DOCUMENT RESUME

ED 481 107 TM 035 268

AUTHOR	Sireci, Stephen G.; Gonzalez, Eugenio J.
TITLE	Evaluating the Structural Equivalence of Tests Used in International Comparisons of Educational Achievement.
PUB DATE	2003-04-00
NOTE	29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003). Paper represents Center for Educational Assessment Research Report No. 490.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Comparative Analysis; Difficulty Level; International Education; *International Studies; Multidimensional Scaling; Test Construction; *Test Format; Test Items; *Translation
IDENTIFIERS	*Third International Mathematics and Science Study

ABSTRACT

International comparative educational studies make use of test instruments originally developed in English by international panels of experts, but that are ultimately administered in the language of instruction of the students. The comparability of the different language versions of these assessments is a critical issue in validating the comparative inferences drawn from the test results. This study analyzed data from the 1999 Third International Mathematics and Science Study (TIMSS) science assessment to evaluation the consistency of the structure of the item response data across different test versions. Individual differences multidimensional scaling analyses were used to evaluate data structure. The findings suggest that slight structural differences exist across countries, and that these differences are related to differences in item difficulty. The implications of these findings for better understanding of international comparisons of educational achievement, and for future research in this area, are discussed. (Contains 1 figure, 7 tables, and 26 references.) (Author/SLD)

Evaluating the Structural Equivalence of Tests Used in International Comparisons of Educational Achievement[1]

Stephen G. Sireci

University of Massachusetts Amherst

Eugenio J. Gonzalez

Boston College and University of Massachusetts Amherst

Correspondence concerning this article should be addressed to Stephen G. Sireci, Center for Educational Assessment, School of Education, University of Massachusetts, Amherst, MA, 01003-4140. E-mail correspondence may be sent to Sireci@acad.umass.edu.

[1] Paper presented at the Annual Meeting of the National Council on Measurement in Education as part of the symposium "Test Adaptations and Translations," (Don Klinger, chair), April 24, 2003, Chicago, IL. This paper represents Center for Educational Assessment Research Report No. 490, School of Education, University of Massachusetts, Amherst, MA. Correspondence concerning this article can be sent to Sireci@acad.umass.edu.

Evaluating the Structural Equivalence of Tests Used in International Comparisons of Educational Achievement

Abstract

International comparisons of educational achievement are important for gauging the academic skills of students within a global context. Current studies of educational achievement, such as TIMSS, PIRLS, and PISA administer tests in mathematics, science, and reading to students in participating countries. These international comparative studies make use of test instruments originally developed in English by international panels of experts, but that are ultimately administered in the language of instruction of the students. The comparability of the different language versions of these assessments is a critical issue in validating the comparative inferences drawn from the test results. In this paper, we analyze data from the 1999 TIMSS Science assessment to evaluate the consistency of the structure of the item response data across different test versions. Individual differences multidimensional scaling analyses were used to evaluate data structure. The findings suggest that slight structural differences exist across countries and that these differences are related to differences in item difficulty. The implications of these findings for better understanding international comparisons of educational achievement, and for future research in this area, are discussed.

Keywords: cross-cultural assessment, dimensionality, multidimensional scaling, test translations, validity,

Evaluating the Structural Equivalence of Tests Used in International Comparisons of Educational Achievement

Educational assessments are an important component of educational systems throughout the world. Such assessments are used for many purposes such as certifying student competence and evaluating education reform movements. Over the past few decades, several international comparisons of students' educational achievement have occurred. Examples include the Third International Mathematics and Science Study (TIMSS) (repeated in 1999 in 38 countries and currently being repeated again in 2003), the Program for International Student Assessment (PISA, (Organization for Economic Co-operation and Development, 2000)), which assesses the reading, math, and literacy skills of 15-year olds in 32 countries, and the Progress in International Reading Literacy Study (PIRLS, (Campbell, et al., 2001)), which assesses the reading skills of fourth-grade students in approximately 40 countries.

An example of the type of information provided by these studies is presented in Table 1. The data in Table 1 come from the 1999 TIMSS Science Assessment for thirteen year-old students. The major findings reported in the newspapers and quoted by politicians is the rank ordering of the countries. One thing to note is that the top five countries all speak different languages. To compare individuals who speak different languages, tests must be translated (adapted) for use across multiple languages. Although such adaptations make cross-lingual assessment possible, the use of different versions of a test confounds group achievement differences with test differences. For this reason, many researchers argue that the comparability of test scores across languages must be established before making comparative inferences across groups (Bechger, van den Wittenboer, Hox, & De Glopper, 1999; Geisinger, 1994; Hambleton, 1994; Sireci, 1997; van de Vijver & Tanzer, 1998). This concern is echoed by the *Standards for*

Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), which states "when a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested" (p. 99).

