as well as using established statistical indices to evaluate their efficacy. For consistency, any new methods will be evaluated similarly to verify the similarity of their results. However, the most important part of the project is the analysis of the comparability of the tests in their language versions.

Definitions and comparability

Comparability of multilingual instruments does not necessitate the one to one relation of ideas or testing concepts. Rather, it is quality and relative equality of the measures associated with the instruments after translation or adaptation. *Adaptation* is the cultural or linguistic adjustment of a test to fit the culture or language in which the instrument will be delivered. Hence, comparability is much more than a translation exercise, wherein a construct is tested the same way in the instrument's *target languages* (i.e., the languages in which a multilingual instrument is to be delivered). Comparability of measures is certainly of vital importance to a multilingual instrument. Without good comparability, the inferences made from the delivery of these instruments are likely to be invalidated.

There are many methods that have been employed by test developers to help multilingual instruments have comparable measurement. These methods can mostly be

[2] The TASS is a test of academic skills (math, reading, science, etc.) which uses Spanish and other minority language tests from grade 3 through 5, but phases out minority language testing by grade 6.

categorized as either expert-dependent methods or psychometric methods. *Expert dependent methods* involve a carefully trained individual with experience in both the source and target language or languages for which a multilingual instrument is designed. *Psychometric methods* rely on statistical procedures used to determine whether the performance of groups in a multilingual instrument differs greatly from the expected.

The comparability of a multilingual instrument's measures is dependent upon a professional's expertise with expert dependent methods. A professional uses specialized skills to attempt to enhance the validity of the instrument's measures. Methods employed by experts include evaluative methods or creative efforts to better comparability. Methods used include those with little quality control, such as direct translation, to methods whose quality is aided by back translation[3] or using expert test-takers[4].

Certainly, there are advantages and disadvantages of expert dependent methods. An important advantage of these expert dependent methods is that they allow both comparability and adaptability of tests and instruments. Thus, an instrument can be adapted appropriately to its target language or culture, and theoretically its measures will maintain validity cross-culturally. However, by their very nature, these expert dependent methods are subjective in nature as they rely on the judgment of those experts.

In another sense, comparability of an instrument must also be thought of in a statistical manner. Given two populations, who are of theoretically normal or matched

---

[3] *Back-translation* is a method that helps to verify proper translation by translating from the original language to the target language and returning to the original language. This is done to see if the content of the back-translated material matches the original concept or the "original intent" of its authors.

[4] The *expert test taker strategy* for CLIs is primarily employed to evaluate the test translation rather than aid the test creation process. Knowledgeable test takers, sometimes monolingual in the target language and sometimes bilingual, are given the translated form of the test and asked to spot-check it for language appropriateness and content similarity.

abilities, one should expect comparable results if the test has been adequately adapted. Hence, another indicator of non-comparability is statistically differing outcomes.

Two major statistical analyses for detecting these differences are differential item functioning (known as DIF) and an analysis of the dimensionality of the multilingual instrument. Dimensionality analysis may give us a broader understanding of what is happening on the language versions of a multilingual assessment. However, current research shows that it is essential to evaluate at the item level, as well (Zumbo, 2003).

DIF flags items where the outcomes of examinees of purportedly similar abilities, but from different backgrounds (e.g., different gender or language background), are not the same. If an item or items are flagged by the DIF procedure, the item is sometimes removed or its substance is further examined for potential bias. Hence, the use of a statistical technique still often requires the use of an expert's judgment to identify if the difference in performance on item/s is the desired result of measurement, or an item/s that need to be revised to ensure comparability.

Another statistical method that is often employed is a dimensionality analysis. Dimensionality analyses are used to detect *multidimensionality*, which is the presence of multiple dimensions of measurement. These multiple dimensions assess different latent traits or constructs, which can at least partially be the consequence of language differences. If the dimensional structures of two or more languages are different, this is further evidence that the instrument is measuring something different between language versions. In such cases, the multiple versions of the multilingual instrument would not produce comparable information about examinees. Though statistical indices can be used to show differences between the dimensional structure of language versions, a

dimensionality analysis is only a first step because it, too, requires the judgment of an expert as to whether the different dimensions tested in the language versions are desirable or not and, more importantly, whether they impact the comparability of the instrument's language versions. Consequently, though the great advantage of a statistical index is to remove the subjectivity of judgment inherent in expert dependent methods, statistical methods also rely on experts for their interpretation.

In sum, the objective psychometric methods which lack explanatory power must be complemented by subjective and nuanced expert-dependent methods. The more important matter, however, is that neither expert dependent methods nor psychometric methods sufficiently address Messick's (1989) call for theory driven work. In this seminal work, Messick calls upon psychometricians and test developers to provide theoretical rationales for their work, rather than merely create cookbooks for better instruments.

Current Thought on Instrument Comparability

The current perspectives of instrument comparability are reflected by the adoption of the International Testing Commission's (ITC) Test Translation and Adaptation Guidelines. These guidelines were approved over ten years ago, and over the years numerous researchers have examined them to comment on their strengths and weaknesses. In Hambleton's work (2001), he takes a look at where the theory of comparability in translated/adapted tests has gone in the first seven years since the ITC publication.

The original ITC guidelines consisted of four broad categories made up of 22 individual guidelines. The four categories are as follows: 1) instrument context, 2) test

development and adaptation, 3) test administration, and 4) score interpretation and documentation (the full ITC guidelines can be found in Appendix 1). Instrument context refers to the concerns that test developers should take into account regarding construct equivalence between the linguistic or cultural groups being tested. Test development and adaptation guidelines involve choosing translators with suitable qualifications and picking the appropriate statistical method for analyzing score equivalence or comparability. Test administration guidelines are suggested for how best to administer a test or instrument to multiple languages or cultural groups. This includes taking into account item formats, time allotted, etc. that should be handled differently based on cultural expectations or needs. Lastly, score interpretation and documentation contains guidelines for providing evidence for the validity of the adaptation (Hambleton, 2001). As a consequence of several investigations and practical try-outs, Hambleton (2001) summarizes many of the initial changes that have been suggested by researchers.

With respect to the *context* of translated/adapted tests, Hambleton reports on a Tanzer and Sim study (1999), which recommends that the guidelines be expanded to incorporate the acknowledgement of linguistic differences. This expansion allows for the application to cross-cultural studies. Furthermore, it should be argued that a natural extension of the Tanzer and Sim (1999) argument would be to a linguistic analysis of differences arising in translated/adaptive test versions. In effect, the questions raised by this area of the ITC guidelines are: Is a construct being understood the same way by all linguistic and cultural groups? Is there any overlap of *definitions* of the construct in language/culture groups and the overlap in the *actual manifestation* of the construct in the language/cultural groups (Tanzer & Sim, 1999)?

Test development and adaptation procedures adopted by the ITC have faced scrutiny through the years, and Hambleton gives us some examples of where the guidelines in this category have needed revision or expansion. First, more stringent ideas of who is an expert and who is most capable of performing an adaptation have been recommended by several studies (Grisay, 1998, 1999; Jeanre & Bertrand, 1999). The recommendations include the assurance that experts be able to function professionally in both source and target languages, defined as *functionally bilingual.* Moreover, Jeanre and Bertrand (1999) suggest advancing the ITC guidelines to a formalization of the D.2 and D.5 guidelines, which refer to the documentation of the compilation of the adaptation process (see appendix 1 for full text). Jeanre and Bertrand (1999) additionally offer guidelines and rating scales for enhancing multilingual assessment comparability by using a simple linguistic rating scale. The aim of these rating scales is to target and better the validity of the inferences made from the adapted/translated instruments. As a consequence, Hambleton writes that test developers need to use systematic and judgmental evidence (including both linguistic and psychological examination) to aid comparability and provide linguistic/cultural validity for test users' inferences.

The third category of the ITC guidelines details the Commission's perspectives on the *administration* of translated/adapted instruments. The aim of these guidelines is to keep the testing environment as similar as possible, to create culturally appropriate materials and translated instructions, and to consider the effects of test-wiseness (or unfamiliarity of item types) which can lead to speededness problems. Since these concerns overlap greatly with most existing standards, such as the *AERA, APA and*

NCME Standards (1999), Hambleton suggests that this section should be eliminated as unnecessary and not unique to ITC concerns.

The last section of the ITC guidelines refers to the *interpretation of scores* and the appropriate *documentation* of the translation/adaptation procedures. According to Hambleton (2001, 165), "Typically, little documentation of the adaptation process and evidence to support the validity of an adapted test is provided, and misinterpretations of scores from tests in multiple languages are common." Hence, the inclusion of this section is of utmost value toward the goal of ensuring the comparability of multilingual or translated/adapted tests to prevent such misinterpretations (e.g., not taking into account curriculum differences between countries on a translated/adapted test).

Research since the release of the 2001 update on the ITC Guidelines has focused on reporting these findings to differing audiences (e.g. language testers and various international psychological journals and associations, see Hambleton & de Jong 2003; Hambleton 2002; Sireci & Allalouf 2003). However, in 2006 the ITC is gathering once more to re-examine the Guidelines for Translation and Adaptation.

The problem

In sum, the consequence of Hambleton's review of the ITC's guidelines is that they, like many standards, are a work in progress. Much of the guidelines has served quite well to enhance test developers and users' collective awareness of what needs to done in a multi-lingual/cultural testing setting. However, there are areas of the guidelines that researchers point out need more work. The areas that seem most plausible for investigation are as follows: 1) a compilation of the findings from multiple test adaptation studies, and 2) better procedures for flagging potentially flawed items, such as an

extension of the use of logistic regression modeling to handle multiple languages/cultures.

Despite the great strides made toward creating much improved and better comparable multilingual translated/adapted tests, there is one central criticism to that can be made of the ITC's guidelines. The ITC guidelines are not based in theory, but rather in testing practice. This is not in and of itself a problem, rather an observation of the state of affairs. Much of testing and psychometrics has been performed in this manner (e.g., a practice to theory approach, rather than a theory to practice approach).

As evidence of this, Sireci (1998) says that much of the current practice in detecting functional non-equivalence ignores the theoretical aspect of validity analyses advocated by Messick (1989). Many studies rely primarily on the statistical indices, and some only follow up with an examination of the items or the item development categories. This contrary to what is called for in Messick's treatise (1989) calls for. Therefore, one of the goals of this dissertation will be to create a theory to aid assessment adaptation/translation. This will be performed as a consequence of performing the compilation of studies that Hambleton (2001) advocates.

CHAPTER 2: REVIEW OF THE LITERATURE

Comparability and measurement equivalence/invariance

Comparability in multilingual tests refers to the quality and relative equality of the measures associated with the instruments after translation or adaptation. Hence, it is not necessarily the one-to-one relationship of ideas between language versions of an instrument. Rather, popular practice (Hambleton, 2001) advocates the *adaptation* of instruments to their target languages or cultures. The quality of the instrument's adaptation or its comparability can be evaluated in many ways, with techniques ranging from statistical inference to expert judgment of the comparability of the items. Thus, there are two main categories of ensuring the comparability of multilingual instruments; expert-dependent methodologies and psychometric methodologies.

As previously stated, there are numerous ways to evaluate the psychometric comparability of language versions of an instrument. In fact, an entire field, known as Measurement Equivalence/Invariance (ME/I), is dedicated to the pursuit of this end. *Measurement equivalence* exists when psychometric properties from multiple groups have similar qualities (Mullen, 1995). If measurement equivalence were to be lost, the validity and generalizability of inferences based upon a multilingual instrument could be considered questionable. This is due to multiple languages or cultures possibly misunderstanding the items, which results in a manifestation of a different underlying construct. The consequence of this loss of comparability may be a systematic bias of a population which reduces the reliability and validity of the inferences made from the instrument, thereby leading to inappropriate comparative conclusions between groups (Cunningham, Cunningham, & Green, 1977; England & Harpaz, 1983).

Another primary concern of ME/I investigation has been the insufficient attention given to the underlying measurement properties of instruments (Donovan, Drasgow & Probst, 2000; King & Miles, 1995). In fact, Chan (2000) generalizes by saying that "There has been little attempt to predict *a priori* what factors result in a failure to support invariance" (p. 172). For this reason, an investigation into these matters would be warranted.

Clearly, measurement equivalence has been written about for several decades (e.g., Cunningham, Cunningham & Green, 1977; England & Harpaz, 1983; Horn & McArdle, 1992). Interest in cross-national or multilingual testing has roots several decades old as well (Butcher & Garcia, 1978; Brislin, 1986; Ellis 1989). Despite this scattering of studies on the subject, interest in the specifics of instrument comparability on a multilingual front did not really come into the spotlight until they were formalized by the International Test Commission's Translation and Adaptation Guidelines (Hambleton, 1994), which emphasizes the importance of and consequences of ensuring comparability across cultures languages.

In recent publications, van de Vijver and Poortinga (2005) point to three major sources of bias, which can lead to nonequivalence. *Bias*, which is essentially the presence of nuisance factors (i.e., those the instrument was not designed to measure), and is closely linked with nonequivalence or incomparability. Nonequivalence is the actual manifestation of bias in the application of an item in a given cultural or linguistic setting. Hence, incomparability is a measurement issue, rather than a trait of a given item or instrument (van de Vijver & Poortinga, 2005). Van de Vijver and Poortinga go on to describe how bias is exhibited. They categorize bias into three categories: construct bias,

method bias and item bias. *Construct bias* is the difference in the measured construct across cultures or languages. *Method bias* involves both instrument bias and administration bias. *Instrument bias* refers to all instrument properties that are not the target of the study. It is sometimes referred to as "test-wiseness", or a basic familiarity with the stimulus and response formats of an instrument's items. *Administration bias is* nonequivalence that results from miscommunication between tester and testee regarding the use of test items or instrument. Lastly, *item bias* is an item which exhibits DIF where group membership is not related to the construct of interest. Both of these terms essentially deal with threats to the validity of the measures affecting individual items, often displayed by items that are poorly translated (van de Vijver & Poortinga, 2005).

Common analysis methods

Given the relative length of history of the study of instrument ME/I and comparability, it should come as no surprise that there exist a multitude of methodologies to enhance and evaluate multilingual instruments. These techniques can be categorized as either expert-dependent or psychometric methods.

Expert-dependent methodologies.

One way to ensure that an item is properly converted to another language is to carefully translate the item. Consequently, the necessity of bilingual or multilingual experts becomes clear. *Bilingual* or *multilingual experts*, as defined in Joldersma (2004):

> "...are knowledgeable in both the source and target language or languages. Their expertise is crucial for the comparability of the multilingual instruments, since it is vital to have someone who is

> intimately familiar with the intricacies of both the source and target languages. An additional qualification that is desirable of comparability experts is a strong foundation in the CLI's [Cross-Lingual Instrument] subject matter. This content familiarity would enable the expert to make judgments regarding the comparability of the CLI's items. Hence, a faithful replication of the original construct, which is essential to CLI comparability, would be greatly supported by having bilingual subject matter experts verify the constructs of the CLI's language versions."

What is critically important about any translation is that it is done professionally and by qualified personnel. To reiterate, Jeanre and Bertrand (1999) state that a bilingual or multilingual expert must be able to function professionally in both source and target languages in order to be considered *functionally bilingual.*

There are essentially two methods of translation that are commonly performed with multilingual assessments: forward translation and back translation. Both methods are designed with the intent to create an instrument that has equivalent language on source and target instruments.

Forward translation.

Forward translation is the direct translation from source language to target language. One can easily recognize the attraction of implementing forward translation for instrument and item comparability. Since every language version of a test is theoretically translated from the original to each target language version, there should be some stability in terms of the content and constructs evaluated by each item. There are, of course, some important caveats to make in

the case of forward translation. Of primary concern is that "linguistic and psychological criteria for good translations ... converge" (Van de Vijer & Poortinga 2005, p. 52). Regarding linguistic concerns, the focus tends to be on the semantic similarity, comprehension, readability and style of items. The psychological criteria focus on the pragmatics of language, essentially the presence or absence of types of bias (e.g., construct, method or item, as previously described). A major disadvantage of forward translation is that it quite often does not account for cultural or linguistic differences in the target versions of the instrument. The problem of some words or concepts being nonequivalent or simply not existing in the target also arises.

