Running Head: CROSS LINGUISTIC INSTRUMENT COMPARABILITY

Cross-Linguistic Instrument Comparability

Kevin B. Joldersma

Michigan State University

Abstract

The role of test translation and test adaptation for Cross-Linguistic Instruments (CLIs)or multilingual tests is of vital importance given the increasing international use of high stakes assessments which affect and help make educational policy. Test developers have traditionally relied upon either expert-dependent or psychometric methods to create comparable CLIs. The problem with expert-dependent methods is that they are perhaps subjective in nature. Psychometric tests remove subjectivity, but they also remove the valuable insights of experts that account for the multi-faceted problem of CLI comparability. This paper proposes linguistic analysis as a method to further CLI comparability.

Cross Linguistic Instrument Validation and Reliability

In recent years, interest in cross-linguistic testing and the comparability of different versions of a test created by translating one test into multiple languages has become a sensitive issue. This is a result of high-stakes international tests, such as TIMSS (Trends in International Mathematics and Science Study), that have an impact on national pride, educational policy and allow test users to make "universal" comparisons of research results across language groups. The role of test translation, or adaptation to multiple cultures and languages, has thus worked its way into the collective conscience of language testers and creators of large-scale assessments (see Hambleton & Patsula 2000, Sireci 1997, Auchter & Stansfield 1997). As such, it behooves testing experts to be aware of test adaptation issues and the best methods for creating valid Cross Linguistic Instruments (CLIs).

Test translation and adaptation (the cultural adjustment of a test, rather than a literal translation of a test) are often assumed to be the best way to establish consistent, equivalent, fair, reliable and valid tests across languages. However, a major hurdle in translating or adapting tests is ensuring the cross-linguistic comparability of test scores in the two languages, particularly at the item level. Item complexity, and by consequence difficulty, may vary from one language to another. This can happen between Spanish and English phrasing, for example—because Spanish speakers are more tolerant of longer sentence structure, English speakers might be disadvantaged by a test translated from Spanish to English (Gutierrez-Clellen & Hofstetter 1994). These problems manifest themselves via differential functioning among examinees on different language versions of CLIs (Allalouf 2000, Price 1999, Sireci & Swaminathan 1996, and Price & Oshima 1998). Thus, poor translation or adaptation procedures may create additional sources of bias and non-comparability of instruments for examinees, which may result in invalid inferences made from scores of such CLIs.

There are two general methods for ensuring and evaluating cross-language comparability: expert judgments and psychometrics. Both of these methods have their shortcomings: a) expert judgments are problematic because they invite subjectivity into attempts to create comparable tests (Hambleton & Patsula 2000) and b) psychometric methods fail to capture substantive insights of experts that could better inform attempts to create comparable tests and identify reasons for performance differences of different language groups (Auchter & Stansfield 1997). This study attempts to reconcile this problem by creating a new

methodology that incorporates both of these sources of information in an attempt to minimize linguistic bias in translated tests.

A logical next step in the evolution of test translation practice is the merging of translation/adaptation systems that incorporate the strengths of these two methodologies. To date, only modest attempts have been made to do so (Sireci & Khaliq 2002; Maggi 2001). Such a system would improve current practice by using language variety and differences to inform test adaptation procedures. This would, in effect, allow for increased fairness and buoy the validity of the inferences of CLIs, since a CLI's validity is a reflection of its ability to measure the same construct across cultures. One can both measure differences in performance between language groups and make an evaluative judgment regarding the equivalency of the construct tested in the language versions of a CLI. Thus, a CLI's validity is dependent upon Messick's definition of validity, wherein validity is an integrated evaluation based upon both theoretical and empirical evidence (Messick 1989).

As indicated by Sireci (1998), much of the current practice in detecting functional non-equivalence ignores the theoretical aspect of validity analyses advocated by Messick. Many studies rely primarily on the statistical indices, and some only follow up with an examination of the items or the item development categories. Linguistic analysis, proposed by this study goes well beyond these methods. Linguistic analysis is a procedure much like a content analysis and makes use of textual data and linguistic features, such as syntax and morphology, to create coding schemes and categorize data for further examination. This provides an important theoretical foundation which supports the statistical findings of psychometric methods.

The aim of this paper is to lay the basis for a technique that aids both instrument development and evaluation for Cross-Linguistic Instruments. As previously mentioned, many of the current methods for creating comparable multi-language instruments are expert-dependent, a characteristic that invites subjectivity into these efforts. Moreover, while available psychometric methods remove subjectivity, they also remove the valuable insights of experts that account for the multi-faceted problem of test translation. The author proposes linguistic analysis of language features that may prove to reduce bias which arises from linguistic diversity and translation or adaptation issues (hereafter linguistic bias). To this end, the

study seeks to answer to what degree the results of linguistic analyses are related to difficulty-based indicators of test item non-comparability.

Ensuring Cross Language Comparability

One step toward maintaining CLI comparability is to eliminate item bias. *Item bias* occurs when examinees of a particular group are less likely answer an item correctly than examinees of another group due to some aspect of the item or testing situation which is not relevant to the purpose of measurement (Clauser 1998). Of particular interest to the present study of CLI comparability is the ability to detect linguistic bias. Linguistic bias is a particular manifestation of *language bias*, which happens when items unintentionally distinguish between examinees of different language backgrounds. Linguistic bias itself is a more strictly-defined depiction of the parts of language that may explain why this differential functioning occurs. *Linguistic bias*, thus, is the presence of item bias that differentiates unintentionally between speakers of different linguistic varieties (be they entirely different languages or dialects within a language). By extension, linguistic bias in a CLI is related to item bias that discriminates unintentionally between language groups or language versions of the instrument. For example, a multiple-choice item on a translated version of a CLI may exhibit bias when there is a careless translation error which results in none of the response options being a correct answer. Such an item becomes useless as an indicator of the construct in question in the translated language.

There are many ways to detect and prevent incomparability across CLIs. Especially important to this study are procedures for detecting *Differential Item Functioning* (DIF)–statistical procedures that indicate when examinees of different demographic groups with identical abilities have different probabilities of answering an item correctly. DIF is certainly not the only method for detecting item bias, and others methods are addressed below. There are, however, essentially two groups of methodologies that test developers use to ensure the comparability of CLIs. Most techniques fall into either an expert dependent or a psychometric category.

Expert dependent methodology

Expert dependent methods rely on a professional's judgment and specialized skills to enhance the cross-language validity of measures created from CLIs. These methods have generally taken the form of evaluative methods or creative efforts to enhance the comparability of the instrument. These methods

range from those with little quality control, such as direct translation, to methods whose quality is aided by back translation[1] or using expert test-takers[2]. These actions, taken during and after instrument construction, are essential to good CLI comparability.

Expert dependent techniques necessitate the presence of bilingual experts. *Bilingual experts* are knowledgeable in both the source and target languages. Their expertise is crucial for the comparability of the CLIs, 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 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.

Psychometric Methodology

Another perspective on maintaining instrument comparability is that a more statistically-based system would allow test designers to have a relatively bias-free tool for instrument creation (Hambleton and Patsula, 2000). Such procedures include using conditional probabilities to detect DIF (differential item functioning) and finding dimensionality differences (e.g., Principal Component Analysis or Confirmatory Factor Analysis) between scores from translated tests.