For the high-profile international assessments such as TIMSS and PIRLS, the methods used to adapt tests across languages are comprehensive and include several quality control steps (O'Connor & Malak, 2000). However, the complex nature of these assessments makes it difficult to statistically compare the psychometric characteristics of the different test versions. For example, the balanced incomplete blocks (BIB) test booklet administration spiraling design, which makes it possible to administer a larger number of items to students within each country, makes it difficult to evaluate data structure due to student-by-item data matrices with large amounts of missing data. Coupled with the fact that cross-lingual assessment requires comparing different individuals who took different items with no common anchor, evaluating psychometric comparability is not straightforward (Sireci, 1997). A primary purpose of this study is to illustrate some methodological options for conducting such studies.

Issues and Standards in Cross-lingual Assessment

In addition to requiring that scores from adapted tests possess adequate reliability and validity *within* each language, the *Standards for Educational and Psychological Testing* also state "when multiple language versions of a test are intended to be comparable, test developers should provide evidence of score comparability[2]" (p. 99). Van de Vijver and his colleagues stated that at least three types of bias can lead to non-comparability of scores across languages:

construct bias, method bias, and item bias (van de Vijver & Poortinga, 1997; van de Vijver & Tanzer, 1998). Construct bias refers to the situation where the construct measured, as operationally defined by the assessment, is nonexistent in one or more cultures or is significantly different across cultures. This source of bias is typically not of concern in most large-scale comparisons of educational achievement[3]. Method bias refers to a systematic source of construct-irrelevant variance that manifests itself at the test score level. Examples of method bias include improper test administration conditions, inappropriate or unfamiliar item formats, or improper test translations that make all items easier or difficult in one language, relative to another language. Item bias refers to construct-irrelevant variance that affects performance at the item level. In this paper, we focus on the evaluation of method bias, which may manifest itself through differences in the dimensionality (structure) of the data obtained from different versions of a test.

Proper test translations go a long way toward avoiding method and item bias. However, empirical studies are needed to rule out such biases. For this reason, the *Guidelines for Adapting Educational and Psychological Tests* (developed by the International Test Commission, see Hambleton, 1994, 2001) underscore the need for statistical procedures to evaluate test comparability across cultures and languages. The *Guidelines* encourage test developers to use appropriate statistical techniques to evaluate item equivalence and to identify areas of a test that may be inadequate for one or more of the intended groups. For example, the *Guidelines* recommend that test developers conduct differential item functioning (DIF) analyses to evaluate test items designed to be used in two are more cultural or language groups. DIF analyses evaluate whether examinees from different groups (e.g., LEP or non-LEP) who are of comparable ability have equal probabilities of success on an item. Although DIF analyses are useful for identifying problematic items, an evaluation of the dimensionality of adapted tests is prerequisite for ruling out systematic biases at the total test score level that are not detectable at the item level (Sireci, 1997, in press; van de Vijver & Tanzer, 1998).

In this study, we illustrate how the recommendations of the AERA et al. *Standards* and the ITC *Guidelines* can be met on complex international comparisons of educational achievement. Specifically, we evaluate the comparability of test structure across languages and countries to explore the possible presence of method bias due to test translation. The results of such analyses should provide information useful for (a) evaluating the comparability of the different versions of a test used in international comparisons of educational achievement, and (b) understanding how structural differences may relate to educational differences across countries.

Method

Test Data

Our illustration of methods that can be used for evaluating different language versions of a test used in international assessments involves the publicly available 1999 TIMSS Science Assessment data (Gonzalez & Miles, 2001). TIMSS 1999 represents the continuation of a long series of studies conducted by the International Association for the Evaluation of Educational Achievement (IEA). The Third International Mathematics and Science Study (TIMSS), conducted in 1994-1995 assessed mathematics and science at third and fourth grades, seventh and eighth grades, and the final year of secondary school. In 1998-1999, TIMSS again assessed eighth-grade students in both mathematics and science to measure trends in student achievement since 1995. TIMSS 1999 is also known as TIMSS-Repeat or TIMSS-R.