As an example of a simple translation error, examine the following. An item is presented in English reading, "Point to the picture of the embarrassed lady." This may be translated incorrectly into Spanish, "Señale a la mujer embarazada" (Point to the pregnant woman). In this case, Spanish speakers will not have the appropriate answer from which to choose if there is no picture of a pregnant woman.

Back translation.

One technique used in an effort at quality control is back translation. *Back translation* is the translation of the original language to the target language and returning via re-translation back to the original language. This is done to see if the content of the back-translated material closely matches the original concept or the "original intent" of its authors. One great strength of this method is that it allows for an independent evaluation of the comparability of items between

languages if multiple and separate translators are employed. Additionally, an individual not familiar with the target language can perform the end comparability judgment. However, there are some disadvantages to this method, as well. The first potential pitfall comes from the reliance on now two rather than one translation. Hence, there are now twice as many opportunities for error in translation. Similarly, one could argue that a back translation does nothing to resolve the issue of nonequivalent or inexistent words or concepts.

While the "embarrassing" mistake from forward translation may be caught easily by back-translation, other issues may arise. For example, a culturally and linguistically appropriate translation of "potluck dinner" may be rather difficult. A potential source of incomparability then could also be lengthy circumlocution, with longer items that generally lead to lower scores. A potluck may be translated as "a communal dining event in which every person is responsible for bringing a 'dish to pass'".

Content analysis.

Another method to aid comparability is content analysis. *Content analysis* is the use of textual data to create coding schemes and categorize data in an effort to systematically understand and make decisions. In a comparability context, the decision made after coding is the extent to which items or instruments are or are not comparable. The advantage to content analysis is it that it is a quantifiable methodology for showing comparability, or perhaps indicating which components of an item or instrument need to be adjusted to create better comparability. Content analysis is typically an extremely time-consuming method, which often

necessitates the use of multiple coders or raters. Thus, problems could arise with inconsistent coding. However, multiple individuals often do, with proper training, achieve a consistently applied coding scheme.

Other expert-dependent methods to improve comparability.

None of the above approaches truly deals with the issue of nonequivalent or inexistent words or concepts. However, this is not their purpose either; they are used to evaluate an existing instrument, rather to aid the comparability of an instrument yet to be constructed. To aid the equivalence of a multilingual instrument, there are essentially three possibilities: 1) application, 2) adaptation or 3) assembly. *Application* refers to the ideal situation where an existing instrument is both linguistically and psychologically appropriate for use in source and target languages/cultures. This is the ideal situation, though does not seem to be a likely scenario. Instrument *adaptation* is the cultural adjustment of an instrument. Items are often altered to fit the target culture. In the *assembly* option, neither the source nor the target instruments have been constructed. Rather, the idea is to cooperatively develop instruments in all cultures and languages in which the instrument will be delivered. (Van de Vijjver & Poortinga 2005). Between assembly and adaptation, the major difference is between a *post hoc* and an *ad hoc* use of tests in multilingual environs. The intuitive best choice would appear to be the assembly option. This is because assembly allows for the input of all intended cultures/languages before problems may arise. The advantage then is to have the amalgamation of the construct in question, rather than a linguistically or culturally biased version of it.

In an article by van de Vijver & Tanzer (1997), the authors list the following additional strategies for addressing comparability issues:

- Decentering (i.e., simultaneously developing the same instrument in several cultures)
- Convergence approach (i.e., independent within-culture development of instruments and subsequent cross-cultural administration of all instruments)
- Use of informants with expertise in local culture and language
- Use of samples of bilingual subjects
- Use of local surveys (e.g., content analyses of free-response questions)
- Non-standard instrument administration (e.g., "think-aloud")
- Cross-cultural comparison of nomological networks (e.g., convergent/discriminant validity studies, monotrait-multimethod studies, connotation of key phrases) (p. 272).

Psychometric methodologies.

Despite the great advantage of experts, whose skills offer explanation to the nuanced and difficult issues of comparability, critics may point out the highly subjective nature of their work. Rightly so, it can be said that much expert-dependent work is based on intuition, feeling and judgment based on experience. In an effort to remove subjectivity statistical techniques are often applied to the dilemma of multilingual instrument comparability.

Descriptive statistics.

One simple method of making a comparability decision is to look at the raw descriptive statistics for an instrument. This can be done at either the test or item level. Essentially, language versions can be compared with a number of potential indicators of nonequivalent measurement. At the test level, mean total scores are likely indicators of equivalence. Mean difficulties for items, known as *p-values*, are quick indicators of comparability. For either of these statistics to aid in test comparability two

considerations must be made. First, are the examinees taking the same item and same number of items? Second, is variance similar on language versions both at item and test level? Like all statistics, a decision must be made on the tolerance for errors. Though these statistics have the advantage of being objective measures, their great disadvantage is a lack of explanatory power. Additionally, despite the simplicity of calculation, these statistics also cannot account for examinees of differing ability levels.

Dimensionality assessment.

Current techniques for assessing the construct equivalency at an instrument level generally deal with dimensionality analysis. There are a number of ways to explore and analyze the *dimensionality* or the data similarities present in the response strings of examinees. Among the methods popularly employed in the measurement equivalence field are Exploratory Factor Analysis, Confirmatory Factor Analysis, Multidimensional Scaling and Comparison of Nomological Networks.

Exploratory Factor Analysis.

Sireci, Patsula & Hambleton (2005) write that Exploratory Factor Analysis (EFA) is the most widely used statistical technique to assess the cross-culture equivalence of constructs. It is typically used to assess the frequency and structure that a construct is present in different languages or cultures. The authors continue by saying that although the approach is intuitive, there are truly no standardized methods for deciding what level of comparability is acceptable or, at a more basic level, whether those structures are equivalent at all. Hence, a more complex data analysis method is more profitable for these purposes (Sireci, Patsula & Hambleton, 2005).

Confirmatory Factor Analysis.

A method that does help to take on the challenge of multiple groups and simultaneous analysis is Confirmatory Factor Analysis (CFA). This is done by allowing a hypothesized format for the data; i.e., allowing the different language or cultural groups to be modeled. The necessity of specifying a data structure may be considered one of the difficulties of working with CFA. Despite this, Sireci, Patsula & Hambleton (2005) list multiple studies that have used this approach to evaluate cross-cultural construct similarity. They also continue by saying that CFA works well in most scenarios and is appealing because it can handle multiple groups simultaneously, fit indices are available and statistical tests of model fits likewise exist. A problem, does however, arise for this method when using dichotomous data. Since the underlying models of CFA are linear in nature, and dichotomous data are often non-linear, this may cause problems for analysis. The specific problem deals with the issue of traditional factor analysis yielding factors that are highly related to difficulty, rather than content. In effect, easy items load on one factor and hard items load on another factor. This is a consequence of using the Pearson correlation for dichotomous data, rather than the more appropriate tetrachoric correlation. Sireci, Patsula & Hambleton (2005) suggest that grouping items together in "parcels" before analysis will remedy this problem. Other testing professionals use TESTFACT or full information factor analysis (Reckase, personal communication, November 2, 2005).

Multidimensional Scaling.

Another statistical method that can be used in establishing construct equivalence is Multidimensional Scaling. The great advantage of Multidimensional Scaling (MDS) is

that it, like EFA, does not require specifying test structure before analysis. Additionally, multiple groups can be analyzed as in CFA. On top of this, MDS can handle both linear and non-linear data. From the perspective of cross-cultural construct equivalence analysis, MDS appears to have the greatest potential (Sireci, Patsula & Hambleton, 2005; Sireci, Bastari & Allalouf 1998; Carroll & Chang 1970).

Differential Functioning.

Other statistical tests address the concept of *differential functioning*, where examinees of purportedly similar ability levels perform differently. Indices exist for examining the overall comparable functioning of examinees at the test level (see Raju et al, 1995; Ellis &Mead 2000; Shealy & Stout 1993). One may take the perspective that a test is not biased unless it is biased at the point at which decision are made (i.e., the test level). Items may exhibit DIF or an instrument may exhibit DTF (differential test functioning). However, it seems impractical if not unfair to make a decision based on a test or set of items that are known to lack psychometric or substantive equivalence. Additionally, current research indicates that the analysis at the instrument level may not be appropriate for a cross-cultural equivalence studies because it misses details of inequivalency at the item level (Zumbo, 2003).

There is no universally accepted manner for assessing differential functioning at the item level or Differential Item Functioning (DIF). The advantage of multiple methods is that an analyst can choose the method most appropriate to the data. Below in Table 2.1, Sireci Patsula, Hambleton (2005) illustrate many techniques for DIF analysis, where it was first presented, what data it can be applied to, and studies that have applied these methods to cross-cultural/lingual assessment.

Table 2.1. Methods for Detecting Differential Item Functioning (from Sireci Patsula, Hambleton, 2005)

Method	*Sources*	*Appropriate*	*Applications to Cross-Lingual Assessment*
Delta	Angoff (1972, 1993)	Dichotomous data	Anghoff & Modu (1973) Cook (1996) Muniz et al (2001) Robin, Sireci & Hambleton (2003)
Standardization	Dorans & Kulick (1986); Dorans & Holland (1993)	Dichotomous data	Sireci, Fitzgerald & Xing (1998)
Mantel-Haenszel	Holland & Thayer (1988); Dorans & Holland (1993)	Dichotomous data	Allalouf et al (1999) Bugell et al (1995) Muniz et al (2001)
Logistic Regression	Swaminathan & Rogers (1990)	Dichotomous data Polytomous data Multivariate matching	Allalouf et al (1999) Gierl et al (1999)
Lord's Chi-square	Lord (1980)	Dichotomous data	Anghoff & Cook (1988)
IRT Area	Raju (1988, 1990)	Dichotomous data Polytomous data	Budgell et al (1995)
IRT Likelihood Ratio	Thissen et al (1988) Thissen et al (1993)	Dichotomous data Polytomous data	Sireci & Berberoglu (2000)
SIBTEST	Shealy & Stout (1993)	Dichotomous data	

Logistic regression has the advantages of data flexibility and effect size computation. It is important when working with secondary data collection, as is the case in this project, to be able to work with multiple data types. Additionally, the ability to compute an effect size with relative ease (using Zumbo's R-squared method, Zumbo, 1999) is of vital importance given the aims of the project.

Comparison of Nomological Networks.

While evidence for the structural equivalence of the constructs can be provided via EFA, CFA and MDS, we cannot necessarily conclude that the constructs are indeed the same. Van de Vijver & Tanzer (1997) suggest going beyond these techniques to a more global approach of equivalence. Returning to the original work by Cronbach and Meehl (1955), Van de Vijver & Tanzer stress that the concept of construct equivalence was introduced simultaneously with the term "nomological network", which emphasizes that a test's inferences cannot be validated using a single criterion. In the past, the difficulty of establishing and measuring test scores in relation to multiple external factors (e.g. linguistic and cultural factors) has been more mostly prohibitive. Despite this, Van de Vijver and Tanzer strongly encourage a search for multiple sources of convergent and discriminant validity evidence for all cultural or linguistic groups evaluated by an instrument. Assuming that this evidence can be obtained, the question then remains: How do we take a decision based on this information? The application of such a technique appears to be unwieldy at best. This dissertation seeks to implement a version of a nomological network to show how it may be done.

Summary of Advantages and Disadvantages of Methods

Although each method has certain advantages, each also has disadvantages. Some of these are summarized in Table 2.2. Clearly, each of these techniques is best implemented under the appropriate circumstances, from low stakes tests to high stakes certification and entrance exams. Typically, the more labor-intensive expert-dependent methods are the more complex, and hopefully more sensitive to issues of cultural incomparability (Hambleton & Jones, 1994). Psychometric methods also range in their

complexity, as well. It is the difficulty of computation, however, that stratifies these methods. Therefore, their implementation likely is most impacted by knowledge of test evaluators/designers and again, the stakes of the test involved.

Table 2.2 Advantages and Disadvantages of the Expert-dependent and psychometric methodologies for translation/adaptation evaluation.

	Advantages	Disadvantages
Expert-dependent		
Forward translation	• Easy to implement • Theoretical construct and content stability	• Linguistic and psychological differences plentiful between cultures and languages • Subjective results/interpretation
Back translation	• Relative ease of implementation • Quality control check on translations • Theoretical construct and content stability	• Multiple translations may mean multiple mistakes • linguistic and psychological differences plentiful between cultures and languages • Subjective results/interpretation
Content Analysis	• Quantifiable results • Detailed analysis	• Time consuming • Multiple raters must be trained • Somewhat subjective results/interpretation
Psychometric		
Simple descriptive statistics	• Simple calculation • Objective measurement	• No explanatory power • No standard for decision-making
Dimensionality analysis	• Objective measurement • Multiple methods for multiple data types	• No explanatory power without expert • Methods limited by data type
DIF analysis	• Objective measurement • Multiple methods for multiple data types	• No explanatory power without expert
Nomological Networks	• Multiple sources of validity documentation	• **Difficulty of implementation?**

Despite all the information that can be garnered from all of the above techniques, there is no one commonly accepted method for evaluating cross-cultural or linguistic instruments. Nor would this be desirable, for the reasons previously described of multiple testing environs and needs. There is however, a call in the literature, most notably Messick's foundational work (1989), for multiple sources of validity evidence (Cronbach and Meehl, 1955; Van de Vijver & Tanzer, 1997; Messick, 1989). There is no debate that it is ideal for every instrument to have multiple means of validity analysis. Multilingual and cross-cultural assessments are certainly no exception. In truth, the very fact that an instrument has multiple linguistic and cultural contexts seemingly demands this type of evidence.

The manner of collection and the techniques available is not in question. What is important to address is what to do with the information once it is gathered. Say for instance that the results of multiple expert-dependent techniques do not agree, what is to be done then? Perhaps only one method shows contrary results, do we discount its results? This is, unfortunately, what may happen as a consequence of subjective analysis, wherein a decision is made by the expert to go with the preponderance of the evidence. To this author, losing this information is a rather unsatisfying end after making efforts to have those multiple sources of validity evidence to support arguments of comparability in the first place. Therefore, one must consider how to objectively combine evidence to make a decision while still using as much information as possible to inform that decision.

Methods of Integration of findings

Over the past century, an evolution of the various methods for integrating findings across studies or data-analytic procedures has occurred. Hunter and Schmidt (1990) critique ten different methods for the amalgamation of findings.

Traditional narrative procedure.

The traditional approach to multiple studies and their findings was to provide a narrative of the various studies and findings. This was done to guide the reader through the wealth of information with the end goal of establishing a theory that reconciles the findings. The end result of the process, according to Hunter and Schmidt (1990) is one of three possibilities. The result may be "pedestrian reviewing where verbal synopses of studies are strung out in dizzying lists (Glass, 1976, p. 4)". In effect, no integration is really attempted across studies. The second possibility is that the reviewer will simplify the results by only addressing a small subset of all the findings. The effect of this choice is to limit the information that is used and perhaps to promote a preconceived "truth" based on studies which support that "truth". Alternatively, the final outcome may be that a reviewer attempts to mentally accumulate the findings across studies. The problem with this approach is that the integration is likely to be subjective and unsystematic.

Traditional voting method.

The traditional vote counting method essentially relies on reporting the number of significant and non-significant findings (p-values) in the literature review or studies. Unfortunately, relying on p-values has the tendency to show bias in favor of studies with large sample sizes. Moreover, the actual size of the effect is unknown using this

procedure, since we are only comparing p-values. Lastly, Hunter and Schmidt (1990) report that a study by Hedges and Olkin (1980) shows that in any set of studies where power is less than .50, the likelihood of false conclusions actually increases as the number of studies rises. Hence, this method uses only part of the available information, does not report an effect size and can possibly lead to false conclusions.