Differential Item Functioning

Differential Item Functioning (DIF) is a potential indicator of *item bias*, which is occurs when examinees of equal ability from different groups have differing probabilities or likelihoods of success on an item (Clauser 1998, p. 31). Typically, DIF is used to make comparisons between a *reference group* (the standard against which the group of interest is to be compared) and *focal group* (the group of interest for the purposes of the planned comparison) to determine whether matched samples from each group have equal probabilities of answering each item correctly. Although DIF may be an indicator of item bias, it is

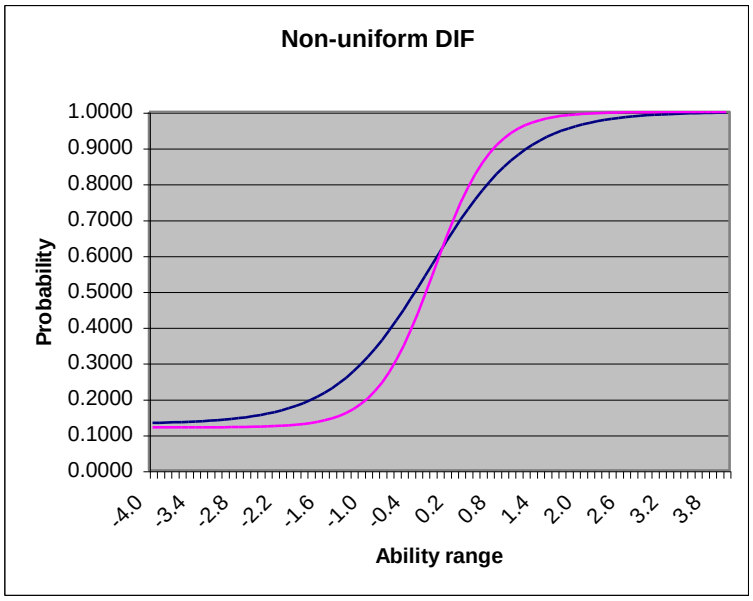
1 *Back-translation* is a method that helps to verify proper translation by translating from the original language to the target language and returning to the original language. This is done to see if the content of the back-translated material matches the original concept or the "original intent" of its authors.
2 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.

sometimes an indicator of true differences in examinee ability. As Clauser instructs, DIF procedures require a combination of expert judgment as well as statistical indices to make appropriate decisions regarding the test items.

Analysts must choose between statistical indicators of uniform or nonuniform DIF. *Uniform DIF*, illustrated in Figure 1, is the presence of differences of conditional item differences that are equal across the range of examinee ability (e.g., a main effect on item difficulty for group membership when controlling for differences in group ability). *Nonuniform DIF*, on the other hand, is depicted in Figure 2 by the presence of differences of conditional item differences that are not equal across the range of examinee ability (e.g., an interaction on item difficulty for group membership when controlling for differences in group ability).

Figure 1. Uniform DIF

Figure 2. Nonuniform DIF



On a CLI, there is reason to suspect the presence of nonuniform DIF due to translation problems or linguistic differences that examinees of differing abilities may compensate for unevenly (Hambleton 2003). For instance, examinees of differing language abilities may be impacted differently due to their ability to tolerate an inappropriate word choice due to poor translation. One DIF procedure that lends itself to detecting non-uniform DIF is logistic regression. *Logistic regression* is a nonlinear modeling technique that estimates the log of the odds (logit) that an examinee from a particular group will answer an item correctly (versus incorrectly), given that examinee's level of ability. Its basic model is given by Hosmer & Lemeshow (2000) as follows:

$$\ln(Y_{odds}) = \beta_0 + \beta_1 X_1 + ... + X_n$$

In this model, β represents a slope and X represents a coding of an independent variable. If we code the variables using 0/1, the log-odds for a reference case is equal to β_0 so that the odds for a reference case is $\exp(\beta_0)$. Thus, β_0 is the intercept of the equation while B_1 represents the effect of independent variable X_1. Additional variables, represented by the ellipsis up to X_n can be added to explain the effect of multiple independent variables.

Thus, the purpose of logistic regression can restated as follows: to compare observed vales of the response variable to predicted values obtained from models with and without the variable in question. We

can compare these values using the likelihood of *saturated model* (one that contains as many parameters as data points) to our theoretical model (Hosmer &Lemeshow 2000). This can be done with the following formula:

$$D = -2\ln\left[\frac{(likelihood\_of\_the\_fitted\_\mathrm{mod}el)}{(likelihood\_of\_the\_saturated\_\mathrm{mod}el)}\right]$$

The quantity inside the brackets is the *likelihood ratio*. This D statistic is then tested with the likelihood ratio test:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]$$

For the purposes of this study, it is especially useful to note that the logistic model accommodates both uniform and nonuniform DIF. This is done by specifying which hypothesis to test and the addition of an interaction term to the basic model, which accounts for nonuniform DIF. The presence of nonuniform DIF is tested by comparing the model fit for the model without the interaction term to the model with the interaction term (Hosmer & Lemeshow 2000).

To evaluate the logistic model, two parts are necessary; a test of the significance of the model and a measure of the effect size of the model. Zumbo and Thomas (as quoted in Zumbo 1999) argue that to effectively evaluate a logistic regression model, one most provide for both the significance test and a measure of effect size in order to avoid ignoring significant effects and over-emphasizing trivial effects. The chi-square is used to test the fit for the model when it is used to test group membership (i.e., language groups, as in this study). The combination of both techniques is essential to DIF testing, especially when one considers the hierarchical nature of uniform and nonuniform DIF that can be accommodated in this manner.

Dimensionality Assessment

Multidimensionality, for the purposes of this study, is the presence of multiple dimensions of measurement, which assess different latent traits which can, at least partially, be attributed to the presence of language differences. Essentially, the multiple dimensions, if detected, will be evidence to suggest that examinees may not be evaluated in the same manner on different language versions of a test. Essentially, examinees of different language backgrounds may have different dimensional structures due to various

cultural differences, differences in educational systems, etc. In such cases, the multiple versions of the CLI would not produce comparable information about examinees. Evaluating whether dimensionality differences exist is a necessary first step in this investigation because such differences may implicate the need to perform additional analyses, such as the DIF procedures described above.

Dimensionality differences between groups can be discovered through the use of Exploratory Factor Analysis (EFA). EFA is a procedure that uncovers the latent variables that exist within a dataset by allowing the empirical relationships to define "factors." An EFA or other dimensionality analyses are a critical part of CLI comparability studies due to their ability to partition out items that a traditional DIF may not flag (Sireci 1998). EFA may also identify the dimensional structure of the CLI, which will help the investigator to determine if linguistic factors are indeed present in the data of the study's instrument.

An EFA, as explained by Lattin et al (2003), assumes that observed variance is attributable to a relatively small number of common factors (an unobservable characteristic common to two or more variables) and a single specific factor (unrelated to any other underlying factor in the model, like error variance for a particular factor). This can be illustrated as follows:

$$X_i = \lambda_{il}\xi_1 + \lambda_{i2}\xi_2 + ... + \lambda_{ic}\xi_c + \delta_i$$

where,

X_i is the variable being measured

λ_i is the coefficient associated with the extent to which each measure X reflects the underlying common factor

ξ_i is the common factor

δ is the specific factor

c is the number of common factors in the model

One problem with EFA is that it is sometimes difficult to account for the factor loadings via substantive interpretation. This could possibly be due to the fact that there may not be a simple structure present within the data. It may also be due in part to the infinite number of solutions possible for the common factor. In order to make these infinite solutions more comprehensible, *rotations* are sometimes performed, which improve interpretability by more closely approximating *simple structure* (Thurstone 1947) wherein items will theoretically be most strongly correlated with only one latent factor of the factor-

loading matrix. Rotation procedures may be *orthogonal* (creating non-correlated components) or *oblique* (creating correlated components).

Sireci (1998) argues for the importance of doing both dimensionality analyses and DIF analyses when considering a CLI. Dimensionality analyses are necessary when evaluating a CLI because they are able to partition items into factors that a traditional DIF analysis cannot. While dimensionality can indicate groups of items that have common underlying factors, these analyses do not in themselves indicate either directionality (in favor of one group or another) or the kind of differential item functioning (i.e., uniform and non-uniform DIF) that items may possess. Hence, the necessity, from a statistical standpoint, of both techniques becomes apparent.

An Alternative Method: Linguistic Analysis

Linguistic analysis is a careful review of an instrument's components (items stems and answers, instructions, etc.) that can be performed to minimize potential sources of bias on a test. This can be done by using linguistic features of the translated tests to indicate how language differences may cause differential functioning. Similar to a content analysis, a linguistic analysis often makes use of textual data to create coding schemes and categorize data.

Some experts have performed detailed linguistic analyses using expert judgment about the structure and syntax of language to show where tests of language have fallen short (Durán et al 1987). This does not specifically apply to CLIs, where the situation is one of adaptation to many languages, rather than accommodation within a single language version of a test, as in Durán's work. Others have examined the manner in which the "ideal reader" of a test item may perform normally on an evaluation, while a less than ideal reader (due to sociolinguistic issues) performs quite poorly (Kay 1987). However, there do not appear to be any studies that specifically attempt to show the explicit connection between psychometric methods and linguistic analyses of instruments, let alone do so in a context of translation and CLI comparability.