The curriculum framework underlying the TIMSS Science tests was developed in 1995 by groups of science educators with input from the TIMSS National Research Coordinators. The two dimensions of the Science curriculum framework relevant to this study are presented in Table 2. The *content* dimension represents the subject matter content of school science. The *performance expectations* dimension describes, in a non-hierarchical way, the many kinds of performance or behavior that might be expected of students in school science.

Table 3 presents the six content areas in the TIMSS 1999 science test, together with the number of items and score points in each area, broken out by item type. There were a total of 146 items across the eight test booklets. About one-fourth of the items were in the free-response format, requiring students to generate and write their own answers. Some free-response questions asked for short answers while others required extended responses with students showing their work or providing explanations for their answers. The remaining questions were in the multiple-choice format. Correct answers to most questions were worth one point. Consistent with longer response times for the constructed-response questions, however, responses to some of these questions (particularly those requiring extended responses) were evaluated for partial credit, with a fully correct answer being awarded two points. Thus, the number of possible score points available for analysis exceeds the number of items.

TIMSS Test Design

To ensure broad subject matter coverage without overburdening students, TIMSS used a rotated test booklet design that included both the mathematics and science items (Gonzalez & Miles, 2001). Thus, mathematics and science items appeared in different sections of the same test booklet. The 1999 assessment consisted of eight booklets, each requiring 90 minutes of response time. Each student was assigned one booklet only. In accordance with the design, the mathematics and science items were assembled into 26 clusters (labeled A through Z). Each booklet had about 60-70 items, half of which were science items.

Test Translation and Verification

The TIMSS 1999 instruments were prepared in English and translated into 33 languages. Ten of the 38 countries collected data in two languages. In addition, the international versions sometimes needed to be modified for cultural reasons, even in the nine countries that tested in English. This process represented an enormous effort for the national centers, with many checks along the way. The translation effort included developing explicit guidelines for translation and cultural adaptation; translation of the instruments by the national centers in accordance with the guidelines, using two or more independent translators; consultation with subject matter experts on cultural adaptations to ensure that the meaning and difficulty of items did not change; verification of translation quality by professional translators from an independent translation company; corrections by the national centers in accordance with the suggestions made; verification by the International Study Center that corrections were made; and a series of statistical checks after the testing to detect items that did not perform comparably across countries.

Student Data

In 1995 one of the TIMSS target populations was students enrolled in the two adjacent grades that contained the largest proportion of 13-year-old students at the time of testing, corresponding to seventh- and eighth-grade students in most countries. TIMSS in 1999 used the same definition to identify the target grades, but assessed students in the upper of the two grades only, the eighth grade in most countries. For the purposes of this study, we focused on the data from nine countries: Belgium (Flemish), Canada, England, Hong Kong, Italy, Japan, Korea, Russia, and the United States. These countries were chosen because they represent the countries members of the G7 group (i.e., economic partners of the U.S.), plus the 4 highest achievers from the 1995 assessment. Both English- and French-language versions of the test were used in Canada, and three language versions of the test were used in Hong Kong[4]. Due to the relatively small sample sizes for the tests answered in English or both languages, only the Chinese data were used to represent Hong Kong. Thus, our comparisons involved ten groups: Belgium (Flemish), Canadian-English, Canadian-French, England, Hong Kong (Chinese), Italy, Japan, Korea, Russia, and the United States. The sample sizes for these groups ranged from 2,437 (French Canadian) to 9,072 (United States).

IRT Scaling and Data Analysis

The reporting of the TIMSS achievement data was based primarily on item response theory (IRT) scaling methods. The mathematics and science results were summarized using a family of 2-parameter and 3-parameter IRT models for dichotomously scored items (right or wrong), and 2-parameter generalized partial credit models for items with more than 1 available score point. The IRT scaling method produces a score by averaging the responses of each student to the items in the student's test booklet in a way that takes into account the difficulty and discriminating power of each item. The method used in TIMSS includes refinements that enable reliable scores to be produced even though individual students responded to relatively small subsets of the total item pool. Achievement scales were produced for six content areas (earth science, life science, physics, chemistry, environmental and resource issues, and scientific inquiry and the nature of science), as well as for science overall.

To allow more accurate estimation of summary statistics for student subpopulations, the TIMSS scaling made use of plausible-value technology (Yamamoto & Kulick, 2000), whereby five separate estimates of each student's score were generated on each scale, based on the responses to the items in the student's booklet and the student's background characteristics. The five score estimates are known as “plausible values,” and the variability between them encapsulates the uncertainty inherent in the measuement.