Accumulation of p-values across studies.

The next step in the progression of accumulation of studies after reporting p-values could be to mathematically cumulate those p-values across studies. In effect, the results of all the studies are combined into a single p-value. If the value is small enough, the conclusion of this method is that there is evidence for an effect. Though the p-value may be significant and provide evidence of an effect, there is no measure of the magnitude of that effect. One solution is to average effect sizes across studies. Still, even after effect size averaging, this method lacks the information about the variability of effects across studies.

Statistically correct vote-counting methods yielding only significance levels.

One of two statistically correct methods of vote-counting methods is one which yields only significance levels. Within this category are three different procedures for cumulating results. One option is to count only positive significant findings. Another option is to count positive results in general. Lastly, a researcher might choose to count both positive and negative significant results. All of these methods have the disadvantages of only using part of available information and not reporting an effect size.

Statistically correct vote-counting methods yielding estimates of effect sizes.

These methods are an improvement over methods which do not report effect sizes. However, the uncertain quality of effect size estimate, due to partial use of information available, creates a problem. If the researcher assumes that effect sizes across all studies are equal, we can only have an approximate estimate of mean effect size, rather than a truly integrated result across studies.

Meta-analytic methods.

Glass (1976) coined the word 'meta-analsyis'. His *Glassian meta-analysis method* is the first true form of systematic integration of research findings. One advantage of the method is that it uses more of the available information in the studies involved. It also includes a more accurate estimate of effect size. This estimate allows effect sizes to vary across studies, thus providing an estimate of variance of effect sizes. Additionally, the method also allows for correlating effect sizes with study characteristics in order to examine the causes of this variation.

Study effects meta-analysis has the advantage of making clearer conclusions about relationships between specific independent and dependent variable constructs. This allows for finer tests of scientific hypotheses. The important differences between study effects meta-analysis and Glassian meta-analysis are that Glass' method only allows for one effect from each study to be certain of statistical independence and also asks the meta-analyst to make judgments about the quality of the studies involved in order to exclude studies which may distort outcomes.

Schmidt and Hunter (1990) claim that *homogeneity test-based meta-analysis* is often less useful than Glassian methods. The method does make efforts to test for

moderator variables. Essentially, multiple tests are performed by grouping all the studies into ever smaller groups until all remaining variance is attributable to sampling error. The problem with this approach, according to Schmidt and Hunter (1990) is that it is once again a return to p-value based integration, with all of its shortcomings.

The *Schmidt-Hunter meta-analysis methods* (Hunter, J.E., Schmidt F.L. & Jackson, G. 1982) for validity generalization provide a more accurate estimate of effect sizes. This is enabled by weighting procedures. The method also corrects effect size estimates by removing the effects of instrument unreliability and range restriction. Additionally, there is a test of the hypothesis that variance in the effect sizes is due to artifacts. One shortfall of this method is that it lacks a step for correlating effect sizes with study characteristics in order to examine causes of this variation when S^2_{ES} (the variance of the effect sizes) cannot be accounted for by artifacts alone.

Finally, the *Schmidt-Hunter psychometric meta-analysis* (Viswesvaran & One 1995) is a method which does allow for artifact examination. It can also be extended for use with both *d* and *r* statistics. The method also provides for the correction of additional artifacts. The only drawback of this method occurs when effect sizes are regressed on study characteristics. Though not always a necessary step, there may be problems of capitalization on chance and low statistical power in this case. Thus, the decision to subgroup studies is best done using *a priori* reasoning as opposed to an *ad hoc* decision based on the results of a Q test for homogeneity. Moreover, if the Q test does not indicate homogeneity, there is no indication as to which studies should be grouped.

Table 2.3. Summary of Integration techniques.

Method	**Comments**
Traditional Narrative Procedure	Subjective and unsystematic
Traditional Voting Method	Uses only part of available information, no effect size reported, can lead to false conclusions
Cumulation of p-Values Across Studies	Uses only part of available information, no effect size reported
Statistically Correct vote-counting methods yielding only significance levels	Uses only part of available information, no effect size reported
Statistically Correct vote-counting methods yielding estimates of effect sizes	Uncertain quality of effect size estimate due to partial use of information (assumption that effect sizes across all studies are equal. When assumption violated, yields approximate estimate of mean effect size.)
Glassian meta-analysis methods	Uses more of available information, more accurate estimate of effect size (effect sizes can vary across studies), provides estimate of variance of effect sizes, allows for correlating effect sizes with study characteristics in order to examine causes of this variation
Study effects meta-analysis	Clearer conclusions about relationships between specific independent and dependent variable constructs (allowing finer tests of scientific hypothesis)
Homogeneity test-based meta-analysis	Less useful than Glassian methods
Schmidt-Hunter meta-analysis methods: validity generalization	More accurate estimate of effect sizes (by weighting), corrects effect size estimates by removing effects of instrument unreliability and range restriction, also provides tests of hypothesis that variance in effect sizes is due to artifacts Lacks step for correlating effect sizes with study characteristics in order to examine causes of this variation when S^2_{ES} cannot be accounted for by artifacts alone
Hunter-Schmidt psychometric meta-analysis	Allows for artifact examination. Extension to *d* and *r* statistics. Provision for correction of additional artifacts. If effect sizes are regressed on study characteristics (not always necessary) there may be problems of capitalization on chance and low statistical power.

Criticisms of Meta-analysis

Meta-analysis is a widely used methodology for objectively combining the findings of multiple studies. Commonly, it is used to merge the findings of both quantitative and qualitative work, as is the case of in this study. Additionally, it has the ability to effectively combine multiple and often conflicting results. Despite its great power for objectively merging findings, meta-analysis is not without its detractors. Various texts report four major criticisms of meta-analysis (Wolf, 1986; Hunter & Schmidt, 1990). Glass et al (quoted in Wolf, 1986, p. 14) effectively summarizes these criticisms as follows:

1) Logical conclusions cannot be drawn by comparing and aggregating studies that include different measuring techniques, definitions of variables (e.g., treatments, outcomes), and subjects because they are too dissimilar.

2) Results of meta-analyses are uninterpretable because results from "poorly" designed studies are included with results from "good" studies.

3) Published research is biased in favor of significant findings because nonsignificant findings are rarely published; this in turn leads to biased meta-analysis results.

4) Multiple results from the same study are often used which may bias or invalidate the meta-analysis and make the results appear more reliable than they really are, because these results are not independent.

Wolf continues by adding that each of these criticisms has been addressed in meta-analytic methodological research. This first criticism of meta-analysis is that it is

much like comparing apples and oranges. In effect, critics claim that the fact that the methods are different and the constructs may differ slightly in their definitions may create an inability to fairly combine the results from the different studies. However, meta-analysis is able to compensate for this possibility by empirical means. Studies can be coded by characteristic to confirm whether or not these differences are related to the meta-analytic findings. The second criticism too is dealt with by empirical coding. Design quality, Wolf (1986) argues, is not related to the effect size magnitude of the outcome. The third criticism, that of bias in favor of positive results, is a more difficult issue. Two solutions are proffered: to look for unpublished work or to estimate the number of unpublished studies to disprove the findings in published work (known as a fail-safe N). The final criticism of multiple results from a single study has been addressed in multiple manners as well. Approaches range from including each outcome variable to addressing each individually. However, researchers suggest that the use of GLM or multiple-regression may sufficiently address the interdependency, interaction and covariance caused by using multiple results from a single study (Wolf 1986).

In sum, meta-analysis, despite its detractors and competitors, appears to be the most effective way to objectively combine the results from multiple analyses of the translated/adapted instrument evaluated in this study. Its ability to account for quantitative and qualitative data suggests that it is the ideal method for making a validity argument for the comparability of a translated/adapted test. From a practical standpoint, meta-analysis is also applicable to the types of data that is generated from comparability analysis.

Theoretical framework of characteristics impacting item comparability

It is clear that previous research has impacted the design of the ITC guidelines for translation and adaptation. This research is explicit in its details of which items will be more or less comparable. This comparability may be impacted by several important factors, including item characteristics and/or degrees of linguistic and cultural similarity.

According to Jeanre & Bertrand (1999), item comparability can be assured by paying attention to content, conceptual equivalence and linguistic equivalence. For content, this means that an item should ideally have appropriate symbols and situations for the respondents. Conceptual equivalence checks whether an item exists in the target culture and additionally whether it has been meaningfully translated. The original meaning of an item is also assessed to assure that it was not significantly altered. Linguistically, Jeanre & Bertrand (1999) are concerned that words, tenses, and idioms, etc. across languages be comparable, as well.

Other research by Reckase, Xin & Joldersma (2004) points to item presentation as a major factor in maintaining cross-linguistic comparability. This paper shows that apart from major factors played by language and culture, item presentation needs to be culturally appropriate for an adaptation process to be successful. For example, in addition to linguistic and cultural factors mentioned by Jeanre & Bertrand (1999), other discrepancies in cross-cultural testing may be caused by typographical errors, improperly adjusted graphics or item length.

One's first inclination might be to assign more weight to certain features as more influential in determining the comparability of an item. The fact is, however, that there is not a well-defined indication of which of factors mentioned by the ITC and the above

research might be more or less important. Additionally, there is no clear indication as to which of the methods previously described might be better or worse for assessing cross-cultural/linguistic comparability. Thus, it does not make sense to assign more importance to any one of the item characteristics. Additionally, since no one method seems to emerge as better or more useful, it seems that each method should be treated equally, as well.

The Current Study

This literature review has established that psychometric and expert-dependent methodologies are both integral components of an assessment of multilingual instrument comparability. Moreover, the literature indicates that evidence for validity is derived via a nomological network (i.e., a body of evidence). Hence, psychometricians are obligated to provide such evidence, if available. The next step, then, is to make the best decision we can regarding that body of evidence. The current study proposes meta-analysis as that technique, since it appears to be a highly efficient manner of combining multiple findings or analyses. This study will additionally seek to answer what the impact of such a technique may be.

1. How does one generate a 'body of evidence', as called for in Cronbach and Meehl (1955), for the validity of a cross-cultural translation/adaptation?
2. Assuming that such evidence of cross-cultural adaptation/translation can be generated, is meta-analysis an effective methodology for the integration of the findings across methods within this body of evidence?

3. Within that body of evidence, which methods are likely to produce smaller or larger effects for the evaluation of cross-lingual/cultural comparability? Which of these methods are better suited for the assessment of cross-cultural translation/adaptation?
4. What type(s) of items or content are likely to trigger differences or difficulties in cross-cultural translation/adaptation?

CHAPTER 3: METHODOLOGY

Overview

In order to address the research questions proposed for the current study, it behooves us to examine how they may be answered.

1. How does one generate a 'body of evidence', as called for in Cronbach and Meehl (1955), for the validity of a cross-cultural translation/adaptation?

This question is essentially addressed in the literature review, which details multiple methods both expert-dependent and psychometric in nature that may be used to compile evidence of comparability or lack thereof for the instrument in question. The important 'next step' that must be taken is to choose which methods are best suited to a comparability decision-making process. This choice is directed by the author's decision to use meta-analysis to arrive at an overall decision for item comparability.

2. Assuming that such evidence of cross-cultural adaptation/translation can be generated, is meta-analysis an effective methodology for the integration of the findings across methods within this body of evidence?

Once the data is generated, the aim of this project is to assess meta-analysis as an effective means of integrating findings of a nomological network. It will be important to consider the implications for combining upon the information in terms of the gains and losses in information and how that impacts the quality of the decisions that are made. Evidence that meta-analysis makes an effective decision shall be judged by the ability of the combined results to either flag and/or explain items lack of comparability.

3. Within that body of evidence, which methods are likely to produce smaller or larger effects for the evaluation of cross-lingual/cultural comparability? Which of these methods are better suited for the assessment of cross-cultural translation/adaptation?

In order to improve meta-analysis as a technique, it will be useful to identify which methods contribute greater or lesser effects to the over-all effect size. This will, in part, enable user to determine which methods may be best suited based on two factors. Another factor that will aid the process is a measure of the overall stability of a given method's measures of comparability (variance). Together this information should aid our analysis of the body of evidence.

4. What type(s) of items or content are likely to trigger differences or difficulties in cross-cultural translation/adaptation?

The utility of knowing what types of cultural differences amount to psychometric irrelevancies is indisputable. Ideally, subsequent analysis (post meta-analysis) illustrates further which items and content are key to aiding cross-cultural comparability. Items which measure medium or large effects should aid this process are likely good indicators of which type of content or item might be troublesome.

Sample

The participants in this study are children who have participated in the Preprimary project, known as the PPP (IEA-PPP, 1994). The data for this study come from the cognitive and language development measures associated with this multipart assessment. For the Phase III data used in this study, children were approximately 7-9 years of age. The three countries selected for this analysis are the United States, Spain and Italy;

chosen as a convenience sample due to the author's language familiarity and availability of language professionals.

Descriptive statistics and demographic information are available for these tests by way of a scaling study done by Wolfe and Manalo (2002), shown in Table 1. The Italian and American children in the sample are roughly the same age, with Italian children being slightly older (8 years compared to 7.7 years). The sample also consisted of nearly identical percentages of males and females (54% male and 46% female in Italy and 53% male and 47% female in the US samples).

Table 3.1. Sample for developmental status measures by type for PPP.

Country/Territory	Phase 3 *N* Child	% Male	Mean Age (Years)
Italy	246	54	8
United States (6 sites)		53	7.7
Head Start centers	59		
Public school preschools	66		
Other organized programs	122		
Family day care homes	55		
Own homes	61		
(Total)	(363)		

Though the sample is from a population of primary-school age children, it is hoped that the results of this study generalize to many types of multilingual assessments of all people of all demographic and linguistic backgrounds.

Instrumentation

The instrument used in this research is the PPP (Preprimary project). The Preprimary project is an international assessment that collects many types of data, including children's cognitive and language development. The portion of the PPP used in this study is the cognitive assessment, which contains several subcategories of items.

The cognitive test reflects five subscales of ability, including: a) quantitative, b) spatial relations, c) and time perception, d) memory, and e) problem solving. The Cognitive Developmental Status Measure uses prompts that require children to demonstrate understanding of a wide variety of concepts by performing an action, pointing to a picture or responding verbally, sequencing events or pictures, or completing drawings (Claxton 2003). Claxton gives the following examples for each subcategory:

Spatial Relations

Procedure: The child is asked to indicate which one of a set of pictures fits the description provided for the test item.

"Look at the boxes. One box has an animal in it. *Skip* a box and point to the next one."

Quantity

Procedure: The child is asked to indicate which one of a set of pictures fits the description provided for the test item.

"Look at the shapes divided into three parts. Point to the shape that is divided into three *equal* parts."

Time

Procedures: Show the child the four sentence cards and say, **"Each of these cards has one sentence on it that explains a part of how you do something. Read each of the sentences and put them in the correct order."**

"What do you do when you put on your socks and shoes?"

Scoring: The child receives 2 points if all the sentences are arranged in the correct order, 1 point for attempting to put them in order and 0 points if no attempt is made to put the sentence cards in order.

Memory

Procedure: Say to the child, **"I am going to say a sentence. Listen carefully and say the sentence exactly as I say it."**

"The shape of a leaf tells what kind of tree it is."

Scoring: The child receives 1 point for each sentence repeated correctly in the order presented. Points are not deducted for articulation or speech errors. However, any omission, grammatical error, substitution, or deviation from the word order results in an incorrect response.