Several of the large branches of linguistic study may prove to demonstrate this connection, or at the very least extend linguistic analyses to the study of CLI comparability. Linguistic fields ripe for this venture include syntax, morphology and lexicography. These disciplines have been used to create the

coding system to make further inferences and analyses of the study's target instrument (reasons for these selections are explained in the Coding System below).

A Complimentary Approach: Linguistic Analysis and DIF as a Hybrid

The aforementioned methods, while effective when used by themselves, are not a panacea. When used in isolation, neither linguistic analysis nor DIF are sufficient. A hybrid methodology would take advantage of the qualitative merits of linguistic analysis and the quantitative advantages of DIF procedures such as MH. If linguistic analysis is shown to work, these two techniques for instrument analysis complement each other by first efficiently flagging suspicious items, and then providing a satisfactory substantive analysis of the problematic items. An informed decision can then be made regarding the instrument; be it to keep, revise or remove the item, depending on the purposes of the test.

To arrive at that informed decision, the author advocates taking care to methodically analyze any CLI's comparability using the following system:

1. Analyze the dimensionality of the instrument to test for the presence of multidimensionality, which may be an indicator of linguistic bias.
2. Perform a linguistic analysis by:
    a. Identifying plausible linguistic features and items that may detract from an instrument's comparability across languages,
    b. Creating a coding scheme based on those features, and
    c. Categorizing items.
3. Flag items with psychometric methods (logistic regression to detect DIF) and compare with linguistic analysis findings (step 2).
4. Provide an analytical explanation and connect it to relevant linguistic theory and psychometric measures.

These steps are detailed in the Methods section that follows.

The Coding System

To investigate the data from the target languages of this study, a coding scheme of potential linguistic pitfalls has been created in order to flag items from the different language versions of the test. The coding system includes flagging items for morphological, syntactic and lexical problems.

Morphological problems would include verb tenses, gender and number agreement, case and marked possessives that may cue respondent's answers. Examples of syntactic items that may be flagged include items with different comparison structures, word placement, question formation, and underlying syntactic complexity (greater numbers of subordinate clauses, for example). Possible lexical causes of incomparability in CLIs could be false cognates incorrectly used in translation, a word or concept which does not easily translate, periphrastic verbs (verbs which require helping words to make sense) or words that have multiple meanings within a language.

Improving each of these facets will likely improve the overall quality of CLIs. However, the stated goal of this particular investigation is to determine whether a prediction can be made as to which items will exhibit dif, rather than to actually test whether improvements take place. Table 1 details many of the potential sources of linguistic incomparability and is designed to be generalizable to any language.

Table 1. Possible Linguistic Causes of DIF.

Category	**DIF source**	**Coding Symbol**
Morphology:	verb tenses	MVBT
	gender	MGEN
	number	MNUM
	case	MCAS
	possessive	MPOS
Syntax:	comparisons	SCOMP
	adjective placement	SADJ
	transitive and intransitive verbs	STI
	reflexive vs. non-reflexive	SRNR
	wh-words and question	SWH
	subordinate clauses	SSC

Lexicon:	self-evident meaning	LSEM
	word does not exist	LNON
	divergent/convergent word	LDIV
	borrowed words	LBW
	false 'friends' cognates	LFF
	periphrastic verb/adjective	LPV

Morphology

The morphology of a language can be of great aid to a test taker depending on the purpose of the items. Where one language may mark gender and person with complex indications of either first, second or third and/or singular and plural markers, others may have more simplified markers (or none at all) that are either understood in context or lexically marked. In Spanish, for example, one can indicate first person singular, present tense, indicative mood; all with the one morpheme (-o). English does not contain all this rich information in its morpheme.

Additionally, morphology can help a reader or test taker with the number involved with the verb. It can also help a speaker by indicating the gender of a subject or object in question. Cues such as number and gender may unintentionally give test-takers of one language an unfair advantage because they will already have a morphological clue as to what the object of the action might be.

Another way that morphology can provide additional unintentional information to test-takers is by way of case markings contained in a verb's morphology. This is demonstrated in German (Flippo 2003), where the article is marked with accusative morphology, and *Der* changes to *den* to indicate an object or recipient of an action. It seems logical that these denotations of case may trigger answers more easily for speakers of case-marking languages when they are asked object identification items.

The last source of morphological aid that considered for the purposes of this investigation is possessive marking morphology. A difference between English and Spanish illustrates this point. In English, an apostrophe followed by an "s" most commonly marks possession. However, Spanish marks possession syntactically, rather than morphologically. This difference may also lead to differential item functioning.

Syntax

Syntactic differences between languages also have the possibility of allowing one language group to outperform another. Some types of syntactic differences are comparisons, adjective placement, transitive and intransitive verbs, reflexive and non-reflexive verbs (which may also be marked morphologically), wh-words and question formation, and subordinate clauses. Other syntactic differences exist, but the aforementioned types conform to the question types and may be likely to appear in the IEA test (the subject of this investigation) based on the prompts.

Adjective placement can be a crucial factor in translation. Placement of the adjective in some languages (e.g. Spanish) can completely change the meaning of the sentence. The placement of *viejo* 'old' before or after *amigo* can change the meaning of the word to mean either a friend of advanced age in years or a friend that you have known for a long time.

Another syntactic feature that may trigger DIF between languages is in comparative sentences. The syntactic structure may differ in its level of complexity, facilitating easier comprehension in one language or another. In the comparisons found in the appendix, the syntactic structure is very similar between English and Spanish. However, Spanish cues the reader that a comparison is being made sooner than English. This may give an edge to Spanish speakers since scholars (such as VanPatten 1993) show that initial words are processed and comprehended more readily.

Verbs may vary between transitive and intransitive among languages. As such, the nature of a verb and its capacity to have a compliment or not may influence its translation or representation in the other target language. Thus, a speaker may process an item less readily, and therefore answer an item with more difficulty.

Question formation may be another area where differing syntactic structure occurs. Some languages are more flexible with their question formation types, while other are rather regimented in what is permissible. Awareness of this tolerance (or intolerance) allows for better construction of items.

The amount of subordination that a language allows can also influence how speakers might perform on different language versions of a test. Studies have shown, for example, that Spanish speakers have a much higher tolerance than English speakers for information embedded in subordinate clauses

(Gutierrez-Clellen & Hofstetter 1994). As a result, a differing amount of linguistic opacity permissible in a particular language may also account for items that exhibit DIF between languages.

Lexicon

Lexical items have a very high likelihood of giving unintentional cues in a translated test. One such example that has been passed around is an analogy translated from English to Swedish. The analogy included a phrase which was designed to compare diving flippers and a duck's webbed feet. The Swedish word for diving flippers, apparently, is quite literally "webbed feet" (class notes, Reckase 2003). The self-evident meaning of the Swedish word gives an advantage to the Swedish language test-takers.

Another type of lexical issue that may add to CLI development difficulties is when a particular lexical item does not exist. A famous example from Eskimo comes to mind. Though proven false, a long-held belief was that the Eskimo language contained dozens of words for what English speakers simply refer to as snow (Martin 1986).

In the same vein, multiple lexical items may be collapsed from one language to another, while in the same two languages other single lexical items may expand to have multiple meanings with multiple lexical items in the target language. Spanish and English have some words with divergent (multiple) meanings. For example, the word *malo* in Spanish has a divergent meaning in English and may mean either *bad* or *evil*.

Borrowed words from other languages also have the potential to cause lexical DIF on a CLI. On a sociolinguistic note, members of certain social strata may be more prone to use borrowed words. Borrowed words may gain new meanings in the language the have been borrowed into, thus creating potential pitfalls for translators.

False friends or false cognates are another danger to a CLI. Where one may believe that a word is correctly translated into the target language, it may indeed have a very different meaning. One clear example of this is the Spanish word *embarazado*, which appears quite similar to the English word *embarrassed*, but actually means *pregnant*.

Lastly, this study will consider the potential for periphrastic verbs and adjectives to create DIF on CLIs. A periphrastic verb or adjective is a combination of multiple words which creates a unique meaning

other than that implied by one of the words on its own. The words *subir* and *bajar* are single lexical items, whereas their respective English meanings are the periphrastic verbs *to go up* and *to go down*.