Data Analyses

The structural analyses were constrained by the incomplete student-by-item data matrices that resulted from the BIB spiral design. For the structural analyses, we used principal components analysis, principal axis factoring, and multidimensional scaling (MDS).

Analysis of Test Structure

Our analysis of test structure focused on comparing the structure of students’ item response data across the different versions of the tests administered in the different countries. These analyses were conducted as a means of discovering whether any method bias was present. Two potential sources of such bias are unintended differences introduced through the test adaptation process or to non-uniform test administration conditions. When the structure of test data differs across groups of examinees, the test may not be measuring the same constructs across groups.

Given the incomplete data, there are several options for analyzing the structure of the data, none of which are perfect. To derive a matrix of inter-item correlations, inter-item correlations or distances could be computed pairwise for the eight major content clusters, because the BIB design paired each of these clusters with one another across the eight test booklets. We looked at test structure across the eight major content clusters using pairwise inter-item correlations (for the factor analyses) and inter-item distances (for the MDS analyses). There were 46 items across the eight major clusters that were taken by all ten groups of examinees.

Principal components and common factor analyses

Principal components analysis (PCA) and principal axis factor analysis (FA) were used to acquire a general idea of the dimensionality of the data across all groups. For these analyses, inter-item (Pearson) correlations were computed, ignoring the subgroups that differed by language of the test. A limitation of these analyses is that they must be done separately for each group (cf. Zumbo, Sireci, & Hambleton, 2003) and so these preliminary analyses were done to see if there were any observable multidimensionality in the group of 46 items from the eight major content clusters. To evaluate the dimensionality of the data, the size of the eigenvalues associated with the first and subsequent factors, and the percentages of variance accounted for by these factors, were compared. The factor loading matrix was also evaluated for interpretability.

Multidimensional scaling analyses

The purpose of the MDS analyses was to discover the structure of the data simultaneously across groups while also accounting for differences in structure across the groups. To do this, a weighted multidimensional scaling (WMDS) procedure was used. WMDS analyzes several matrices of "dissimilarity" data to derive both a common structure that best represents the data for all groups and individual group weights for adjusting this common structure to best fit the data for each specific group. In this study, the groups were defined by country or language of the test within a country. The weights for each group were used to compare the relevance of the dimensional structure across groups.

For the WMDS analyses, Euclidean distances were computed among the items separately for each group of students[5]. This process provided a symmetric inter-item distance matrix for each group. Davison (1985) and Davison and Skay (1991) showed that when using MDS to analyze inter-item correlations, if a general factor were present, it drops out of the MDS solution.

Given that these tests were constructed to be unidimensional, and given that De Ayala and Hertzog (1991) and Meara, Robin, and Sireci (2000) found that the MDS analysis using inter-item Euclidean distances performed well in uncovering dimensionality, inter-item distances, rather than inter-item correlations, were used.

The INDSCAL WMDS model (Carroll & Chang, 1970) was used for all WMDS analyses. This model specifies a weighted Euclidean distance formula to scale the items:

$$d_{ijk} = \sqrt{\sum_{a=1}^{r} w_{ka} (x_{ia} - x_{ja})^2} \qquad [1]$$

where: d_{ijk}=the Euclidean distance between items *i* and *j* for group *k*, w_{ka} is the weight for group *k* on dimension *a*, x_{ia}=the coordinate of item *i* on dimension *a*, and *r*=the dimensionality of the model. A common structural space, called the group stimulus space, is derived for the stimuli. The "personal" distances for each group are related to the common stimulus space by:

$$x_{kia} = \sqrt{w_{ka}}\, x_{ia} \qquad [2]$$

where x_{kia} represents the coordinate for item *i* on dimension *a* in the personal space for group *k*, w_{ka} represents the weight of group *k* on dimension *a*, and x_{ia} represents the coordinate of stimulus *i* on dimension *a* in the common stimulus space.

Differences in dimensional structure across groups are reflected in the group weights (i.e., w_{ka}). The larger a weight on a dimension (*a*), the more that dimension is necessary for accounting for the variation in the data for the specific group (*k*). All analyses were implemented using the ALSCAL program in SPSS, version 11.1 (Young & Harris, 1993). The nonmetric option was used, to maximize the fit of the data to the MDS model. In the INDSCAL model implemented in SPSS, the group weights can range from zero to one. A weight of zero indicates the dimension is completely irrelevant to the data for the group. A weight of one indicates the MDS coordinates on that dimension completely account for the variation in the data for that group. Using simulated data, Sireci, Bastari, & Allalouf (1998) found that when structural differences exist across groups on one or more dimensions, one or more groups will have weights near zero, while other groups will have noticeably larger weights. They concluded non-equivalence of the structure of an assessment across groups should be obvious via inspection of the MDS weights.