Problem Solving

Procedure: Say to the child, **"A friend gave this girl a box of carrot seeds and a box of daisy seeds, but she forgot to write the names of the seeds on the boxes. She wants to plant the carrot seeds in her garden and the daisy seeds around her house. How can she find out which seeds are in each box? Think of as many things as you can that the child could do to find out which seeds are in each box."**

Scoring: The child receives two points for giving two to three logical causes of or solutions to the problem. The child receives only one point for giving one cause of or solution to the problem. Some appropriate answers are, "Ask her friend," "Ask an adult," "Open it up to look at the seeds and get a picture to see what they are," "Plant them and see what will grow," "Take them to the store and ask them," etc.

The instrumentation selection and development for the PPP was the result of an international collaboration. First, the PPP steering committee and research coordinators from the participating countries defined specific areas of measurement for each of the variables of interest. Some existing instruments were reviewed to develop the measures for the PPP. These instruments had to meet the criteria of multi-cultural suitability in order to be considered. Additionally, the instruments needed to have an appropriate level of difficulty and be easy to administer in a one-on-one situation. Moreover, the instrumentation in this study received substantial input from the countries involved in the

PPP over a period of years. This included two rounds of pilot-testing in each country with revision in between (Claxton, 2003).

Data collection procedures

This study uses secondary data collected as part of the PPP Phase III assessments. As such, the PPP report written by Claxton (2003) summarizes specific details of the data collection. In that report, Claxton describes the process as follows:

> "[the test designers] developed a common set of training procedures and recommendations for all countries participating in the study. Although training sessions varied from country to country in presentation and style, all countries were required to meet minimum observation system training standards. The data collectors selected were persons with experience in early childhood, such as teachers or graduate students in the field. Data collectors in each country had to reach or exceed an interrater reliability of 80% on the observation instruments." (Claxton, 2003)

This demonstrates the great lengths that test developers went to ensure consistent data collection. After a series of observations and interviews, the data collection for the cognitive developmental assessment was performed. Data collectors did this in one-on-one interview situations with the children whereupon they were asked the questions as exemplified in the Instrumentation section.

Analysis

As demonstrated in the literature review, the combination of expert dependent and psychometric methodologies is necessitated for good instrument development or evaluation. Accordingly, the methods chosen for the analysis of this project's instrument

will be of both types. Moreover, since meta-analysis seems to be indicated as the most feasible means in which to combine the findings from these analyses, it will be described as well. Finally, comparability criteria for the items will be detailed. Thus, the analysis section will be divided as follows:

1. Expert dependent analysis
2. Psychometric analysis
3. Meta-analysis
4. Comparability Criteria and Data Analysis

Expert dependent analyses

Forward translation—As the name implies, this analysis is performed by a one-way translation (e.g., source language to target language). The items can then be verified for their comparability to the existing translated/adapted versions. Three bilingual individuals (3 language teaching professionals with expertise in both source and target languages in this study) evaluate the language versions of the items for comparability. Two of the three experts were native speakers of Italian, while 1 was a native Romanian speaker. The experts are asked to spot-check the instrument for linguistic and cultural appropriateness as well as content similarity. Each item evaluated is coded using the labels detailed in Table 3.2 below (adapted Jeanrie & Bertrand, 1999).

Table 3.2. Expert-dependent coding.

Area of comparability	Rating
Content—situations or symbols are appropriate for the cultures.	(1) identical (2) very similar (3) somewhat similar (4) somewhat different (5) very different
Conceptual (construct) equivalence—concepts represented in item are:	
(1) in existence in the languages or culture, and	(1) equally existent in languages/cultures (2) mostly existent in languages/cultures (3) somewhat existent in languages/cultures (4) somewhat inexistent in languages/cultures (5) inexistent in languages/cultures
(2) meaningfully translated into the target language or culture.	(1) meaningful in languages/cultures (2) mostly meaningful in languages/cultures (3) somewhat meaningful in languages/cultures (4) somewhat meaningless in languages/cultures (5) meaningless in languages/cultures
Original intent—the meaning of the target item, when compared to the source item is:	(1) identical (2) very similar (3) somewhat similar (4) somewhat different (5) very different
Linguistic equivalence—considering the original tenses, markedness (gender, number, case appropriate), words choice, idioms, etc., the translated items:	(1) use perfectly equivalent language in its form and meaning (2) use mostly equivalent language in its form and meaning (3) use somewhat equivalent language in its form and meaning (4) use somewhat nonequivalent language in its form only (5) use nonequivalent language.
Item presentation—items are typographically accurate, of similar sentence length, use appropriate layout, use appropriate graphics (charts, graphs, etc.)	(1) appropriate in languages/cultures (2) mostly appropriate in languages/cultures (3) somewhat appropriate in languages/cultures (4) somewhat inappropriate in languages/cultures (5) inappropriate in languages/cultures
Holistic equivalence—in your judgment, the item as a whole is:	(1) highly comparable (2) very comparable (3) moderately comparable (4) somewhat comparable (5) not comparable

The three judges' ratings are averaged for each category and that value is assigned to that item. To ensure a balanced representation of the categories deemed important by the literature review, the six major categories above are used as six separate characteristics. Thus, within construct equivalence, subsections 1 and 2 are combined (averaged) to produce one rating. To ensure relatively reliable results from the experts, a test of inter-rater reliability will be performed, using PRAM software (Skymeg Software, 2005) to calculate percent agreement and Holti's coefficient of reliability (Holsti, 1969).

The equation for Holsti's coefficient of reliability is as follows:

C.R. = 2 M/ N1 + N2, where

C.R. = coefficient for reliability

M = number of coding agreements between the judge and

N1 = number of coding decisions made by judge 1

N2 = number of coding decisions made by each judge 2.

Back translation—Back translation of the instrument will begin with the existing translated/adapted instrument (e.g., target to source language). A bilingual expert then translates each selected item into the original source language. These items are then compared to the original version of the items in the source language and coded again as shown in Table 3.2 above. Again, the results from this procedure are reported in table format and converted to 6 average values for each characteristic of each item.

Psychometric analysis

Simple descriptive statistics—The following statistics are calculated for each item in each language version: p-values, point biserial correlation, internal consistency and

reliability. Large discrepancies will be noted for further examination. Results are presented in table format with standard associated statistics (N, s.d., $\overline{X}$,etc.). Most critical is the p-value (proportion correct). The p-value is used to generate the effect for each item by comparing the proportions correct on that item for the target language versus all test takers.

DIF assessment and analysis—Differential item functioning is evaluated for each item using logistic regression, because of its flexibility to work with multiple data types, and more importantly, existing literature on how to calculate an effect size using Zumbo's R-squared method (detailed below).

Logistic Regression with Binary Items.

The method of DIF analysis chosen for this data analysis is logistic regression (LR). In addition to previous reasons given, Zumbo (1999, p 22) states that "one of the most effective and recommended methods for detecting DIF is through the use of logistic regression (Clauser & Mazor, 1998; Swaminathan & Rogers, 1990)". Additionally, LR works well with binary or dichotomous item types, which are the type present on the instrument which is the subject of this analysis. Logistic regression is a statistical model which accounts for the probability of responding correctly to an item based on group membership. Group membership in the present study is the difference between the reference group (American English test takers of the original instrument) and the focal group (test-takers in all other language versions of the test, specifically the Italian version). This difference is conditioned upon a criterion variable, which in this case is the total score.

The LR procedure uses the item response (0 or 1) as the dependent variable, with grouping variable (dummy coded as 1=reference, 2=focal), total scale score for each subject (characterized as variable TOT) and a group by TOT interaction as independent variables. This method provides a test of DIF conditioned on the relationship between the item response and the total scale score, testing the effects of group for uniform DIF, and the interaction of group and TOT to assess non-uniform DIF.

The logistic regression equation is

$$Y = b0 + b1TOT + b2LANGUAGE + b3TOT* LANGUAGE.$$

where Y is a natural log of the odds ratio. That is, the equation

$$\ln\left[\frac{p_i}{(1-p_i)}\right] = b0 + b1\,tot + b2\,group + b3(tot* group),$$

where p is the proportion of individuals that endorse the item in the direction of the latent variable. One can then test the 2-degree of freedom Chi-Square test for both uniform and non-uniform DIF (Zumbo 1999).

Tests of Significance for DIF.

In order for us to determine whether an item should be flagged for DIF or not, there is a test of significance for LR. There is a natural hierarchy for entering variables into the DIF model as follows: 1) enter the conditioning variable (total score), 2) the group variable is entered, and 3) the interaction term is entered into the equation.

This information is all that is needed to compute the statistical tests for DIF in LR. In effect, the Chi-square values from step 3 are simply subtracted from those in step 1. This value is in turn compared to a Chi-square distribution with 2 degrees of freedom. The resulting two-degree of freedom Chi-squared test is a simultaneous test of uniform

and non-uniform DIF (Swaminathan & Rogers, 1990). This modeling strategy is essential to test whether the group and interaction variables are statistically significant over-and-above the conditioning (i.e., matching) variable.

Measures of the Magnitude of DIF (Effect size).

Measuring the magnitude of DIF in the context of multi- cultural/linguistic instrumentation will need to be done using the Cohen guidelines (Cohen, 1992), since there currently is no set standard as to how large or small these effects will be. To generate the effect size, the process is similar to that used for the statistical hypothesis test. The major difference is that R-squared values are used at each phase.

Zumbo and Thomas (1997) state that both the 2-df Chi-square test (of the likelihood ratio statistics) in logistic regression and a measure of effect size are needed to identify DIF. This is done to prevent overemphasizing trivial effects which are statistically significant when the DIF test is based on a large sample size. The Zumbo-Thomas measure of effect size for R^2 parallels effect size measures available for other statistics.

There are essentially two criteria for an item to be classified as displaying DIF using Zumbo and Thomas method:

- The two-degree-of-freedom Chi-squared test in logistic regression should have a p-value less than or equal to 0.01 (set at this level because of the multiple hypotheses tested), *and*
- The Zumbo-Thomas effect size measure had to be at least an R-squared of 0.13 (which is essentially a reconstituted form of Cohen's 1992 guidelines).

R-squared Measures for DIF.

Table 3.3 shows the R-squared measures to measure the magnitude of DIF. The measure most appropriate for the current study is dependent on the data we have, which are dichotomous items. The items are achievement-type questions scored where 1 is a reflection of more of an ability of trait and 0 is less of that trait. Hence, the 1 and 0 scoring represents a collapsed continuum forced into two values.

Table 3.3. R-squared Measures for DIF

Item Scoring	**Measure**	**Notes**
Ordinal	R-squared for ordinal	McKelvey & Zavoina (1975)
Binary (nominal)	Nagelkerke R-squared	Nagelkerke (c.f., Thomas & Zumbo, 1998)
Binary (nominal)	Weighted-least-squares R-squared	Thomas & Zumbo (1998)
Binary (ordinal)	R-squared for ordinal (i.e., same as above)	McKelvey & Zavoina (1975)

The latent trait underlying the items is important to our statistical analysis because the technique we choose depends on the nature of the items. The ordinal logistic regression decomposes the variation in y*, the latent continuous variable defined in the LR model, into "explained" and "unexplained" components. Zumbo (1999, p. 29) says:

> "...as per the typical use of regression, this squared multiple correlation then represents the proportion of variation in the dependent variable captured by the regression and is defined as the regression sum of squares over the total sum of squares. Therefore, the R-squared values arising from the application of ordinal logistic regression are typical in magnitude

to those found in behavioral and social science research and Cohen (1992) and Kirk's (1996) guidelines may be useful in interpretation."

Meta-analysis

The meta-analyses for each of the items analyzed will be carried out using the following procedures (adapted from Hunter & Schmidt, 1990, p 485-487):

1. Establish the basic facts (variables and values) surmised from the above mini-studies.
2. Express key findings in a common statistic (correlation or d "difference" statistic for effect sizes).
3. Correct for study artifacts.
4. After artifact investigation, check for variation across study findings: if large discrepancies exist, search for moderator variables.
5. Statistically combine results from studies.

The first task of a traditional meta-analysis is to draw out the important variables that are relevant to the research question. In this adaptation of meta-analysis, this is straight-forward, since all the mini-studies are designed specifically to investigate the basic comparability of items between language formats.

The next step of the meta-analysis is to express the findings of the individual studies in a common metric. Typically, results from statistics (such as a correlation, T-test, Chi-square, etc.) can easily be converted to either the "d" statistic or a correlation (see Appendix 2 for conversion formulas). This works well for the proposed

psychometric analyses. However, finding an effect size or correlation with coded or categorical data, as in the expert-dependent studies, is slightly more difficult.

In this particular study, each of the comparability studies needs to have an effect size or correlation to compare with the others. Below are the details of how this will be accomplished for each phase of the procedure:

1. *Forward translation*—to generate the effect size, the average rating for each item is taken. This average is across rating categories and raters. The effect size is a t-statistic, with each item characteristic evaluated for its distance from a perfect match for each items and all raters. The standard deviation is based on the pooled within rater variance. The following formula is used:

$$t = \frac{\overline{X}_i - 1}{\sigma}$$

   In effect, rating scales are computed as follows: value for item, subtracted by 1 (null or reference of "equally comparable for each group"), divided by the standard deviation for all items across the categories. Because of the small sample size (3 raters), this information will highly sensitive to individual characteristics of the raters such as native tongue or potentially gender.

2. *Back translation*—as in forward translation (1), each item is evaluated by a testing expert after being back-translated from target to source. The effect size, similarly, is generated in the same manner as forward translation (1 above). It is similarly also subject to rater characteristics, as are all expert-dependent methods.

3. *Simple descriptive statistics*—p-values (probability of correct response) can be used to create an effect size measures between the language versions. This is done in a similar manner to both 1 and 3 above. The formula is nearly identical:

$$t = \frac{\overline{X}_i \text{ exp} - \overline{X}_i con}{\sigma}$$

In this formula, the control group p-value (US) is subtracted from the p-value for the experimental groups (Italy), which is divided by the pooled standard deviation of the groups.

4. *DIF assessment and analysis*—an effect size is computed using Zumbo's R-squared method (1999), as described above in the psychometric methods section. The square root is a comparable measure for meta-analysis which can easily be converted to a *d* value for inclusion in an effect size meta-analysis.

In summary, each of the above analyses will be summarized as a relationship or group difference between the language versions of the translated/adapted test. Table 3.4 below shows the resultant effect size from each analysis, how it is derived, along with comments on how these effects are be used to illustrate comparability of language versions of the translated/adapted test.

Table 3.4 Summary of effects generated by mini-studies.

Study type	Effect size	Derivation	Comments
Forward translation Back translation	*d*	5 level unipolar coding used with reference point of 1, divided by s.d. of coding across all items.	*d* will be used in the MA
Simple descriptive	*d*	mean difficulty (p-values) for test items using a t-test of difference	*d* will be used in the MA
DIF analysis	*d*	logistic regression r^2 (Zumbo 1999)	square root for r value, then conversion to *d* for MA

Meta-analytic studies may encounter several artifacts that may alter the value of the outcome measures. Many of these artifacts may be dealt with using existing meta-

analysis techniques. Some examples of correctable artifacts are: a) sampling error, b) error of measurement of the dependent or independent variable, c) dichotomization of a continuous dependent or independent variable, and d) range variation.