Hence, as a result of looking at aforementioned linguistic factors, we have numerous potential sources for CLI non-comparability. Consequently, it is the aim of this study to determine whether linguistic analyses and psychometric analyses are sufficiently interrelated to make it possible to integrate them in the analysis and development of CLIs. As such, the investigation seeks to answer the following questions:

1. Using factor analytic and content analysis procedures, can one identify substantive explanations for observed differences in the differential dimensional structure of translated tests?
2. Can language structure analysis be used to predict DIF in translated tests, specifically with respect to language (a) morphology, (b) syntax, and (c) lexical structure?

Methods

This study endeavors to use linguistic analysis, as detailed below, to further CLI comparability. The source data come from sections of the IEA (International Education Assessment). The particular focus within the IEA is the comparability of two languages: Italian and English. These language versions will be analyzed using a combination of expert-dependent linguistic analysis and psychometric methods more fully described below.

Sample

The examinees for the IEA (International Education Assessment, described in Instrumentation) are children from 3 to 5 years old. The data sets used for this analysis came from the cognitive tests given in the United States and Italy. Descriptive statistics and demographic information is available for these tests by way of a scaling study done by Wolfe and Manalo (2002). Table 2 details mean logits for the cognitive ability measurements and their standard deviations in parentheses. Two subscales of quantitative and spatial items are included as subscales of the children's abilities.

The cognitive test shows that, in this sub-sample, the Italian and American children in the sample are roughly the same age, with nearly equal proportions of males and females. The discrepancy in the overall cognitive score between the Italian and American examinees is notable. However, the logistic

regression techniques for evaluating DIF will accommodate this by controlling for the examinees differences in ability.

Table 2. Cognitive development

Variable	***Italy***	***US***
Age (s.d.)	4.52 (0.18)	4.50 (0.30)
% Male	54	50
N	498	557
Cognitive (s.d.)	0.52 (1.22)	0.05 (0.93)
Quantitative (s.d.)	0.55 (1.37)	-0.07 (0.98)
Spatial (s.d.)	0.61 (1.12)	0.24 (1.07)

Instrumentation

The Instrument used in this research is the IEA (International Education Assessment) Preprimary Cognitive Developmental Status Measure. The Preprimary project is an international assessment that collects data on children's cognitive and language development. The portion of the IEA used in this study is the cognitive assessment, which contains both quantitative and spatial relations questions. The cognitive test reflects three subscales of ability with quantitative , spatial relations and time perception examined. 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 (Claxton (2003). The following is an example drawn from the spatial relations section:

"Look at the birds and windows. Point to the bird that is flying ***toward*** the window . . . Point to the bird that is flying ***toward*** the window."

The instrumentation selection and development for the IEA was the result of an international collaboration. First, the IEA 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 IEA. 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 IEA 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 IEA Phase II assessments. As such, the IEA 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.

Data analysis procedures

The analysis for this project was comprised of three major parts, 1) a tabulation and analysis of the items that were predicted to exhibit DIF based on the linguistic coding system, 2) a psychometric analysis of the items, and 3) a substantive explanation of why the coding system did or did not work to effectively predict DIF. To perform these major parts, the researcher has broken them down into several smaller steps as follows:

1. Analyze the dimensionality of the instrument to test for the presence of multidimensionality, which may be an indicator of linguistic bias
2. Perform a linguistic analysis
    a. Identify plausible linguistic features and items that may detract from an instrument's comparability across languages

b. Create a coding scheme of linguistic categories based on those features

c. Categorize and code data (at the item level)

3. Flag items with psychometric DIF methods (logistic regression) and compare with linguistic analysis findings (step 2)
4. Provide an analytical explanation and connect it to relevant linguistic theory show how it supports the psychometric measures.

Details of these steps follow.

Analyze the dimensionality.

An EFA was performed using SAS software (SAS Institute Inc. 2004). The EFA may be a primary indication of whether or not language or cultural differences are likely to be found in the analysis. Dimensionality analyses are only used to flag differences in factor structure. If multidimensionality is confirmed, a substantive analysis is indicated as an appropriate next step followed by the DIF analyses that follow below. As noted in the description of rotation, factors may need to be rotated to make more substantive sense.

Three rotation methods were chosen in addition to the original Principal Component Analysis. *Principal Component Analysis* (PCA) is a method that to reduce the dimensionality of multivariate data. It allows researchers to re-express the multivariate data by taking linear combination of the original variables. This is done so that the first few resulting new variables, called components in PCA, account for as much of the available information as possible. (Lattin et al 2003). The rotations selected were promax, varimax and oblimin.

Promax is designed to perform an orthogonal rotation similar to the original solution, but designed to maximize loadings for substantively defensible interpretations of the factor structures.

Varimax is an orthogonal rotation that maximizes the variance for each component (rather than just the first component). Consequently, the eigenvalues for a PCA will be more evenly distributed after applying a varimax rotation. This procedure usually has the result of having items load highly or very lowly with each factor in the analysis when compared to other solutions.

Oblimin rotation may allow better interpretability of the data or the simple structure by assuming orthogonality. In other words, oblimin performs an oblique rotation wherein the factors do not need to be

uncorrelated. This also has the advantage of providing more satisfactory explanations of the factors extracted due to loadings which have more substantively interpretable results.

The differences between oblique and orthogonal rotation interpretations are not always great, but when they are different, there are no criteria for choosing one rotated set over another. A further complication is that oblique rotations are more difficult to interpret. This is because both the factor pattern matrix (standardized loadings to arrive at factor scores) and the factor structure matrix (correlations between Xs and Ys) need to be considered when considering the results of these rotations. In orthogonal rotations, the factor pattern matrix and the factor structure matrix are the same.

Once factorial structures are obtained and rotated from the different language samples, an objective manner of evaluating whether the solutions obtained are fundamentally equivalent needs to be done. The technique that serves this purpose is the *coefficient of congruence* (Tucker 1951), which is a statistical method used to evaluate factorial structures obtained from different samples. This technique can be applied to either a situation of one instrument given to different samples (as is the case of this study) or applied to a situation where different instruments are given to the same sample. This study will make use of the coefficient of congruence in order to evaluate the differences between the factorial structures of the two language versions of the instrument. The coefficient, for two samples on same variables, is formulated as follows:

$$CC_{pq} = \frac{\sum_{i=1}^{n} a_{ip} \cdot a_{iq}}{\sqrt{\sum_{i=1}^{n} a_{ip}^{\,2} \cdot \sum_{i=1}^{n} a_{iq}^{\,2}}}$$

where:

CC_{pq}: is the coefficient congruence.

n : number of variables in two samples.

p : number of factor in first sample.

q : number of factor in second sample.

a : factorial loading in first sample.

b : factorial loading in second sample.

The coefficient is interpreted as having a distributed range between 0 (total discrepancy) until 1 (similar solutions).

Linguistic Analysis.

The first step of the proposed methodology was to perform a content analysis. In the study, this was done in the following manner: First, a cursory analysis of the language versions of the test was carried out in order to verify the number of sections and items for each of the two tests. Next, the instruments were carefully read and the key features of the test identified and labeled. This identification includes the construct measured by the instrument at a general level, as well as the measurement purpose of each item's answers, stem and instructions (if applicable). This in part validates the instrument's content by checking to see whether items appear to measure the intended construct of measurement. If the item does not meet the stated or implicit purposes of the instrument, a review of that item is recommendable. Lastly, the apparent success or failure of the translation was evaluated by comparing the content of each item across languages. This is the heart of the analysis, and was done in order to see if there have been potential breakdowns of the CLI's comparability.

Identify plausible linguistic features.

After the content analysis is performed, the next step is to identify plausible linguistic features and items that may detract from an instrument's comparability across languages. This initially consisted of the researcher's own judgment and previous study of the linguistic differences between the languages, but also included some of the potential sources of DIF mentioned earlier in the literature review. The second phase was to use the results of the content analysis to see where the language versions differed. The language differences between the versions may be the results of translation or adaptation failures, thus leading us to possible clues as to where potential problems lie.

Categorize data.

The above list of plausible linguistic features was organized into categories to facilitate coding and interpretation. For simplicity, the author chose the major branches of linguistics for this categorization. These categories include: syntax, morphology and lexicography/semantics. An explanation for their inclusion can be found in the previous section entitled Coding System.