The STRESS and R^2 fit indices were used to identify solutions that provide reasonable fit. STRESS represents the square root of the normalized residual variance of the monotonic regression of the MDS distances on the transformed item dissimilarity data. Thus, lower values of STRESS indicate better fit. The R^2 index reflects proportion of variance of the transformed dissimilarity data accounted for by the MDS distances. Thus, higher values of R^2 indicate better fit. There are no absolute guidelines for determining adequate fit in MDS, but simulation research conducted by MacCallum (1981) for weighted, non-metric MDS fitted using ALSCAL provides some guidance. For example, MacCallum (1981) provides an equation for estimating the expected level of STRESS for random data, given the number of items scaled, the number of dissimilarity matrices, and the number of MDS dimensions (from two through five dimensions). MacCallum's formula was used to compute expected levels of STRESS, where possible.

Results

As mentioned earlier, a major focus of our analysis was on the eight content clusters that were paired with one another in at least one test booklet. These analyses involved 46 items that represented all of the content distinctions present in the test frameworks.

PCA and FA

The PCA and FA gave very similar results, even considering oblique and orthogonal rotations of the FA solution. In general, the results suggested a dominant general factor that accounted for about 12% of the variance in the observed inter-item correlations, along with several minor factors that accounted for about 2-3% of the variance. A summary of the PCA results is presented in Table 4 (the FA results were nearly identical). The ratio of the first eigenvalue to the second was about 4:1. These results suggest the presence of a weak, but dominant first factor along with several minor factors.

To help interpret the PCA and FA results, data summarizing item characteristics were correlated with the component and factor loadings. The item characteristic data included dummy variables representing content area, performance expectations, and IRT item difficulty estimates for each country. The dummy variable that distinguished between Physics items and other items had a small correlation with the first component (r=-.38) and item some of the country-specific IRT difficulty estimates exhibited large-to-moderate correlations with the secondary components. The largest of these correlations were the correlation between the difficulty estimates calibrated from the U.S. data and the item loadings on component 2 (r=-.78) and the correlation between the difficulty estimates calculated from Belgium and the item loadings on component 3 (r=.84). Group-specific item difficulty factors were more salient in the MDS analyses and so we temporarily forestall discussion of this issue. No other correlations were statistically significant, which eliminated interpretation of the minor multidimensionality from a content perspective.

MDS Results

A summary of the MDS fit statistics is presented in Table 5. The two-dimensional solution had the greatest difference from the STRESS expected from random data and was associated with the greatest improvement in STRESS from a lower-dimensional solution. However, the STRESS and R^2 values indicate that there is still much variation among the items that is not accounted for using these two dimensions. An analysis of the group weights across the two- through six-dimensional solutions suggested that the improvement in fit was due to improved modeling of the item difficulties across languages.

As with the PCA, the data summarizing item characteristics were correlated with the WMDS item coordinates. None of the content area or performance expectation dummy variables exhibited significant correlations with the dimensions. However, the country-specific item difficulty estimates were all significantly correlated with the coordinates on each dimension. These difficulty/coordinate correlations are presented in Table 6, along with the group weights. With respect to the group weights, all ten groups had weights above zero in the first dimension, but Japan, Korea, and Belgium have higher weights on the second dimension. There is an interesting positive relationship between the group weights and the country-specific item difficulty estimates. In general, the difficulty/coordinate correlations are larger on the dimension on which the group has the larger weight. In fact, the correlation between the group-specific difficulty/coordinate correlations and the group weights is .97 for Dimension 1 and .91 for Dimension 2. This finding suggests that the multidimensionality is associated with differences in the rank ordering of the item difficulties across groups.

The relationship between group weights on the MDS dimensions and group-specific item difficulty estimates increases with dimensionality. Table 7 presents the group weights and difficulty/coordinate correlations for the five-dimensional MDS solution. A similar pattern of weights and correlations is evident for the English versions of the exam (i.e., Canadian English, England, and U.S.) and some countries have idiosyncratically large weights on a single dimension (i.e., Hong Kong on Dimension 2, Japan on Dimension 5, Russia on Dimension 4, and Belgium on Dimension 3). Across the five dimensions, the correlations between the difficulty/coordinate correlations and the group weights ranged from .85 (Dimension 3) to .99 (Dimension 2). It is interesting to note that the Canadian English and Canadian French versions of the exam have very similar patterns of MDS weights and coordinate/item difficulty correlations.