Sampling error and error of measurement are present in nearly every study (Hunter & Schmidt, 1990), and thus they will be accounted for in the present study as follows:

Sampling error, (p. 108) is estimated via the following formula:

$\sigma^2{}_\rho = \sigma^2{}_r - \sigma^2{}_e = \sigma^2{}_r - (1-\bar{r}^2)^2/(\overline{N}-1)$, where

$\sigma^2{}_\rho$ is the variance present in the population correlation

$\sigma^2{}_r$ is the observed variance of the correlation

$\sigma^2{}_e$ is the error in the variance of the correlation

$\bar{r}$ is the average correlation in observed studies

N is the sample size

The error of measurement (p. 117) is accounted for in the following manner:

$r_c = \rho_c + e_c$, where e_c is the sampling error in the corrected correlation r_c, and ρ_c is population value for true correlation without error in measurement. The value r_c is found as follows:

$r_c = \dfrac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$, where r_{xy} is the correlation and r_{xx} and r_{yy} represent reliabilities. Next,

e_c can be obtained from: $e_c = \dfrac{e}{\sqrt{r_{xx}\sqrt{r_{yy}}}}$, in which e is the sampling error in r_{xy}. Thus,

by subtraction, the population value for true correlation can be derived as follows:

$$\rho_c = r_c - e_c .$$

Moderator variables can be tested for with meta-analysis. If moderator variables are present, a common solution is to split studies into subsets and analyze separately before combining them together. One such possibility (of potential theoretical importance as well) would be the split in analysis types between psychometric methods and expert-dependent methods. The outcomes between the categories of the studies can be tested with a Chi-square (2×2, 1 df) between the studies of the various categories and whether they have significant findings or not. Alternatively, since there are no existing studies of this nature, another approach is to perform a cluster analysis on each item to see which study types go together.

The final step to each meta-analysis is to combine the effects from each mini-study. This has commonly been done using either a Fisher, Winer or Stouffer combined test, depending on the metrics and variables involved. However, many methods exist for this final step of meta-analysis. The most appropriate for the present situation is either a Glassian (Glass, 1976) approach or that proposed by Schmidt and Hunter (Schmidt & Hunter, 1977), since both primarily focus on effect sizes (as *r*'s or *d*'s), rather than p-values as a means for comparison. The decision between the two depends on simplicity of calculation (Glassian) or accounting for sampling error, unreliability and range restriction (Schmidt & Hunter). In the present study, a combination of the two is used. This is due to the use of META-Programs software (Schwarzer, 1989). META automatically generates the results of a weighted integration method as well as a random effects model. Both results have been used to derive the end effect sizes that are

reported, since the weighted integration method works only when homogeneity is not violated, and random effects models are only necessary when that assumption is violated.

Comparability Criteria and Data Analysis

Once the meta-analyses are computed, the next step is to interpret the results. The data generated by the above analyses suggests two areas of focus, namely, the means and standard deviations of the results. Additionally, there are also two natural sub-groupings within the data that lend themselves to further analysis; these groups are the items themselves and the methods of analysis. Hence, items are first categorized in terms of their effect sizes and then their standard deviations.

Since there is no current literature that points to what effect-size might be of note in cross-lingual testing, the investigation follows the general lead of Cohen's generalizations for effect sizes (Cohen, 1992), where effects of 0.2, 0.5 and 0.8 and classified as small, medium and large, respectively. Following this categorization, items and methods found to be representative of a particular category are further investigated for similar content or substantive explanation. After the examination of the effect sizes, items and methods are also investigated in terms of their standard deviations, noting similar content or substantive explanation for greater of lesser amounts of deviation.

CHAPTER 4: RESULTS

Introduction

The results of this study are broken down into the four major sections of the research. Results begin with 1) Expert-dependent analysis results, continue with 2) Psychometric analysis results, then 3) Meta-analysis results, and finally 4) a report of where the items fall based on comparability criteria previously established.

Expert dependent analysis

Forward translation.

The forward translation results found in Table 4.1 detail the effect sizes for each item are detailed below. The ID tags indicate the label for each item based on its content and ordering within the content grouping. Thus, item S1 is the first ordered item in the spatial section of the instrument, Q6 is the sixth item in the quantitative section, etc. The following columns contain the effect size for each item, averaged across raters for each category (content equivalence, construct equivalence, etc.). Finally, the combined (average) effect size across rating categories for each item is in the last column. All characteristics are given equal weighting since there is no research indicating which of these factors might be more or less important. However, it is logical to combine the characteristics, as they are all components of a greater measure of comparability.

Table 4.1 Forward Translation Results

Forward Translation

ID	*Content*	*Conceptual (Construct Exists)*	*Conceptual (Construct Translates)*	*Original Intent*	*Linguistic Equivalence*	*Item Presentation*	*Holistic Equivalence*	*Combined: d*
S1	0.57	0.00	0.00	0.00	1.35	0.99	0.00	0.42
S2	0.00	0.00	0.00	2.57	1.35	0.00	3.69	1.09
S3	0.00	0.00	0.00	0.00	2.03	0.00	0.00	0.29
S4	0.00	0.00	0.00	0.00	1.35	0.99	0.00	0.33
S5	0.00	0.00	0.00	0.00	1.35	0.99	0.00	0.33
S6	0.00	0.00	0.00	0.00	1.35	0.00	0.00	0.19
S7	0.00	0.00	0.00	0.00	2.03	0.00	0.00	0.29
S8	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.10
S9	1.70	0.00	0.71	0.64	1.35	0.99	0.00	0.77
Q1	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.10
Q2	0.00	0.00	0.00	1.28	0.68	0.99	0.92	0.55
Q3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q4	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.10
Q5	0.00	0.00	1.43	0.00	0.68	0.00	0.00	0.30
Q6	4.52	0.00	3.57	2.57	0.00	0.00	0.00	1.52
T1	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.10
T2	0.00	0.00	0.00	0.00	0.68	0.99	0.00	0.24
T3	0.00	0.00	0.71	0.00	0.68	0.00	0.00	0.20
T4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T5	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.10
T6	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.14
T7	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.14
M1	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.14
M2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M4	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.10
M5	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.10
M6	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.14
M7	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.14
M8	0.00	0.00	0.00	0.00	0.68	0.99	0.00	0.24
M9	0.00	0.00	0.00	0.00	0.00	1.97	0.00	0.28
M10	0.00	0.00	0.00	0.00	0.00	1.97	0.00	0.28
M11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P7	0.00	2.03	0.00	0.00	2.03	0.00	0.00	0.58
P8	0.00	2.03	0.00	0.00	0.00	0.00	0.00	0.29
P9	0.00	0.00	1.43	0.00	0.68	0.00	0.92	0.43
P10	0.57	2.03	0.71	0.64	1.35	0.00	0.00	0.76
P11	0.57	2.03	1.43	0.64	0.68	1.97	0.00	1.05
P12	0.57	0.00	0.71	0.00	0.00	0.99	0.00	0.32
P13	2.83	0.00	1.43	0.64	1.35	0.00	0.92	1.02
average	0.24	0.17	0.25	0.19	0.55	0.39	0.13	0.27
stdev	0.80	0.57	0.65	0.56	0.64	0.60	0.57	0.34

Across all items, it can be noted that, among the rated categories, linguistic equivalence shows the largest effect size (0.55) while content shows the greatest standard deviation (0.80). This is potentially due to two factors. First, it is somewhat predictable that linguistic equivalence as a category should be slightly larger since the languages are indeed different from one another. Secondly, the greater standard deviation in the content related effect sizes shows that there tend to be items of very high comparability and very low comparability, since content overall shows a smaller effect size of 0.24. After linguistic equivalence, item presentation comes next; indicating possibly that the way items are presented is the second most important factor to the expert judges. The other factors seemed to be of lesser importance to the raters, with relatively smaller average effects. The degree of agreement between raters was fairly high, with Holsti's Coefficient of Reliability at 0.855 (where 0.8 and higher is a good degree of reliability).

Back translation results.

The back translation results are displayed in the same format as the forward translation results from the previous section.

Table 4.2 Back Translation Results.

	Back Translation							
ID	Content	Conceptual (Construct Exists)	Conceptual (Construct Translates)	Original Intent	Linguistic Equivalence	Item Presentation	Holistic Equivalence	Combined: d
S1	0.00	0.00	0.00	0.00	1.50	1.50	0.00	0.43
S2	0.00	0.00	0.00	5.98	1.50	0.00	5.98	1.92
S3	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
S4	1.50	0.00	0.00	0.00	1.50	1.50	1.50	0.85
S5	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
S6	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
S7	0.00	0.00	0.00	0.00	2.99	0.00	0.00	0.43
S8	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.21
S9	0.00	0.00	1.50	0.00	1.50	1.50	1.50	0.85
Q1	0.00	0.00	0.00	0.00	2.99	0.00	0.00	0.43
Q2	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
Q3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q4	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
Q5	0.00	0.00	0.00	0.00	2.99	0.00	0.00	0.43
Q6	0.00	0.00	1.50	0.00	1.50	0.00	1.50	0.64
T1	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
T2	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
T3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T4	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
T5	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
T6	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.21
T7	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.21
M1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M6	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.21
M7	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.21
M8	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.21
M9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P7	0.00	1.43	1.50	1.50	1.50	0.00	1.50	0.85
P8	0.00	1.43	0.00	0.00	1.50	0.00	0.00	0.21
P9	4.49	0.00	1.50	1.50	1.50	0.00	1.50	1.50
P10	0.00	1.43	0.00	0.00	1.50	0.00	0.00	0.21
P11	0.00	1.43	0.00	0.00	1.50	0.00	0.00	0.21
P12	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.21
P13	0.00	0.00	0.00	4.49	1.50	0.00	1.50	1.07
Average	0.12	0.12	0.12	0.28	0.84	0.28	0.31	0.28
stdev	0.68	0.40	0.42	1.10	0.92	0.59	0.97	0.40

These back translation results closely parallel those of forward translation in many regards. Again, the largest effect size belongs to linguistic equivalence, which is to be expected given that we are dealing with different languages. However, in this instance, the back-translation yields slightly different results for measures of deviation. In this case, the original intent is the category that shows the greatest variation. The other measures of comparability seem to indicate that holistic comparability gains a greater relationship to linguistic equivalence, most likely due to this expert's background in linguistics. Item presentation sinks to fourth in relative ranking, again emphasizing this expert's knowledge of linguistics over cultural presentation of items. The other effects change little in comparison to their relative ranking in forward translation. In sum, it should be reiterated that the natural tendency of any expert-dependent method with relatively small sample sizes would be influenced by rater characteristics.

Psychometric analysis

The psychometric/statistical findings for the item analyses are detailed below in subsections on the simple descriptive statistical findings and the logistic regression findings.

Simple descriptive statistics.

The results for the comparability of simple descriptive statistics across languages are reported below in Table 4.3. This table includes p-values for each language group by item, as well as the associated effect size or *g*-statistic (Hedges and Olkin, 1985, p. 80), which is a biased estimator of the *d*-difference statistic between the language groups.

Table 4.3 Simple Descriptive Statistical Results.

Descriptive Statistics

ID	*p-value Italy*	*p-value USA*	*effect*
S1	0 84	0.81	0.04
S2	0.83	0.78	0.06
S3	0.96	0.86	0.11
S4	0.96	0 96	0.01
S5	0 91	0.92	-0.01
S6	0.56	0.80	-0.29
S7	0.89	0.86	0.04
S8	0.99	0.97	0.02
S9	0.80	0.84	-0.04
Q2	0.94	0.94	0.00
Q3	0.89	0.90	0.00
Q4	0.53	0.70	-0.21
Q5	0.93	0.90	0.03
Q6	0.89	0.81	0.08
T1	0.33	0.38	-0.09
T3	0.67	0.62	0.07
T4	0.69	0.80	-0.13
T5	0.68	0.63	0.05
M1	0.72	0.91	-0.21
M2	0 61	0.76	-0.18
M4	0.77	0.78	-0.01
M5	0.24	0.44	-0.34
M6	0.12	0 32	-0.43
M7	0.88	0.82	0.06
M8	0.32	0.32	0.00
M9	0.95	0.91	0.03
M10	0.84	0.83	0.01
M11	0.59	0.73	-0.17
M12	0.93	0.94	-0.01
M13	0.53	0.68	-0.20
P1	0.80	0.81	-0.01
P2	0.63	0.71	-0.10
P3	0.49	0 49	-0.01
P5	0.77	0.88	-0.12
P6	0.27	0.28	-0.01
P7	0.48	0.63	-0.19
P8	0.51	0.32	0.29
P9	0.44	0.31	0.21
P10	0.27	0.23	0.08
P11	0.91	0.68	0.25
P13	0.65	0.52	0.18
Average	0.68	0.70	-0.03
Stdev	0.24	0.22	0.15

Logistic regression results.

Results for the logistic regression DIF analysis of each item are reported below in Table 4.4. The table contains the three model progression as advocated in Zumbo's R-squared methodology (1999): 1) the model with intercept only, then 2) add group (language of test item) and total score modeled, and finally 3) add an interaction term to

the model. Following the model details, the DIF R-squared generated by the procedure is presented, with the final column the conversion from R-squared to *d* for inclusion in the subsequent meta-analyses.

Table 4.4 LR-DIF Results.

Logistic Regression Results

ID	*intercept only*	*group and total*	*grp/ttl/interaction*	*DIF R-square*	*d--from r to d conversion*
S1	67.73	69.92	69 92	0.02	0.30
S2	56.64	61.17	61.39	0.05	0.45
S3	58.81	85.69	86.50	0.28	1.24
S4	66.79	66.95	72.83	0.06	0.51
S5	61.84	61.93	62.04	0.00	0.09
S6	96.45	140.14	140.24	0.44	1.77
S7	74.09	77.33	77.33	0.03	0.37
S9	101.70	102.77	102.97	0.01	0.23
Q2	66.29	66.31	66.37	0.00	0.06
Q3	82.23	82.23	84.09	0.02	0.28
Q4	84.64	101.95	102.30	0.18	0.93
Q5	54.42	56.45	60.87	0.06	0.53
Q6	67.65	77.81	77.84	0.10	0.67
T1	101.36	101 74	105.02	0.04	0.39
T3	149 47	155.73	156.04	0.07	0.53
T4	60.17	70 28	71.15	0.11	0.70
T5	93.80	97 27	97.42	0.04	0.39
M1	101.01	147.26	147.56	0.47	1.87
M2	154.42	170.30	171.35	0.17	0.90
M4	115.30	115.37	115.38	0.00	0.06
M5	146.53	171.36	171.95	0.25	1.17
M6	101.54	136.19	138.32	0.37	1.53
M7	89.96	95.64	98.46	0.09	0.61
M8	150.16	151.51	156.86	0.07	0.54
M9	41.48	45.11	46.59	0.05	0.46
M10	100.17	100.71	100.72	0.01	0 15
M11	90.79	103.17	103.26	0.12	0.76
M12	19.73	19.78	20.95	0.01	0.22
M13	66.37	81.09	83.59	0.17	0.91
P1	110.62	110 67	110 67	0.00	0.05
P2	123.54	126.58	126.87	0.03	0.37
P3	112.76	113.25	113.43	0.01	0.16
P5	44.98	58.04	58.09	0.13	0.78
P6	89 35	89.79	90.05	0.01	0.17
P7	101.43	112.47	113.07	0.12	0.73
P8	74.62	109.16	109.69	0.35	1.47
P9	85.57	105.49	110.43	0.25	1.15
P10	49.25	52 55	53.93	0.05	0.44
P11	15.14	72.17	77.21	0.62	2.56
P13	120.99	142.33	148.36	0.27	1.23
Average					0.69

On the whole, the instrument appears to be slightly incomparable, with an effect size of 0.69 from the reference of being completely equivalent or comparable. But as noted by Zumbo (2003), instrument-level analysis can subsume interesting and often important item level features. Hence, we should view this conclusion somewhat

skeptically and the item-level analysis that follows will be more critical to assessing the instrument's cross-lingual comparability.

Meta-analyses

As a reminder, these are the steps carried out as part of the meta-analytic results that follow.

1. Establish the basic facts (variables and values) surmised from the above mini-studies.
2. Express key findings in a common statistic (correlation or d "difference" statistic for effect sizes).
3. Correct for study artifacts.
4. After artifact investigation, check for variation across study findings: if large discrepancies exist, search for moderator variables.
5. Statistically combine results from studies.

The basic facts are contained in the previous subsection of expert-dependent and psychometric results. All the above results were subsequently converted to *d* statistics to have a common metric. The META program (Schwarzer, 1989), used for these analyses, corrects for study artifacts, such as measurement error and/or sampling error. The search for moderator variables was not a principal component of this dissertation, though the analysis can be found in Appendix 3. Finally, the results of statistical combination of the above mini-study findings are presented in combined form below in Table 4.5.