Coding scheme.

After the list of plausible linguistic features is identified, a coding scheme based on those features will be created. This coding scheme is detailed above in the Coding System section. Features will be coded for item instructions, stems and answers as appropriate.

Items will then be coded as likely to exhibit DIF (1) or unlikely to have DIF (0) for each of the three linguistic categories identified above (syntax, morphology and lexical). An item may have more than one category of linguistic bias indicated, and will then be coded for each category in which it may exhibit DIF.

An independent coder tested the interrater reliability of all the items in order to verify the coding system. Results from the untrained rater were marginal (interrater reliability = 0.76), indicating that additional training of future raters may be required.

Flag and Compare.

Items were flagged using the linguistic coding system created above. The criteria for evaluating an item as having potential linguistic bias, for the purposes of the study, was to label it at the item level using expert judgment as being likely or unlikely to exhibit DIF (coded 1 and 0, respectively).

Items were also flagged using psychometric methods to compare with the linguistic analysis findings. This was done through logistic regression procedures. As previously stated, the study uses Zumbo's criteria of a combination of a Chi-square (with 2df) and the Zumbo-Thomas effect size measure (R squared of at least 0.130) to indicate the presence of significant DIF. SAS software (SAS Institute Inc. 2004) was used to do logistic regression where the reference group is the American sample and the focal group is the Italian sample. Items were then be flagged for uniform and nonuniform DIF.

The comparison between the linguistic analysis and the psychometric methods was be comprised of a logistic regression wherein each linguistic category is an independent variable on the outcome of being flagged or not (coded 1 0) in the previous analysis. Each of the three linguistic categories were tested singly, to verify their ability to predict some of the variance in the outcome. Then, each of the categories was combined to see if there is an independent and additive effect. Each effect was then evaluated for significance at the .05 level based on the effect size indicated by the logistic regression.

Analytical Explanation

The goal of the analytical explanation is to provide a connection to relevant linguistic theory and psychometric measures. Efforts are made to explain possible causes for cross-language noncomparability in the items flagged via both methods. This is done by identifying associations that exist between the DIF and linguistic categories and testing the associations with a chi-square. In essence, this allowed the development of a substantive model that may explain why the DIF exists.

Results

Dimensionality of the CLI

The analysis of the IEA's dimensionality was performed using several rotation trials to see which solution provided an explanation for the greatest amount of variation while keeping simple structure in mind. The rotations performed, as previously described are: a non-rotated principal component analysis, varimax, promax, oblmin.

The Principal Component Analysis serves as the baseline analysis for comparing how much variation might be accounted for before any rotation. Table 3 shows the factor loadings of the Varimax rotation, which indicate that one dominant factor has been extracted from the item responses. This holds true for all the other rotations as well. Varimax results have slightly larger communality estimates for the Italian version, but smaller estimates for the US language version. Promax results are identical to those of varimax for all factors extracted. Oblimin, however, differs slightly, but still yields nearly identical results. Hence, it appears that rotation does little to change the amount of variance explained by the model. This finding seems to reflect the complex multidimensional nature of the CLI, where a simple structure may not be immediately apparent.

Table 3. Factor Loadings Using Varimax: U.S. and Italy.

Italian Factor Analysis--Varimax Rotation					
Item	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**
i3	-0.01802	0.15744	0.43605	-0.17517	0.42236
i4	0.36596	0.17054	0.14797	-0.28318	0.12983
i5	0.10893	0.03257	0.188	0.02249	0.21319
i6	0.37559	0.28238	0.10381	0.03928	-0.19148
i7	0.04267	0.37624	0.20093	-0.07927	-0.1691

i8	-0.14221	0.02484	0.7782	-0.02967	-0.08544
i9	0.45408	-0.04128	0.36984	-0.05924	-0.09287
i10	0.28092	0.06651	0.53493	-0.15316	0.14633
i11	0.15767	-0.07749	0.57313	0.33645	0.08991
i12	-0.00003	0.35077	-0.04804	0.15231	-0.02978
i13	0.2038	0.18345	0.06453	0.09439	-0.29063
i14	0.24289	0.40172	0.07836	0.03354	-0.02151
i15	0.14428	0.08707	0.00662	0.02622	-0.15212
i16	0.45502	-0.00804	0.16759	0.12774	0.13766
i17	0.24618	0.35774	0.11844	-0.36415	-0.11981
i18	0.24732	-0.28902	0.35492	0.34537	0.10258
i19	0.03625	0.02023	0.059	0.00018	0.78246
i20	0.26093	0.47271	-0.03371	-0.16601	-0.09582
i21	0.10052	-0.12338	0.02094	0.04132	0.61291
i22	0.37443	0.12205	-0.14676	0.1367	-0.11801
i23	0.19858	0.47149	0.08562	0.12599	0.34887
i24	0.13116	0.27393	0.04802	-0.0751	-0.1128
i25	0.28313	0.1596	-0.0513	0.04814	0.36508
i26	0.27059	0.47772	0.0017	0.16806	0.0963
i27	0.41123	0.12894	0.00147	0.1191	0.02136
i28	0.13389	-0.16324	-0.08481	0.15529	0.07602
i29	0.01248	0.3179	0.02731	-0.11506	0.36375
i30	0.35548	0.45632	-0.15456	0.01905	0.0518
i31	0.37595	0.03313	0.00774	0.01715	-0.14698
i32	0.49211	0.20591	0.01474	0.08863	0.05118
i33	0.54152	0.01461	0.0319	-0.00107	0.01343
i34	0.39165	0.17747	0.004	-0.14265	0.10691

i35	0.49356	0.07664	0.07771	-0.08096	0.05655
i36	0.57902	-0.02526	0.0512	0.00453	0.13736
i37	0.0616	-0.0795	0.11625	0.64751	-0.04076
i38	0.15553	0.01994	0.00797	0.39653	-0.05633
i39	0.02869	0.43205	0.03144	-0.01076	0.09091
i40	-0.07889	0.40604	0.36434	0.28979	0.04514
i41	-0.17095	0.09642	0.57634	0.43454	-0.20727
i42	0.36458	0.155	0.24095	0.23875	-0.16431
i43	0.46303	-0.09207	0.00434	-0.05773	0.03205
i44	-0.15452	0.42777	-0.00017	0.33914	0.14235
i45	-0.00175	0.25338	0.00177	0.50504	-0.09703
i46	0.19264	0.29172	-0.06691	0.41521	0.21356
i47	0.11695	0.05896	0.39189	-0.02381	0.03996
i48	0.40415	0.20285	0.03882	0.04495	0.0366
i49	0.4535	0.02984	0.03928	0.04982	-0.02036
i50	0.45798	0.14499	0.15455	0.23312	0.19563
i51	0.35129	0.22357	0.01389	0.11729	0.08125

USA Factor Analysis--Varimax Rotation					
Item	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**
i3	0.6475	0.06917	0.09055	0.0242	-0.10523
i4	0.57233	0.29024	-0.0395	0.14325	-0.24941
i5	0.53938	0.25155	0.00653	0.08021	-0.23758
i6	0.44692	0.07089	-0.15636	0.08931	0.05969
i7	0.24339	-0.09521	0.23195	0.00786	0.02927
i8	0.51651	-0.15497	0.36048	0.04959	0.25854
i9	0.47324	-0.04678	-0.05911	0.20677	0.26914

i10	0.27572	0.0286	0.21661	0.27691	0.02069
i11	0.44192	0.06311	0.15665	0.08075	0.13293
i12	-0.00443	0.40713	-0.13195	0.26212	0.21075
i13	-0.02828	0.31599	0.02659	0.07285	0.16601
i14	0.26826	0.37178	-0.16191	0.10739	0.13545
i15	0.00469	0.50328	0.13839	-0.13485	0.0518
i16	0.43815	0.06015	-0.12034	0.24373	0.16416
i17	0.28335	-0.00228	-0.07794	0.2485	0.24917
i18	-0.10388	0.21907	0.50657	-0.00025	0.44654
i19	0.45703	0.03075	0.18603	-0.26247	0.22828
i20	0.22653	0.4024	0.26586	-0.14232	0.29954
i21	0.03516	0.10761	0.51692	-0.07539	0.18695
i22	0.03025	0.26635	0.32211	0.17864	0.00658
i23	0.26685	0.54019	0.12871	-0.15234	0.10802
i24	0.04242	0.27076	0.19213	0.07119	0.45163
i25	0.5157	0.17247	0.15353	0.00126	0.10627
i26	0.04948	0.2693	0.24106	0.04865	0.45455
i27	0.14663	0.08481	0.0805	-0.05204	0.48352
i28	0.37537	0.2509	0.04181	-0.08946	0.06949
i29	0.17578	-0.095	-0.10103	0.24924	0.41917
i30	-0.01311	0.51846	0.08523	0.00898	-0.03355
i31	0.08583	0.23248	0.00377	-0.08435	0.18024
i32	0.03614	0.48445	0.00527	0.31181	0.04731
i33	0.16576	0.04587	-0.03548	0.49576	0.10384
i34	0.11693	0.04992	-0.10981	0.3165	0.05327
i35	0.10952	0.38017	0.18488	0.1565	-0.0216
i36	0.24555	-0.04552	0.037	0.11866	-0.0391

i37	0.2459	0.20074	0.09551	0.26878	-0.01015
i38	0.04874	0.10163	0.60722	-0.01555	0.01971
i39	0.06469	0.01368	0.51906	0.20503	-0.15195
i40	0.17352	0.3574	0.04887	0.14294	-0.03615
i41	0.01649	0.31326	0.49625	0.02592	0.2489
i42	0.07923	0.03788	0.24188	0.03469	-0.10939
i43	-0.03119	-0.0625	0.16683	0.44227	-0.20958
i44	-0.08841	0.39411	0.27646	0.00409	0.02457
i45	0.07683	0.2533	0.23761	0.46688	0.03718
i46	0.17142	0.24421	0.25145	0.38173	0.06796
i47	-0.01924	0.11962	0.38178	0.31278	0.15271
i48	0.24136	0.317	0.23014	0.17242	-0.04229
i49	-0.00454	-0.05425	0.12516	0.42029	0.08402
i50	0.1118	0.25829	-0.01702	0.2698	0.11404
i51	-0.07732	0.04228	-0.0761	0.16973	0.41224

Table 4 contains the congruence coefficient for all rotations and illustrates that there is one factor that is common to both language versions of the test. The congruency coefficient of ~.87 shows that this factor has a high degree of congruency across the language versions. The table also conveys that the next 4 factors extracted do not have good indicators of congruency. There is likely a large discrepancy between the two languages in terms of the congruency of how these last four factors are expressed in the language versions.

Table 4. Principal Component Analyses and Three Rotations

	Factors		**Test of congruence**		
Rotation	Italian(var explained / %comunality)	US (var explained / %comunality)	CCpq	Zcpq	CTpq
None					

1	3.8407127	0.318049	6.0723070	0.441459			
2	2.2427650	0.185723	2.6112765	0.189841			
3	2.2215744	0.183968	1.7857364	0.129824			
4	1.9722444	0.163321	1.6722287	0.121572			
5	1.7985737	0.148939	1.6135478	0.117305			
Total	12.075870		13.755096				
Varimax							
1	4.8285790	0.334393	4.1177326	0.299361	0.87683	1.36189	0.17249
2	3.4364555	0.237985	3.8621639	0.280781	-0.12201	-0.12262	0.34652
3	2.7964166	0.19366	3.2512548	0.236367	-0.31913	-0.33068	0.32669
4	2.3782742	0.164703	2.3174668	0.168481	-0.19994	-0.20267	0.30189
5	2.3566830	0.163207	2.5552916	0.185771	-0.05821	-0.05828	0.27318
Total	14.439813		13.755096				
Promax							
1	4.8285790	0.334393	4.1177326	0.299361	0.87683	1.36189	0.17249
2	3.4364555	0.237985	3.8621639	0.280781	-0.12201	-0.12262	0.34652
3	2.7964166	0.19366	3.2512548	0.236367	-0.31913	-0.33068	0.32669
4	2.3782742	0.164703	2.3174668	0.168481	-0.19994	-0.20267	0.30189
5	2.3566830	0.163207	2.5552916	0.185771	-0.05821	-0.05828	0.27318
Total	14.439813		13.755096				
Oblimin							
1	4.9044478	0.339648	4.1284654	0.300141	0.87683	1.36189	0.17249
2	3.3917164	0.234886	3.8268894	0.278216	-0.12201	-0.12262	0.34652
3	2.9981040	0.207628	3.0726157	0.22338	-0.31913	-0.33068	0.32669
4	2.1496780	0.148872	2.3635456	0.171831	-0.19994	-0.20267	0.30189
5	2.0993026	0.145383	2.4585259	0.178736	-0.05821	-0.05828	0.27318
Total	14.439813		13.755096				

By performing a substantive analysis of the factor analyses of each language, perhaps the reasons for this discrepancy will become apparent. Both Italian and US English versions of the instrument have a larger primary factor, which includes items that deal with three spatial relations skills: object-manipulation, observation of spatial relations, and auto-spatial relations. Object manipulation items involve moving a toy or other object to various positions at the request of the test-giver. Observation of spatial relations includes the ability to identify the correct picture when prompted to pick between pictures showing the object in various locations. Lastly, auto-spatial relations items involve the child (test-taker) moving themselves to a location specified by the test-giver. Interestingly, the Italian children's factor structure shows that items that deal with the three aforementioned item types additionally load with several quantitative items where the children must point to pictures illustrating "many" or "few".

Factor 2 appears quite similar between US and Italian versions, being a mix of spatial items testing words like: through, between, nearest, inside and next to. It also includes comparable quantitative items that test concepts such as "whole" and "most". However, the item content similarity breaks down significantly among factors 3 through 5. Items that load on factors 3 through 5 do not appear to have the same apparent common ground as items from factors 1 and 2.

Likely, we can conclude there is a small core set of similar items that the IEA measures. This core set appears common to both Italian and US versions of the test. This may account for there being one dominant component, which remains similar to all examinees. This consists of the largest factor extracted from the items. The other smaller factors do not seem to align well at all, however. These differences in dimensional structure are quite probably due to, among other causes, the linguistic differences that exist between language versions of the CLI. If these differences in dimensionality can be accounted for by linguistic analysis, linguistic analysis for CLI comparability may be supported as an additional tool for test developers. However, the potential success of that tool has yet to be evaluated. The next section provides details on the implementation of linguistic analysis.

Evaluating Linguistic Analysis

Language structure as an analytical tool needs to be evaluated to answer the author's second research question. Specifically, it desires to know whether linguistic analysis, with respect to language morphology, syntax, and lexical structure can be used to predict DIF in translated tests. To answer this

question, the author has performed the following two sets of analyses: (a) a chi-squared test to demonstrate the relationship between linguistic flags, regardless of type, and their ability to predict DIF, and (b) local chi-squares to show the efficacy of each individual type of linguistic flag, thereby showing whether types matters or does not matter.

The results of the logistic regression indicate that is are a large percentage of items that exhibit DIF of some sort on the IEA. Overall, between uniform and nonuniform DIF, 12 out of 49 or roughly 25% of items were flagged as exhibiting DIF of some type. The linguistic analysis, on the other hand, flagged 23 unique items of the 49 examined (about 47%) as having a likelihood for exhibiting DIF. In regards to the linguistic categories, 17 items were flagged as likely to have syntactic issues, 12 were identified as having morphological differences that could cause non-comparability and 10 items were categorized with marked lexical problems.

Table 5. Items Flagged by Linguistic Analysis

Linguistic Category	**Number of Items Flagged**	**Percent**
Syntax	17	34.69%
Morphology	12	24.50%
Lexicon	10	20.41%

A typical item that linguistic analysis flags for syntactic differences is illustrated by the following item:

> *Guarda questi cagnolini. Fammi vederi quale di questi cagnolini si trova vicino alla sua cuccia… Adesso fammi vedere quale di questi cagnolini si trova* ***vicino*** *alla sua cuccia.*
>
> *Look at the dogs. Point to the dog that is* ***next to*** *his house… Point to the dog that is next to his house.*

The analysis flags this item for a couple of the syntactic reasons mentioned previously. First, we must consider the overall length of the two items. Clearly, the Italian item is longer than its U.S. counterpart, which may also be an indication of greater syntactic complexity in the Italian item. Additionally, the Italian item has a reflexive verb and it also does not bold the accompanying preposition “alla” that would be the equivalent to the English “to”.