Table 4.5 Meta-analytic Results.

		Meta-analyses results		
Item	*D*	*observed var*	*error var*	*pop var*
S1	0.22	0.03	0.02	0.01
S2	0.85	0.65	0.03	0.62
S3	0.47	0.27	0.02	0.25
S4	0.40	0.12	0.03	0.10
S5	0.07	0.02	0.02	0.00
S6	0.48	0.79	0.02	0.76
S7	0.23	0.03	0.02	0.00
S9	0.41	0.18	0.03	0.16
Q2	0.15	0.06	0 02	0.04
Q3	0.12	0.02	0.02	0.00
Q4	0.27	0.23	0.02	0.20
Q5	0.30	0.05	0.02	0.02
Q6	0.70	0.35	0.03	0.32
T1	0.15	0.04	0.02	0.02
T3	0.23	0.06	0.02	0.03
T4	0.21	0.13	0.02	0.11
T5	0.21	0.02	0.02	0.00
M1	0.46	0.91	0.02	0.88
M2	0.20	0.24	0.02	0.21
M4	0.03	0.00	0.02	0.00
M5	0.25	0.41	0.02	0.39
M6	0.38	0.65	0.02	0.62
M7	0 28	0.06	0.02	0.04
M8	0.25	0.05	0.02	0.02
M9	0.21	0.05	0.02	0.02
M10	0.09	0.02	0.02	0.00
M11	0.24	0.17	0.02	0.15
M12	0.09	0.01	0.02	0.00
M13	0.19	0.25	0.02	0.22
P1	0.02	0.00	0.02	0.00
P2	0.10	0.04	0 02	0.02
P3	0.07	0.01	0.02	0.00
P5	0.18	0.17	0.02	0.14
P6	0.07	0.01	0.02	0.00
P7	0.47	0.21	0.03	0.19
P8	0.59	0.36	0.02	0.33
P9	0.81	0.35	0.03	0.32
P10	0.35	0.09	0.02	0.06
P11	1.02	1.19	0.03	1.16
P13	0.86	0.22	0.03	0.19

All meta-analyses had four groups and 1510 data points (k=4, N=1510). The effects range from near perfect comparability (as in item P1 with an effect size of 0.02 and almost no variance) to large discrepancies in comparability indicators (e.g., item P11 with a measured effect size of 1.02 and large variance measures).

Comparability Criteria and Data Analysis

Finally, now that the data has been generated, the final analysis is to categorize the data, as proposed initially. This has been done in two ways. First, items are categorized in terms of their effect sizes and then their standard deviations. Then, the methods are categorized by effect sizes and standard deviations. This categorization has been done principally using Cohen's (1992) classification criteria, into small, medium and large effects. The individual item classification is detailed below in Table 4.6.

Table 4.6 Item Effect Size Classification.

Item	Cohen's Classification small (0.2) 0.0-0.3	medium (0.5) 0.3-0.7	large (0.8) 0.7+	Observed s
S1	0.22			0.03
S2			0.85	0.65
S3		0.47		0.27
S4		0.40		0.12
S5	0.07			0.02
S6		0.48		0.79
S7	0.23			0.03
S9		0.41		0.18
Q2	0.15			0.06
Q3	0.12			0.02
Q4	0.27			0.23
Q5	0.30			0.05
Q6			0.70	0.35
T1	0.15			0.04
T3	0.23			0.06
T4	0.21			0.13
T5	0.21			0.02
M1		0.46		0.91
M2	0.20			0.24
M4	0.03			0.00
M5	0.25			0.41
M6		0.38		0.65
M7	0.28			0.06
M8	0.25			0.05
M9	0.21			0.05
M10	0.09			0.02
M11	0.24			0.17
M12	0.09			0.01
M13	0.19			0.25
P1	0.02			0.00
P2	0.10			0.04
P3	0.07			0.01
P5	0.18			0.17
P6	0.07			0.01
P7		0.47		0.21
P8		0.59		0.36
P9			0.81	0.35
P10		0.35		0.09
P11			1.02	1.19

Following the item classification, we can get a feel for the instrument as a whole by showing how many items fall into the small, medium and large effect size categories. What is shown is that 26 of the 40 items have effects of negligible to small size, 9 items have effects in the medium category and 5 items fall into the large effect size category. The items exhibiting medium and large effect size characteristics will be examined further in Chapter 5.

Table 4.7 Frequency of Item Classifications

	Items Classified by Cohen's Standards		
	Small (0.2)	Medium (0.5)	Large (0.8)
Range of ES	0.0-0.3	0.3-0.7	0.7+
# of items	26	9	5

The variance of items and their ratings/measure of comparability ranges from near zero (0.00058) to quite large (1.18962), with an average variance of 0.212483. To categorize items in terms of having small, medium and large variances, the author has chosen the arbitrary cutoffs that parallel Cohen's pattern for effect sizes. Essentially, effect sizes of 0.0-0.3 are categorized as small, 0.3-0.7 are medium, and 0.7 and greater are deemed large in terms of their variance. The resultant classification is displayed below in Table 4.8.

Table 4.8 Item Variance Classification

	Item Variance Classified by Cohen's Standards		
	Small (0.2)	Medium (0.5)	Large (0.8)
Range of ES	0.0-0.3	0.3-0.7	0.7+
# of items	31	6	3

Similar to item effect size, items exhibiting medium and large variance in their effects sizes are discussed in Chapter 5.

Method Categorization

While the previous Table 4.8 refers to all the items present in the instrument, it behooves us to investigate how the methods classify these items as well. Hence, the methods that were used to assess the comparability of the instrument in this study were also categorized into Cohen's effect size categories as well as labeled by their variance. This was done in part to better understand which methods might be more likely to produce larger effects and which have greater variance. The results appear below in Table 4.9 below.

Table 4.9 Method Classification by Effect Size and Standard Deviations

Method	Methods Classified by Cohen's Standards: Small (0.2) 0.0-0.3	Medium (0.5) 0.3-0.7	Large (0.8) 0.7+	St Dev.
Forward translation	0.27			0.34
Back translation	0.28			0.40
Simple descriptive	-0.021			0.15
DIF analysis		0.69		0.56

What is immediately evident is that simple descriptive statistics nearly cancel out measures of incomparability, since their aggregate effect size is slightly negative for this instrument. In addition, both forward translation and back translation appear to be borderline medium in the effect size measures they produce across the entire instrument. Finally, logistic regression as a method is clearly the largest in terms of overall effects measured across all items in the instrument.[5] The variance categorization parallels this structure as well, where simple statistics show little deviation across items when aggregated. Simple and back-translation, as methods, are categorized as medium in

terms of effect size category along with LR-DIF, which has variance indicators near the top end of the medium category.

[5] The apparent disappearance of the large effects reported earlier is merely the result of averaging.

CHAPTER 5: DISCUSSION AND CONCLUSIONS

Summary of results

In this chapter, the results of the multiple meta-analyses and their subsequent findings are discussed. Items and methods both are described in terms of their variance, mean value and the categories into which they have been placed. Additionally, the theoretical implications of these findings upon methods for translation/adaptation are discussed.

Before continuing with the item level and method level analysis, the reader may be interested to know whether items of different content had differing levels of comparability. The spatial items appear to be somewhat unstable in their comparability, since of the 8 items, 4 items had a medium effect and 1 item had a large effect. In the problem solving section, 4 items showed a medium effect and 3 items showed a large effect. Over half of the items developed for the problem solving section show some signs of incomparability. However, 6 of these 7 items are story problem situations, all of which had at least a medium effect size. No doubt, this is due to the difficulty of translating rather lengthy items and trying to give culturally appropriate examples for examinees and test-givers. In contrast to other content areas, the memory problems had only two items with medium effects (out of 12 items), while the quantitative section had one item with a large effect (of 5 items) and time had no items (out of 4) categorized above small. The average effects by content are shown below in Table 5.1, which also illustrates the amount of variance in the effect sizes measured by content category.

Table 5.1 Average Effects by Content Area

Content Details

	Spatial	Quantitative	Time	Memory	Problem
mean	0.39	0.31	0.20	0.22	0.41
st dev	0.23	0.23	0.034	0.12	0.36

Despite the number of items with medium and large effect sizes, the mean effect sizes for each content category cannot be described as significantly different from each other (ANOVA, F=1.2813, p=0.296). This is notable, because it shows that there is not one particular type of item content that appears to be causing large-scale incomparability.

Interpretation of Results

While these content area results are good in that they do not show any one category of items as any less comparable than another, it is important, as indicated in the literature for the analysis to focus on the item-level results. It is also of interest, from a theoretical standpoint, to examine which methods tended to produce effects of differing sizes and variance.

Methods Discussion.

With respect to methods, the general pattern, as shown in chapter 4, runs from small to large effects in the following order: 1) simple descriptive with a small effect, 2) forward translation and back translation with a borderline medium effect, and 3) logistic regression with a borderline large effect. Similarly, the variance in the effect sizes for these methods followed a similar pattern. Simple descriptive statistics had a small amount of variance while the other three methods had a medium amount of variance.

There are several plausible explanations for the results in the case of the two expert-dependent methods. Since there was an effect of at least a small size or larger in the majority of items with these two methods, it would appear that there is the potential for hypothesis guessing, or at the least experts who thought it would be helpful to find incomparable items. Despite this potential problem, the expert dependent methods did not, in general, lead to extreme values. The potential advantage, then, is having experts who may act as a moderating measure on the more extreme values of LR-DIF.

The LR-DIF analysis has produced some rather extreme values (e.g. an effect size of up to 2.558). This is most likely due to LR-DIF taking into account total test score as a moderator variable. Consequently, the method is more sensitive to items that have an impact on total test score.

Another question relevant to the discussion of these methods is whether there is any connection between item difficulties (p-values) and the ratings produced by these methods. By simple correlation, we can see in Table 5.2 that there is no strong connection between p-values and the comparability ratings. The expert-dependent methods appear to have no relation whatsoever with item difficulty (forward translation corr = 0.8190 and back translation corr = -0.00302). Meanwhile, the psychometric methods have weak relationships with opposite results (simple statistics corr = 0.1811 and LR-DIF corr = -0.2040). Likely, this means that there is another source of variation being picked up by these comparability ratings.

Table 5.2. Correlation of techniques with p-values

	p-value
simple	0.181085
LR-DIF	-0.20402
forward	0.081902
back	-0.00302

Item Discussion.

On an instrument where 40 items were evaluated, there were 9 items with medium effects and 5 with large effects, meaning that 35% of the items evaluated warrant further review.

Table 5.3 Frequency for Item Effect Size Classification.

	Items Classified by Cohen's Standards		
	small (0.2)	**medium (0.5)**	**large (0.8)**
range of ES	**0.0-0.3**	**0.3-0.7**	**0.7+**
# of items	26	9	5

Item S2 (e.s.= 0.85), the second item in the spatial section was classified as having a large effect for a rather good reason. The text of the item reads as follows:

S2:

> Guarda questi animale che vanno in fila. Fammi vedere qual e il **secondo** animale.
>
> Look at the animals walking in a line. Point to the animal behind the **second** animal.

The clear difference between the English and Italian versions is a consequence of translation error. The Italian children are asked to point to the second animal, rather than the animal behind the second animal.

The following items, (S3, S4, S6, S9) all had medium sized effects. Item S3 had slight linguistic differences, but little statistical differences (p-values). However, the LR-DIF effect was quite large (1.24), which generally leads one to a further review of content. However, there does not appear to be a reasonable explanation as to what may be functioning differently with this item. Item S4 with marked as differing in terms of content, language, and presentation, but again had virtually zero difference in p-value and a relatively small LR-DIF value (0.51). The issues with this item are that the item, despite its slight linguistic differences, would be seen differently by the two cultures. The Italian experts claim that the old-fashioned cars would be distracting to young Italians. Item S6 was not picked by experts, but the statistical properties show this item as having a detrimental effect on Italian test takers (Italy p=0.56 vs. US p=.80, LRDIF=1.77). The experts could not explain this large gap. Finally, item S9 was flagged for differences in conceptual and linguistic inequivalence, as well as presentation differences. S9 also is not, however, flagged by statistics. The item was perhaps singled out due to its apparent complexity, which can be seen by looking at the full text:

S9:

> Look at the words dog, flower and tree on the three small cards and look at the lines on the one big card. Place the word dog on the **middle line**, place the word flower on the **right side** of the word dog, and place the word tree on the **left side** of the word dog.

This item clearly has significantly more text than do the previous items which consist of one or two simple sentences.

In item Q6 (e.s.= 0.70) the phrasing of the item in Italian was confusing to all three experts. In fact, all three asked to have the item explained to them so that they could answer it correctly. This could be an indication of a more syntactically complex item in Italian. One might think that the likelihood of children answering this item correctly, when three language professionals could not easily, would be low. However the actual p-values (Italy=0.89 and US=0.81) would seem to contradict this.

Q6:

Guardi i numeri scritti sul cartoncino. Nella sua classe Marco e il quarto in ordine di eta. Fammi vedere il numero che indica **quanti bambini** sono piu grande di Marco.

Look at the numbers on this card. Don is the fourth oldest child in his class. Point to the number that shows **how many children** are older than Don.

Options are: 3, 4, or 5

With the two memory questions that exhibited a medium effect, M1 and M6, it is hard to conceive of why the may function differently. Item M1 is a repeated after me type, with the series: "8, 9, 1, 7, 4". M6 is a repeat after me at the sentence level, e.g. "My dog chases the white cat." Interestingly, both of these items come after new instructions for the children. Perhaps this is an indication that the instructions and the attention given them needs to be reasserted.

Items P7 and P8 also had medium effects. In both instances, the p-values have a gap. On P7, the US children performed much better (Italy p=0.48 US p=0.62, LR-DIF=0.73), while the Italian children performed much better on P8 than the American children (Italy p=0.51 US p=0.32, LR-DIF=1.47). The Italian experts report that having

a pet is not as common an experience for many Italian children as it appears to be for American children, which may help to explain the differences in P7. For item P8, where an imaginary child is stuck in a tree, the content appears to be identical. No cultural explanation was offered by the experts.

Item P9 (e.s.= 0.81) is a problem situation where children are asked to choose the most unusual way to get a ball stuck on a rooftop. The native experts said that Italian children would answer differently than American children given the following choices:

A. get an older person to help them

B. use a ladder to get the ball, or

C. each child can stand on the shoulders of another until they are tall enough to reach the ball.

According to the native speakers, Italian children would think it strange to ask somebody to help them get the ball when they could get it for themselves. Contrarily, it seems that most American children think it stranger to have three children stand on each other's shoulders to retrieve the ball. Despite this potential problem, the p-values show that Italian children were favored over American children in this question (Italy=0.44, USA=0.31). One possible explanation is that the items were keyed differently and already accounted for this cultural difference.

On item P10, there is another situation where an imaginary child in the photo is interacting with a dog. In this case, the children are to explain why the child appears to be angry at the dog. The value of the overall effect is most inflated by the forward-translation score (0.76) where the experts indicated that content, conceptual equivalence and translation, original intent and language were all sources for possible differences.

Again, the issue of pets not being as common an experience to Italian children was mentioned.

The potential problems for Item P11 (e.s.= 1.02) are not immediately clear. The item is, however, exceedingly easy for Italian children (p=0.90 Italy versus p=0.68 USA). One potential cause for difference may be the slightly higher level of specificity in the English language grading criteria.

P11:

> Say to the child, "This boy is sound asleep. What things could he have done to make him so tired? Think of as many things as you can that he could have done to make him so tired. (Some appropriate responses are: stayed up too late, ran fast, worked hard, played sports, played outside all day, did exercises, etc.)

For Item P13 (e.s.= 0.86) the difference is possibly from a cultural adaptation of the item. Instead of using carrot seed as in the English version, the Italian version refers to pumpkin seeds. Another potential explanation is the more sedentary lifestyle of American children as opposed to Italian children, and hence, a higher likelihood that Italian children have been involved in gardening or to have seen vegetable seeds. Lastly, another speculative theory could be that American children could are more accustomed to seeing highly processed food, and not necessarily the source of their food. Finally, and perhaps most interesting, the grading criteria are again more specific in the English version.

Items with larger degrees of effect size variance.

On item S6, a large degree of variance was identified. This was no doubt due to the logistic regression effect size being particularly large here. Perhaps the problem in this item, which favored American examinees over Italian ones, is that the word "people" is used in English and the word "bambini" (children) to describe the people in the pictures for S6. Taking this into consideration, the correct answer shows a decidedly older man with a moustache which may explain why Italian children did not choose the answer as often as their American counterparts. This DIF was not detected by the other methods to be as truly disadvantageous as it truly was, hence the great variation on this item.

For item M1, there is also a great deal of variation in the effects. Again, this is due to the LR-DIF flagging DIF and the other methods not. One possible explanation for this is that this is the first in the memory series. Perhaps this is an unfamiliar task or the instructions are not immediately clear to the examinees. This seems to be a good explanation as it happens twice in the memory section (items M1 and M6, both of which exhibit a medium effect); at the start of digit repetition section and at the beginning of the sentence repeat section.

Lastly, there is also an extreme value in the variance in the effect sizes for item P11. The LR-DIF again has a value that is a good deal larger than the other methods effects, even though forward translation flags this as a potentially problematic item as well. One of the native Italian informants opined that this item appears to give more clear explanation of potential responses to American test-givers.

Findings

After all these analyses, we are left to wonder: Did we gain anything from this method? Let us return to the principal questions of this project:

1. How does one generate a 'body of evidence', as called for in Cronbach and Meehl (1955), for the validity of a cross-cultural translation/adaptation?
2. Assuming that such evidence of cross-cultural adaptation/translation can be generated, is meta-analysis an effective methodology for the integration of the findings across methods within this body of evidence?
3. Within that body of evidence, which methods are likely to produce smaller or larger effects for the evaluation of cross-lingual/cultural comparability? Which of these methods are better suited for the assessment of cross-cultural translation/adaptation?
4. What type(s) of items or content are likely to trigger differences or difficulties in cross-cultural translation/adaptation?

A body of evidence that provides for the validation of translation/adaptation can be generated in any number of ways using a myriad of techniques. Some important factors play a part in this process: 1) the facility of data collection via the proposed method(s), 2) a decision regarding the kind of information to be collected, 3) how the information is collected, 4) deciding what information will be treated as supporting or degrading the degree of comparability of the instrument.

This dissertation demonstrates that this process can be done, and that such a body of evidence can be generated. Data collection was relatively easy for all methodologies involved in the analysis. The decision regarding which kind of information to collect was simplified by the author's decision to use MA as the method for integration. Thus, the

body of evidence was restricted to evidence that was easily included in an MA (i.e., easy conversion to r or d). The manner of collection of the information is dependent upon which methods are chosen. Finally, each individual study of method needs to be examined to determine which show evidence in support of comparability and which do not.

It can be noted that MA is a proven method for the integration of findings across multiple studies. However, how do we show that MA is an effective methodology for the integration of findings for a cross-cultural/linguistic comparability evaluation? One type of evidence is likely to be whether items can be identified as having problems. Further evidence would be the ability of a method to provide explanation for why such items are either comparable or not. As shown in the discussion above, this technique both flags items with problems, and given its reliance on expert-dependent methodology is potentially able to provide at least a partial explanation as to why an item is being flagged as non-comparable. In fact, nearly 80% items flagged as having medium to large effects were explained by the experts (11/14).

With respect to the different methodologies employed by this study, a correlation between LR-DIF and the simple descriptive statistics does not yield a high degree of relationship ($r = -0.0845$). However, the expert dependent methods do seem to have some relationship with LR-DIF (forward $r= 0.4477$ and back $r= 0.3107$), see Table 5.4, below. However, this is not concerning information. The fact that they do not produce redundant information is not a problem; rather it is a strength. It shows that the expert dependent methods more likely yield different results reflecting the language understanding of those experts.

Table 5.4. Correlations Between Comparability Methods.

	Simple	LR-DIF	Forward	Back
Simple	1.0000	-0.0845	0.4477	0.3107
LR-DIF	--	1.0000	0.1862	0.0540
Forward	--	--	1.0000	0.6516
Back	--	--	--	1.0000

Between the methods, it appears that logistic regression is most likely to produce a larger effect size, while simple statistics produce little overall measure. The two translation rating scales seem to produce a more moderate measure. One potential explanation is that the definition of what is a large effect size may vary between the methods, since there is an underlying assumption that effect sizes generated by the different methods are comparable and scale-equivalent. Another plausible explanation is that the T-test style simple statistics do not account for overall test score (ability) of examinees, which makes them somewhat less useful for determining real impact on the language groups. They may also be an inappropriate measure due to the potential for regression to the mean, which produces a canceling effect. Contrarily, LR-DIF seems to be a good choice for the assessment of comparability. This is because it has been shown that logistic regression results are similar to results of other DIF methods (Clauser & Mazor 1998), which have a long history of assessing group differences such as language background. One should note, however, that the greater variance indicates that LR-DIF may isolate extreme items and not pick out the subtleties teased out by forward and back-translation.

Since we have an explanation for nearly 80% items flagged, it seems only logical to try and understand the commonalties between those items. Table 5.5 details the items

that were flagged as having Medium and Large effect sizes, whether experts offered an explanation for the item incomparability, and the reason given.

Table 5.5. Sources for Item Incomparability.

Item	ES (M or L or V)	Explained	Reason Given
S2		Y	Translation error
S3		N	
S4		Y	Cultural difference; picture needs to be adapted
S6	V	Y	Inconsistent translation--> picture not adapted for translation
S9		Y	Complexity
Q6		Y	Syntactically complex translation; confusing
M1	V	Y	1st in series after instructions
M6		N	
P7		Y	Cultural difference; unfamiliar situation
P8		N	None
P9		Y	Cultural differences in approach to problem
P10		Y	Cultural appropriateness of situation; unfamiliar situation
P11	V	Y	Clearer explanations for American test givers; higher degree of grading specificity in American version
P13		Y	Adaptation possibly not appropriate. Changed examples, more/less specific grading criteria

The major causes for item incomparability appear to be of types: true linguistic/cultural differences and technical problems or extra-linguistic issues. Among the technical problems that can noted, there is a single case of a translation error and slight differences in the amount of detail present in the instructions for the language versions. Clearly, the translation glitch would results in different outcomes by not asking the same question. The amount of detail provided test-givers as well as examinees could also impact the outcomes on the test. With respect to the cultural issues, the problems can be labeled as cultural differences, culturally unfamiliar situations, and linguistic

differences. A cultural difference is the occurrence of differing cultural explanations or understandings of a term or different attitudes or approaches that are commonly shared. Culturally unfamiliar situations are those to which a member of a given culture either not generally presented with, or a situation in which the construct being tested is not perceived from the same world-view. Finally, linguistic differences are just that. They are the actual language differences as they manifest themselves on the test (e.g. syntactic complexity where an item in one language must necessarily be overly complex to explain an idea present in the other language).

Theoretical Implications and Connections to the ITC Guidelines

Certainly we have gained in our understanding of which measures and methods for producing these measures seem more appropriate for multilingual comparability assessment. In addition, we have also gained some understanding of what types of items and content may be likely to produce incomparable items across languages[6].

Another gain we have made as a result of this is yet another tool to aid the adaptation/translation process. Though the ITC guidelines detail individual techniques for assessing comparability, it is useful to know that these multiple techniques can be integrated to produce measures that flag items that are clearly problematic. The advantage of these multiple measures of comparability is that the technique provides for the nomological network that Cronbach and Meehl (1955) state is necessary as evidence of construct equivalence. The ITC Guidelines are excellent start to providing guidelines for the creation of such a nomological network. The techniques detailed in this work

[6] This information may be limited to the languages that were the focus of this investigation (Italian and English). Such findings are likely to differ for other languages.

show how that body of evidence can be generated, as well as how one might make objective decisions based on the information gathered.

Returning to the ITC guidelines, let us examine how the meta-analytic technique developed in this paper aids in the implementation of those guidelines. Recall that the ITC guidelines are concerned with appropriate translation and adaptation of multilingual instruments with respect to: 1) instrument context, 2) test development and adaptation, 3) test administration, and 4) score interpretation and documentation. This dissertation addresses components of each of these factors of comparability in different manners.

Let us begin with the impact of this method upon the instrument context. The context refers to the concerns that test developers should take into account regarding construct equivalence between the linguistic or cultural groups being tested. Tanzer and Sim study (1999) recommend that the guidelines be expanded to incorporate the acknowledgement of linguistic differences; this study does this directly by incorporating expert knowledge of bilingual experts. The questions raised in this section of the ITC guidelines are:

> "Is a construct being understood the same way by all linguistic and cultural groups? Is there any overlap of *definitions* of the construct in language/culture groups and the overlap in the *actual manifestation* of the construct in the language/cultural groups (Tanzer and Sim, 1999)?"

This study directly assesses the question of construct overlap by means of expert dependent codification. Specifically, the experts determine whether the construct tested in each item exists and is meaningfully actualized in the test languages and cultures.

With respect to the test development and adaptation section, the author's previous work involving the selection and definition of who is a language expert helps in the evaluation stage of this instrument (Joldersma, 2004). By selecting qualified experts to translate and evaluate the instrument, this method addresses this guideline. The documentation provided by the test developers (see Claxton, 2003) provides additional evidence of the quality of this translation. An additional component of this section is the selection of appropriate statistical method for analyzing score equivalence or comparability. Though many techniques exist, this study had the limitation of needing to select methods that were able to be converted to an effect size measure for inclusion in the meta-analysis. Lastly, Hambleton writes that test developers need to use systematic and judgmental evidence (including both linguistic and psychological examination) to aid comparability and provide linguistic and cultural validity for test users' inferences. This method performs this task directly be having experts perform such an evaluation.

The test administration guidelines suggest how best to administer a test or instrument to multiple languages or cultural groups. This includes taking into account item formats, time allotted, etc. that should be handled differently based on cultural expectations or needs. In this instance, the only evaluation that could be performed was an assessment of the item formats and task familiarity by the experts.

Finally, according to Hambleton (2001), the score interpretation and documentation guidelines are essential for documenting evidence for the validity of the adaptation. Certainly the efforts to include expert input into the results of this method show that meta-analytic methodology can aid in score interpretation and the documentation of adaptations as well. First, scores will more easily be interpreted with a

MA result that includes both expert and statistical information. This is because the information includes the expert results that indicate problematic items as well as the statistics that verify these judgments. Additionally, by going through each of these multiple analyses, a developer is providing excellent documentation for the validity of the adaptation process and how their instrument will function appropriately in the target cultures.

Limitations

The study is likely open to various limitations, be they measurement problems, statistical shortcomings (such as power), or threats to internal or external validity. Each of portions of the study (expert dependent methods, psychometric methods and meta-analysis methods) is subject to its own shortcomings.

Content validity for the study is supported by the fact that the mini-studies all cover similar content of items, since they come from the same adapted/translated test. However, this is in essence, what the study seeks to make an argument for, rather than a pre-existing characteristic of the study.

Criterion-related validity for the study is very much a part of the purpose of the study, addressed by concurrent validity. This concurrent validity is essentially what is sought by the endeavor; i.e. to find out whether the instrument is comparable (e.g., do the findings of the mini-studies/analyses coincide?). The study relies on meta-analysis to provide additional evidence if there are disagreements in the findings of the individual studies, and thus, draw the appropriate conclusion.

Construct-related validity is often established in three steps which check to see whether: 1) the variable is defined clearly, 2) hypotheses, based on a theory of the

underlying variable, are formed about how people who possess a "lot" versus a "little" of the variable will behave in a particular situation; and 3) the hypotheses are tested both logically and empirically. In the present study, this is established as follows:

1) Each variable is defined as essentially having to do with the comparability of some characteristic of each item
2) People of purportedly equal ability levels will perform equally well on the items. Items may thusly be said to have a "lot" of comparability versus "little" comparability.
3) The items' comparability is tested both logically and empirically using a mixture of commonly accepted techniques, and an application of some of these to newer measurement situations.

Naturally, one additional limitation is that this is only a comparison of English to Italian, but the extension to multiple languages should prove to be an interesting future project. On a related note, the implications of children in Italy having familiarity with English (or the source language of a translated/adapted test) was not accounted for in this study, since that type of demographic information was not recorded.

Future Research

The extension of this technique to other and or multiple languages, as opposed to two only is a next step for this research. Also, it will prove interesting to reevaluate the rating scales employed by expert raters. This will help improve the coding system for more accurate evaluation of comparability. Anecdotally, the native speakers (as opposed to the high-functioning non-native) tended to be harsher in their judgments of the item comparability. A rating scale analysis should provide evidence of this and show which

categories are utilized or underutilized. One potential solution for this issue, if this is the case, is to reduce the rating scale from a 5 point to a 3 point scale to bring it closer in line with the range (1 or 2 standard deviations) of effects observed in the psychometric data. Another solution would be to attempt transformations of the effects (e.g., a log function) to produce similarly scaled item. Another line of potential inquiry includes the investigation of other effect sizes which may produce better and/or more appropriate measures for this type of analysis. A route to this analysis would be to test various models predictive capacity using regression techniques.

In the end, despite the systematicity of meta-analysis, comparability is not really a question of statistical certainty, but rather a collection of evidence and a thinking aid that provides support for the inference of the degree to which a multilingual instrument's measures are comparable. However, the consequences of ignoring comparability for multilingual tests are too great to ignore because there is the risk of potentially invalid inferences made from these instruments without proper analyses.

Appendices

Appendix 1. The ITC Guidelines for Test Translation and Adaptation

Context

C.1: Effects of cultural differences that are not relevant or important to the main purposes of the study should be minimized as much as possible.

C.2: The amount of overlap in the constructs 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 are intended.

D.2: Test developers/publishers should provide evidence that the languages 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 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.

D.7: Test developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the test, and (2) identify problematic components or aspects of the test that 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: nonequivalent questions between versions intended for different populations should not be used in preparing a common scale or in comparing these populations. However, the 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.

A.2: Test administrators should be sensitive to a number of factors related to 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 should be made as similar as possible across populations for whom the test is intended.

A.4: Test administration instructions should 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 test and its administration that require scrutiny in the application of the test 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 the test should be followed.

Documentation/Score Interpretations

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

I.2: Scores differences among samples of populations administered the test 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 on the test, and should suggest procedures to account for these effects in the interpretation results.

Appendix 2. Conversion of Statistics to Common Metrics

(Wolf 1986, 35)

Guidelines for converting various statistics to *r*.

Statistic to be converted	*Formula for Transformation to r*	*Comment*
t	$r = \sqrt{\frac{t}{t^2 + df}}$	
F	$r = \sqrt{\frac{F}{F + df(error)}}$	Use only for comparing two group means (i.e., numerator df=1)
χ^2	$r = \sqrt{\frac{\chi^2}{n}}$	n = sample size Use only for 2 X 2 frequency tables (df = 1)
d	$r = \frac{d}{\sqrt{d^2 + 4}}$	

Guidelines for converting various statistics to *d*.

Statistic to be converted	*Formula for Transformation to r*	*Comment*
t	$d = \frac{2t}{\sqrt{df}}$	
F	$d = \frac{2\sqrt{F}}{\sqrt{df(error)}}$	Use only for comparing two group means (i.e., numerator df=1)
r	$d = \frac{2r}{\sqrt{1 - r^2}}$	

Appendix 3. Moderator Variables: Search Results and Percent of Variance Accounted for by Meta-Analyses.