A lexical item that clearly demonstrates how one language may be advantaged can be as follows:

Guarda bene questi bambini e l'acqua. Fammi vedere quale di questi bambini si sta ***allontanando*** *dall'acqua... Fammi vedere quale di questi bambini si sta* ***allontanando*** *dall'acqua. Look at the children and the water. Point to the child who is walking* ***away*** *from the water…Point to the child who is walking* ***away*** *from the water*.

The Italian item is self-evident in its meaning. The verb "sta allontanando" means to go away from in Italian, this possibly giving Italian children an advantage on this item.

Lastly, nearly every item between the Italian and U.S. versions is bound to have morphological differences by mere linguistic fact. These linguistic differences are especially helpful when gender or number is identified or reinforced in the prompts, such as the following item:

Dite al bambino: <<Metti il giocattolo **all'angolo** della sedia>>.

Say to the child, "Put the toy on the **corner** of the chair."

In this item, the gender of the location where the toy must be placed is indicated in Italian both by the noun "sedia" itself and the preposition "della". There is a chance this may partially cue the answer.

Despite these cases, the results of the analysis indicate that as a whole, the linguistic flags tended to slightly over-predict DIF, producing a flag in 39 spots, where 36 should be predicted. These results are displayed in Table 6 below.

Table 6. All Linguistic Flags (Observed)

	Linguistic DIF	No Linguistic DIF		
Logisitc Flag	21	36	57	
No Logisitc Flag				
	18	93	111	
	39	129	168	X2=0.005

Despite over-predicting DIF, a Chi-square shows that the is a relationship between linguistic flag and logistic flag. It seems clear that linguistic flags have an associative, if not predictive, relationship with DIF.

At an individual level as well, each category within linguistic analysis produces statistically significant results. Each category; be it syntax, morphology or lexicon shows sensitivity toward predicting DIF in logistic regression. Each linguistic category proves to be significant, as shown in Table 7.

Table 7. Linguistic Analysis of DIF by Category.

Syntax

	DIF	No DIF		
LR DIF	10	2	12	
NO LR DIF	7	30	37	
	17	32	49	X^2=0.0009

Morphology

	DIF	No DIF		
LR DIF	6	6	12	
NO LR DIF	6	31	37	
	12	37	49	X^2=0.0000

Lexicon

	DIF	No DIF		
LR DIF	5	7	12	
NO LR DIF	5	32	37	
	10	39	49	X^2=0.0000

Discussion

This project sought to answer two questions regarding the potential for a non-psychometric approach to analyzing CLI comparability. First, the research attempted to answer whether factor analytic and content analysis procedures could identify substantive explanations for observed differences in the differential dimensional structure of translated tests. Second, the study also tried to show that language structure analysis may be used to predict DIF in translated tests.

With respect to dimensional differences, the factor analytic study concluded that there is indeed a different structure between the two language versions, and presumably by extension between multiple languages of a CLI. Common ground between the languages on a larger principal factor may have interesting implications upon concepts such as Universal Testing Constructs, as proposed by Reckase, Li and Joldersma (2004, in progress). The remaining factors, which may be attributable to language differences, cannot be conclusively attributed to documented cultural or linguistic phenomenon. Still, the fact that there is difference in the dimensional structure between the languages warrants further study into the matter.

The language structural analysis for DIF identification, had notable results. It showed that the individual categorical (syntactic, morphological and lexical) have a demonstrable relationship both with logistic regression modeling and in a simple Chi-square test of association. This suggests that DIF indices, working cooperatively with linguistic analysis are an effective technique for helping to ensure CLI comparability. As a consequence, linguistic analyses bolster the validity of the results of traditional DIF procedures by providing a theoretical foundation for the items that are flagged by these psychometric methods (as per Messick).

The difficulty is to determine whether this is truly due to linguistic complexity or poor translation. However, by further examining CLI comparability with linguistic analyses, patterns in the DIF exhibited may be detected and rules may be created as a plausible explanation for these differences. These 'rules' may further substantiate and provide theoretical support to the findings of psychometric analyses.

Limitations

A primary limitation to this study is the limited nature of the sample. With only two languages and 49 items to discriminate between, this study is only a first step to providing a more substantial method for CLI comparability. Hopefully, linguistic analysis as a comparability methodology will generalize to other DIF procedures and language contexts (as well as other instruments) once it is implemented properly.

One issue of concern is that some of the items that were flagged as likely to have DIF by linguistic analysis, but were not indeed flagged by the psychometric procedures. It is not known if this is an indication of erroneous flagging on the part of the author or sensitivity to DIF not predicted by traditional methods. This question is beyond the scope of this investigation. Methods for the resolution of conflicts between the two procedures also have potential implications for better understanding of CLI comparability.

Future Research

Future research will likely focus on the implementation of a confirmatory factor analysis to determine if there are substantively interpretable models which better fit CLI data, and thus improve comparability. In addition, another direction for future studies is to do an additional analysis of the linguistic codes for disconfirming cases in order to identify possible elaborations on any rules that might be developed for the purpose of explaining why DIF exists. Additionally, efforts to ensure generalizability across multiple languages, rather than the two focused upon here will be performed to gauge the usefulness

of the linguistic coding system developed. Moreover, the extension of this methodology to other instruments could also further the generalizability of the technique. A major goal of future research would be to provide substantive rules and/or a hierarchy of linguistic features to see if there is systematicity to the DIF exhibited in CLIs. Lastly, the extension to other DIF techniques may prove more beneficial to evaluate the success of linguistic analysis and CLI comparability.

References

Allalouf, Avi. (2000). Retaining Tranlated Verbal Reasoning Items by Revising DIF Items. Israel; 2000-04-00. 23 p. Paper Presented at AERA (New Orleans, LA, April 24-28, 2000).

Auchter, Joan E. & Stansfield, Charles W. (1997). Developing Parallel Tests across Languages: Focus on the Translation and Adaptation Process. Version of a paper presented at the Annual Large Scale Assessment Conference (27th, Colorado Springs, CO, June 15-18, 1997).

Budgell, Glen R. (1995). Analysis of Differential Item Functioning in Translated Assessment Instruments. *Applied Psychological Measurement* v19 n4 p309-21 Dec 1995

Clauser, B.E. & Mazor, K.M. (1998). An NCME Introduction Module on Using Statistical Procedures to Identify Differentially Functioning Test Items. Princeton, NJ: ETS.

Draba, R.E. (1977). The identification and interpretation of item bias (Research Memorandum No. 26). Chicago, IL: University of Chicago.

Durán, R.P., Canale, M., Pennfield, J., Stansfield, C., & Liskin-Gasparro, J.E. (1987). TOEFL From a Communicative Viewpoint on Language Proficiency: A Working Paper. Found in: Cognitive and Linguistic Analyses of Test Performance. Freedle, R.O. and Durán R.P., Eds. (1987). Norwood, New Jersey: Ablex Publishing Company.

Flippo, Hyde. (2003). Date accessed: June 22, 2003. http://german.about.com/library/blcase_acc.htm. The Four German Cases. *About.com*

Gutierrez-Clellen, Vera F. ; Hofstetter, Richard (1994). "Syntactic Complexity in Spanish Narratives: A Developmental Study". *Journal of Speech and Hearing Research* v37 n3 p645-54 Jun 1994.

Hambleton, Ronald K. & Patsula, Liane. (2000). Adapting Tests for Use in Multiple Languages and Cultures. Laboratory of Psychometric and Evaluative Research Report. U,S,; Massachusetts; 2000-01-00. Corp. Authors: Massachusetts University, Amherst. School of Education.

Hambleton, R. K., & de Jong, J. H. A. L. (Eds.). (2003). Advances in translating and adapting educational and psychological tests [Special issue]. Language Testing 20 (2). Hambleton, R. K., & de Jong, J. H. A. L. "Advances in translating and adapting educational and psychological tests" (127-134); Zumbo, B. D. "Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests" (136-147); Sireci, S. G., & Allalouf, A. "Appraising item equivalence across multiple languages and

cultures" (148-166); Lokan, J., & Fleming, M. "Issues in adapting a computer-assisted career guidance system for use in another country" (167-177); Chae, S. "Adaptation of a picture-type creativity test for pre-school children" (178-188); Stansfield, C. W. "Test translation and adaptation in public education in the USA" (189-207); McQueen, J., & Mendelovits, J. "PISA reading: Cultural equivalence in a cross-cultural study" (208-224); Grisay, A. "Translation procedures in OECD/PISA 2000 international assessment" (225-239).