Item	cluster details	k	N	*d*	so	se	sp	Q	Q p-value	% var exp
C12		4	1510	0.22206	0.03256	0.02441	0.00815	8.21089	0.04185	74.96
	cluster 1,2,4	3	851	0.32567	0 00477	0.03049	0 00000	0.51750	0.77201	100.00
	study 3			0.0355						
C13	no clusters	4	1510	0.84791	0.65284	0.03062	0.62222	65.30043	0.00000	4.69
C14		4	1510	0.47463	0.27211	0.02438	0 24773	100.07340	0.00000	8.96
	cluster 1,2,3	3	851	0.14171	0.00797	0.03006	0.00000	0.78889	0.67405	100.00
	study 4			1.23779						
C15	no clusters	4	1510	0.39620	0.12312	0.02510	0.09801	27.32686	0.00001	20.39
C16		4	1510	0.06931	0.02200	0.02412	0.00000	3.20755	0.36072	100.00
C17		4	1510	0.48048	0.78622	0.02463	0.76159	285.97935	0.00000	3.13
	cluster 1,2	2	192	0.20199	0.00020	0.04189	0.00000	0.00479	0.94484	100.00
	study 3			-0.2666						
	study 4			1.76538						
C18		4	1510	0.23027	0.02861	0.02429	0.00431	9.77350	0.02059	84.92
C20		4	1510	0.41085	0.18327	0.02572	0.15754	25.77918	0.00001	14.04
	cluster 1,2		192	0.80579	0.00351	0.04529	0.00000	0.07749	0.78073	100.00
	study 3			-0.0426						
	study 4			0 22656						
C22		4	1510	0.15026	0.06122	0.02438	0.03684	6.69141	0.08241	39.83
	cluster 2,3,4		1414	0.03928	0.01241	0.01807	0.00000	1.07575	0.58399	100.00
	study 1			0.55306						
C23		4	1510	0.11732	0.01918	0.02392	0.00000	7.11574	0.06830	100.00
C24		4	1510	0.26544	0.22821	0.02415	0.20406	99.40218	0.00000	10.58
	cluster 1,2		192	0.15387	0.00673	0.04182	0.00000	0.16093	0.68830	100.00
	study 3			-0.21351						
	study 4			0.92614						
C25		4	1510	0.30078	0.04653	0.02433	0.02220	20.41396	0.00014	52.29
	cluster 1,2,4		851	0.48679	0.01288	0.03039	0.00000	1.16150	0 55948	100.00
	study 3			0.02583						
C26		4	1510	0.70109	0.34676	0.02772	0.31904	51.10009	0.00000	7.99
	cluster 2,4		755	0.66820	0.00068	0.02520	0.00000	0.02706	0 86934	100.00
	study 1			1.52266						
	study 4			0.67372						
C27		4	1510	0.15210	0.03979	0.02401	0.01578	18.32378	0.00038	60.34
	cluster 1,2		192	0 15387	0.00673	0.04182	0.00000	0.16093	0.68830	100.00
	study 3			-0 08812						
	study 4			0.38977						
C29		4	1510	0.23207	0.05503	0.02401	0.03102	18.85221	0.00029	43 63
	cluster 1,2,3		851	0.07832	0.00993	0.02990	0.00000	0.48934	0.78296	100.00
	study 4			0.53049						
C30		4	1510	0.20774	0.13318	0.02406	0 10911	56.10190	0.00000	18.07
	cluster 1,3		755	-0.11220	0.00829	0.02391	0.00000	0 34676	0.55595	100.00
	study 2			0.21372						
	study 4			0 70215						
C31		4	1510	0.21095	0.02205	0.02401	0.00000	9.20296	0.02671	100.00
C34		4	1510	0.45753	0.90712	0.02460	0 88251	298.54210	0.00000	2.71
	cluster 1,2		192	0.06980	0.00977	0.04172	0.00000	0.23413	0.62848	100.00
	study 3			-0.20997						

	study 4			1.86652						
C35		4	1510	0.19571	0.23658	0.02407	0.21251	93.99201	0.00000	10.17
	cluster 1,2		192	0.00000	0.00000	0.00000	0.00000			100.00
	study 3			-0.18012						
	study 4			0.90296						
C37		4	1510	0.02872	0.00244	0.02392	0.00000	0.49418	0.92017	100.00
C38		4	1510	0.24756	0.41045	0.02419	0.38625	165.81408	0 00000	5.89
	cluster 1,2		192	0.04797	0.00461	0.04169	0.00000	0.11051	0.73956	100.00
	study 3			-0.33977						
	study 4			1.16773						
C39		4	1510	0.38346	0.64782	0.02446	0.62336	254.49462	0.00000	3.78
	cluster 1,2		192	0.17583	0.00261	0.04185	0.00000	0.06235	0.80282	100.00
	study 3			-0.42741						
	study 4			1.52562						
C40		4	1510	0.27756	0.05972	0.02407	0.03565	25.15593	0.00001	40.30
	cluster 1,2,3		851	0.08483	0 00593	0.02995	0.00000	0.57441	0.75036	100.00
	study 4			0.60962						
C41		4	1510	0.25265	0.04874	0.02410	0.02464	23.11589	0.00004	49.45
	cluster 1,2		192	0.22388	0.00028	0.04195	0.00000	0.00673	0.93463	100.00
	study 3			-0.00145						
	study 4			0.53595						
C42		4	1510	0.21436	0.04766	0.02405	0.02361	16.18403	0.00104	50.47
	cluster 3,2		755	0.03026	0.00060	0.02390	0.00000	0.02522	0.87383	100.00
	study 1			0.28181						
	study 4			0.46407						
C43		4	1510	0.08726	0.01738	0.02402	0.00000	2.61956	0.45407	100.00
C44		4	1510	0.24256	0.16975	0.02402	0.14573	69.58617	0 00000	14.15
	cluster 1,2			0 00000	0 00000	0.00000	0.00000			100.00
	study 3			-0.17288						
	study 4			0.75503						
C45		4	1510	0.09408	0.01250	0.02391	0.00000	4.66986	0.19763	100.00
C46		4	1510	0.19282	0.24551	0.02407	0 22144	98.35398	0.00000	9.80
	cluster 1,2			0.00000	0.00000	0.00000	0.00000			100.00
	study 3			-0.20088						
	study 4			0.91231						
C47		4	1510	0.01687	0.00058	0.02390	0.00000	0 23929	0.97101	100.00
C48		4	1510	0.09775	0.04240	0.02393	0 01847	18.21908	0.00040	56.44
	cluster 1,2,3		851	-0.07271	0.00296	0.02983	0.00000	0 32862	0.84848	100.00
	study 4			0.37120						
C49		4	1510	0.06887	0.00689	0.02391	0.00000	2.60466	0.45667	100 00
C51		4	1510	0.18156	0.16900	0.02402	0.14498	66.58516	0.00000	14.21
	cluster 1,2,3		851	-0.09026	0.00456	0.02983	0.00000	0.50647	0.77628	100 00
	study 4			0.77670						
C52		4	1510	0.06690	0.00747	0.02391	0.00000	2.92638	0.40312	100.00
C53		4	1510	0.47161	0.21406	0.02547	0.18859	73.51069	0.00000	11.90
	cluster 1,2,4		851	0.72131	0.01855	0.03190	0.00000	0.83877	0.65745	100.00
	study 3			-0.19217						
C54		4	1510	0.58933	0.35775	0.02451	0.33324	106.80224	0.00000	6.85
	cluster 1,2,3		851	0.31027	0.00348	0.03009	0.00000	0.29338	0 86356	100.00
	study 4			1.47012						
C55	no clusters	4	1510	0.81141	0.34950	0.02748	0.32202	76.76499	0.00000	7.86
C56		4	1510	0.34662	0.08678	0.02479	0.06199	15.77705	0.00126	28.57
	cluster 2,3		755	0.09705	0.00869	0.02403	0.00000	0.36150	0.54768	100.00
	study 1			0.75834						
	study 4			0.44321						

C57		4	1510	1.02360	1.18962	0.02672	1.16289	317.21421	0.00000	2.25
	cluster 2,3		755	0.26506	0.00185	0.02406	0.00000	0.07708	0.78130	100.00
	study 1			1.04524						
	study 4			2.55836						
C59		4	1510	0.85597	0.21692	0.02719	0.18973	85.40135	0.00000	12.54
	cluster 1,2,4		851	1.18206	0.01223	0.0342	0.00000	1 15453	0 56143	100.00
	study 3			0.177206765						

References

AERA-APA-NCME. (1999). Standards for educational and psychological testing. Washington D.C.: American Psychological Association.

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

Butcher, J.N. & Garcia R.E. (1978). Cross-national application of psychological tests. *The Personnel and Guidance Journal*, 56(8), 472-275.

Carroll, B.E. & Change, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika*, 35, 283-319.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaptation-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavior Research, 35*,169-200.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Claxton, J. (2003). Sampling, Instrumentation, And Data Collection of the IEA/PPP. Ann Arbor, MI: High/Scope Educational Research Foundation.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cronbach, L.J. & Meehl, P.E. (1955). Construct Validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Cunningham, I.C.A.M, Cunningham, W.H., & Green, R.T. (1973, November). A Cross Cultural Study of Subjective Product Attributes. *Proceedings of the Association of Consumer Research.* 82-98.

Donovan, M.A., Drasgow, F., Probst, T.M. (2000, April). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology, 85*(2), 305-313.

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

Ellis B.B., Mead A.D. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement.* 60(5):787- 807.

England, G.W. & Harpaz, I. (1983). Some methodological and analytic considerations in cross-national comparative research. *Journal of International Business Studies, 14*(3), 597-622.

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5, 3-8.*

Grisay, A. (1998). *Instructions for the translation of the PISA material (OECD/PISA Report).* Melbourne, Australia: Australian Council for Educational Research.

Grisay, A. (1999). *Report of the development of the French source version of the PISA test material (OECD/PISA Report).* Melbourne, Australia: Australian Council for Educational Research.

Hambleton, R.K. (1994). Guidelines for Adapting Educational and Psychological Tests: A Progress Report. *European Journal of Psychological Assessment,* 10, 229-240.

Hambleton, R.K. (2001). The next generation of the ITC Test Translation and Adaptation Guidelines. European journal of psychological assessment Vol 17(3) (2001): 164-172

Hambleton, R.K. (2002). Adapting achievement tests into multiple languages for international assessments. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58-79). Washington, D.C.: National Academy Press.

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

Hayes, A. F. (2005). An SPSS procedure for computing Krippendorff's alpha [Computer software]. Available from http://www.comm.ohio-state.edu/ahayes/macros.htm.

Hedges, L. V. & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*, 359-369.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Holland, Paul W & Wainer, H. (1993). *Differential Item Functioning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Holsti, O.R. (1969). *Content Analysis for the Social Sciences and Humanities.* Reading, MA: Addison-Wesley.

Horn, J.L. & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117-144.

Howell, D.C. (2002). Statistical Methods for Psychology, 5th Ed. Pacific Grove, CA: Duxbury—Thomson Learning.

Hunter, J.E. & Schmidt, F.L. (1990). Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. Newbury Park, CA: Sage.

Hunter, J.E., Schmidt F.L. & Jackson, G. (1982). Advanced Meta-analysis: Quantitative Methods for Cumulating Research Findings A Cross Studies. Beverly Hills, CA: Sage.

IEA Preprimary Project: Phase III. (1994). Yipsilanti, MI: High/Scope Educational Research Foundation.

Jeanre, C. & Bertrand, R. (1999). Translating tests with the International Test Commision Guidelines: Keeping Validity in Mind. *European Journal of Psychological Assessment, 15, 277-283.*

Joldersma, K. (2003). "Cross-Linguistic Instrument Comparability". Unpublished paper.

King, W.C., Jr., & Miles, E.W. (1995). A Quasi-experimental assessment of the effects of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80*(6), 643-651.

Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

McKelvey, R. D., & Zavoina, L (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology, 4,* 103-120.

Messick, S. (1989). Validity. In Linn, Robert L. Ed. (1989). Educational Measurement (3rd ed.); 13-103; New York, NY: MacMillian Publishing Co., Inc.

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 26*: 573-596.

PISA (2000) From: http://www.pisa.oecd.org/.

Raju, N. S., van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19,* 353-368.

Reckase, M., Xin, L., & Joldersma, K. (2004). "An Investigation of Sources of Differential Item Functioning in a Cross-cultural Adaptation of an Achievement Tests". Unpublished paper.

SAS (2004). SAS/STAT software 8e [Software manual]. Cary, NC: SAS Institute, Inc.

Schmidt, F. L. & Hunter J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62, 529-540.*

Schwarzer, R. (1989). Meta-Analysis Programs. Berlin, Germany: Institut für Psychologie (WE 7).

Shealy, R.T. & Stout, W.F. (1993). An Item Response Theory Model for Test Bias and Differential Test Functioning. In Holland, Paul W & Wainer, H (Eds.). (1993). *Differential Item Functioning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S.G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5, 299-321.*

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

Sireci, S.G., Bastari, B. & Allalouf, A. (1998, August). *Evaluating construct equivalence across adapted tests.* Invited paper presented at the meeting of the American Psychological Association, San Francisco.

Sireci, S.G., Patsula, L. & Hambleton, R.K. (2005). Statistical Methods for Identifying Flaws in the Test Adaptation Process. In Hambleton, R.K., Merenda, P.F. & Speilberger, C.D. (Eds.), *Adapting Educational and PsychologicalTests for Cross-Cultural Assessment* (pp. 93-116). Mahway, NJ: Lawrence Earlbaum Associates.

Skymeg Software. (2005). PRAM: a Program for Reliability Assessment with Multiple Coders. Available from http://www.geocities.com/skymegsoftware/pram.html

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 lanaguages and culres: A review of the ITC Guidelines for test adaptation. *European Journal of Psychological Measurement. 15, 258-269.*

Thomas, D. R., & Zumbo, B. D. (1998). *Variable importance in logistic regression based on partitioning an R-squared measure.* Presented at the Psychometric Society Meetings, Urbana, IL.

TIMSS. (1999). Translation and cultural adaptation of the TIMSS instruments, *TIMSS 1999 Technical Report,* International Center at Boston College. Chestnut Hill, MA: Boston College. From: http://timss.bc.edu/timss1999i/pdf/T99_TR_Chap05.pdf

Tucker, L. R. (1951). A Method for syntheses of factor analyses studies (Report No. 984). Washington, DC: Department of the Army, Personnel Research Section.

van de Vijver, F.J.R. & Poortinga, Y.H. (2005). Conceptual and Methodological Issues in Adapting Tests. In Hambleton, R.K., Merenda, P.F. & Speilberger, C.D. (Eds.), *Adapting Educational and PsychologicalTests for Cross-Cultural Assessment* (pp. 39-63). Mahway, NJ: Lawrence Earlbaum Associates.

van de Vijver, F.J.R. & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment. *European Journal of Psychological Assessment*, 8, 17-24.

Viswesvaran, C. & One, D.S. (1995). Theory testing: Combining psychometric meta-analysis and structural equation modeling. *Personnel Psychology*, 48, 865-885.

Wolf, F.M. (1986). Meta-Analysis: Quantitative Methods for Research Synthesis. Beverly Hills, CA: Sage.

Wolfe, E.W., & Manalo, J.R. (2002). Phase III Scaling of the PrePrimary Project Cognitive and Language Development Scales. Ypsilanti, MI: High/Scope Educational Foundation.

Zumbo, B. D., & Thomas, D. R. (1997) *A measure of effect size for a model-based approach for studying DIF*. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.

Zumbo, Bruno D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing* Special Issue: Advances in translating and adapting educational and psychological tests Vol 20(2) (Apr 2003): 136-147