Kay, P. (1987). Three Properties of the Ideal Reader. Found in: Cognitive and Linguistic Analyses of Test Performance. Freedle, R.O. and Durán R.P., Eds. (1987). Norwood, New Jersey: Ablex Publishing Company.

Maggi, Stefania. (2001). "Italian Version of the Self-Description Questionnaire-III". *International Journal of Testing* v1 n3&4 p245-48 2001

Martin, L. (1986). "Eskimo words for snow": A case study in the genesis ans decay of an anthropological example. American Anthropologist, 88:418-423.

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

Price, Larry R. & Oshima, T. C. (1998). Differential Item Functioning and Language Translation: A Cross-National Study with a Test Developed for Certification. Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

Price. (1999). Differential Functioing of Items and Tests versus the Mantel-Haenszel Technique for Detecting Differential Item Functioning in a Translated Test. U.S.; Georgia; 1999-04-00. Paper presented at the Annual Meeting of the American Alliance of Health, Pychical Education, Recreation, and Dance (Boston, MA, April 12-16).

Raju, N., van der Linden, W. & Fleer, P. (1992). As quoted in: Price (1999). Differential Functioning of Items and Tests versus the Mantel-Haenszel Technique for Detecting Differential Item Functioning in a Translated Test. Paper presented at the Annual Meeting of the American Alliance of Health, Physical Education, Recreation, and Dance (Boston, MA, April 12-16, 1999).

SAS software (SAS Institute Inc. 2004).

Sireci, Stephan G. & Swaminathan, Hariharan. (1996). Evaluating Translation Equivalence: So What's the Big Dif? Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 1996).

Sireci, Stephen G & Khaliq, Shameem Nyla. (2002). An Analysis of the Psychometric Properties of Dual Language Test Forms. Corp Author(s): Massachusetts Univ., Amherst. School of Education. U.S.; Massachusetts

Sireci, Stephen G. (1997). Problems and Issues in Linking Assessments across Languages. *Educational Measurement: Issues and Practice* v16 n1 p12-19,29 Spr 1997

Thurstone, L.L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago Press.

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

VanPatten, Bill and Cadierno, Teresa. (1993). Explicit Instruction and Input Processing. *Studies in Second Language Acquisition* v15 n2 p225-43 Jun 1993.

WINSTEPS. (1999). Rasch-Model Computer Program. Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis. Chicago: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). Best test design, Rasch measurement. Chicago: Mesa Press.

Zumbo, Bruno D. (1999). A Handbook on the Theory of Differential Item Functioning (DIF): Logisitc Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

APPRENDIX I: Logistic Regression DIF results

			uniform DIF				**non-uniform DIF**			
Linguistic Analysis Predicted			**country main effect**				**country inter-action**			
	item	**separation**	**Estimate**	**st error**	**wald**	**p-value**	**estimate**	**st error**	**Wald**	**p-value**
0	1	yes								
0	2	yes								
0	3	no				0.5049				0.6744
0	4	no				0.7218	0.0612	0.0368	2.7578	**0.0968**
0	5	no				0.2386				0.5334
2	6	no	3.1774	1.3726	5.3588	**0.0206**	-0.0585	0.0347	2.8394	0.092
1	7	no				0.584				0.6668
2	8	no	-4.7365	2.3598	4.0288	**0.0447**	0.1275	0.064	3.9756	**0.0462**
1	9	no				0.5228				0.5002
1	10	no				0.285				0.3982
0	11	No				0.735				0.7005
1	12	No				0.4352				0.4609
1	13	no				0.7022				0.5845
2	14	no				0.3867				0.1448
1	15	no				0.3965				0.48
0	16	no				0.9741				0.4609
0	17	no				0.508				0.122
1	18	no				0.8875				0.8411

1	19	no				0.752				0.8978
0	20	no				0.3459				0.2251
1	21	no				0.7214				0.8073
0	22	no				0.5084				0.7284
1	23	no				0.1239	-0.104	0.0576	3.2587	**0.071**
0	24	no				0.2397				0.2161
0	25	no				0.4102				0.4773
2	26	no	3.0905	1.5961	3.7491	**0.0528**				0.1937
1	27	no	2.1815	1.3468	2.6238	**0.1053**				0.1773
2	28	no	-5.7975	2.0538	7.9685	**0.0048**	0.1056	0.051	4.287	**0.0384**
0	29	no				0.5784				0.6332
0	30	no				0.5508				0.2352
0	31	no				0.9895				0.4486
0	32	no				0.5983				0.593
0	33	no				0.3633				0.8472
0	34	no				0.2433				0.3224
1	35	no				0.6051				0.7711
2	36	no	3.4788	1.3801	6.3538	**0.0117**	0.0956	0.0336	8.0873	**0.0045**
0	37	no				0.2266				0.1211
0	38	no				0.2885				0.1764
0	39	no				0.641				0.7512
0	40	no				0.4232				0.376
1	41	no	-5.0857	2.1422	5.6358	0.0176	0.1347	0.0581	5.3691	**0.0205**
1	42	no				0.1235	-0.072	0.0319	5.0949	**0.024**
0	43	no	2.7458	1.4052	3.8182	**0.0507**	-0.0599	0.0335	3.1883	0.0742
0	44	no				0.2839				0.5105
2	45	no	-4.3883	1.9642	4.9916	**0.0255**	0.1075	0.0541	3.9512	**0.0468**

0	46	no				0.1222				0.3082
0	47	no				0.2972				0.3734
0	48	no				0.5275				0.4419
2	49	no				0.4896				0.3463
0	50	no				0.2928	-0.0647	0.0361	3.2054	**0.0734**
2	51	no	1.116	0.1781	39.2826	**<.0001**	-0.0407	0.00608	44.9529	**<.0001**

APPENDIX II: Linguistic Coding

Item	**Item name**	**DIF with LR**	**DIF predicted 0,1,2**	**Syntax: 0,1**	**Morphology: 0,1**	**Lexicon: 0,1**
3	S3	0	0	0	0	0
4	S4	0	0	0	0	0
5	S5	0	0	0	0	0
6	S6	1	2	0	1	1
7	S7	0	1	1	0	0
8	S8	1	2	1	1	0
9	S9	0	1	1	0	0
10	S10	0	1	1	0	0
11	S11	0	0	0	0	0
12	S12	0	1	1	0	0
13	S13	0	1	0	1	0
14	S14	0	2	0	1	1
15	S15	0	1	0	0	0
16	S16	0	0	0	0	0
17	S17	0	0	0	0	0
18	S18	0	1	0	1	0
19	S19	0	1	0	0	1
20	S20	0	0	0	1	0
21	S21	0	1	0	1	1
22	S22	0	0	1	0	0
23	S23	1	1	1	0	1
24	S24	0	0	0	0	0
25	S25	0	0	0	0	0

26	S26	1	2	1	0	1
27	S27	0	1	0	0	1
28	S28	1	2	1	0	0
29	Q1	0	0	0	0	0
30	Q2	0	0	0	0	0
31	Q3	0	0	0	0	0
32	Q4	0	0	0	0	0
33	Q5	0	0	0	0	0
34	Q6	0	0	0	0	0
35	Q7	0	1	0	1	0
36	Q8	1	2	1	1	0
37	Q9	0	0	0	0	0
38	Q10	0	0	0	0	0
39	Q11	0	0	0	0	0
40	Q12	0	0	0	0	0
41	Q13	1	1	1	0	0
42	Q14	1	1	1	1	0
43	Q15	1	0	0	0	0
44	T1	0	0	1	0	0
45	T2	1	2	1	1	1
46	T3	0	0	0	0	0
47	T4	0	0	0	0	0
48	T5	0	0	0	0	0
49	T6	0	2	1	0	1
50	T7	1	0	1	0	0
51	T8	1	2	1	1	1
	Total:	12		17	12	10