A two-dimensional sub-space of the five-dimensional MDS weight space is presented in Figure 1. This figure illustrates the importance of Dimension 2 to the Hong Kong data and the relative importance of Dimension 1 to the data from the Romance language versions of the test, particularly English. However, it should be noted that these dimensions are "smaller" than those from the two-dimensional solution in that the inter-item variance accounted for by each dimension ranged between 12% and 18% (see Table 7).

In general, the MDS results seem to suggest that (a) the gross (two-dimensional) structure of the data is fairly consistent across groups and the minor multidimensionality is related to differences in item difficulty across groups, and (b) as more dimensions are fit to the data, more subtle structural differences are revealed and these differences also stem from differences in item difficulty across groups. Given the similar pattern of dimension weights across the three groups who took the English-language version of the exam, item differences introduced through the test adaptation process should be investigated as a source of multidimensionality. However, instructional practices may also be similar across these countries, which could also be linked to multidimensionality.

Discussion

This study represents a relatively new exploration into the similarity of data structure across different language versions of a TIMSS test. From a methodological perspective, we learned MDS is useful for comparing data structure simultaneously across groups, and that when subtle structural differences are of interest, it provides a pretty powerful microscope. The results also suggest that important information regarding structural equivalence can be obtained by analyzing a subset of items. This is a significant finding, given the BIB spiral design used in TIMSS, NAEP, and other large-scale assessments. However, methods for analyzing structure across the whole item pool are needed.

At this juncture, it is difficult to ascertain the magnitude of the structural differences or their substantive importance. The largest structural differences were observed as more dimensions were fit to the data, but the "size" of each dimension diminished accordingly. Thus, the largest group differences were noted on the smallest dimensions. Nevertheless, the fact that the structural differences were related to differences in item difficulty suggests that the multidimensionality stems from a logical cause. Two potential causes are differences in item difficulty due to the test translation process or differences in the familiarity of the item due to differences in curricula. Analysis of the content of the different language versions of the items, and analysis of differential item functioning across languages, may prove illuminating in identifying the causes of multidimensionality (Allalouf, Hambleton, & Sireci, 1999; Sireci & Berberoglu, 2000).

In contrasting the PCA/FA and MDS results, there was some similarity in that minor secondary factors appeared to be present in the PCA and FA results. However, the multidimensionality was much more conspicuous in the MDS solutions. This finding is not surprising given that MDS portrays the items in multidimensional space whereas PCA and FA portray the examinees in factor space (Davison, 1985; Davison & Skay, 1991). Thus, MDS is item-focused while PCA and FA are person-focused. It is also possible that the presence of multidimensionality due to structural differences across groups may be obscured when the data are aggregated across groups of examinees. If an assessment is essentially unidimensional, but the dimensionality is not consistent across groups, multidimensionality exists *across*, but not within, groups (Zumbo et al., 2003). Given that the TIMSS item pool contained 146 items, the primary dimension probably (appropriately) dominates the secondary factors. This dominance would support comparing groups on overall performance. However, the minor dimensions may prove illuminating for discovering finer distinctions across groups with respect to the interaction of specific content and country-specific instructional practices.

In summary, this study illustrated a useful method for evaluating the consistency of test structure across different language versions of a test. The effectiveness of the method is encouraging; however, discovery of such differences is only a first step toward better understanding of educational differences and achievement differences across languages and cultures, which is at the heart of international comparisons of educational achievement.

References

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36,* 185-198.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Bechger, T. M., van den Wittenboer, G., Hox, J. J., & De Glopper, C. D. (1999). The validity of comparative educational studies. *Educational Measurement: Issues and Practices, 18(3)*, 18-26.

Campbell, J. R., Kelly, D. L., Mullis, I.V.S., Martin. M. O., & Sainsbury, M. (2001, March). *International Association for the Evaluation of Educational Achievement: Progress in International Reading Literacy Study.* Chestnut Hill, MA: PIRLS International Study Center, Lynch School of Education, Boston College.

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

Davison, M.L., (1985). Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin*, 97, 94-105.

Davison, M.L., & Skay, C.L. (1991). Multidimensional scaling and factor models of test and item responses. *Psychological Bulletin*, 110, 551-556.

De Ayala, R. J., & Hertzog, M. A. (1991). The assessment of dimensionality in item response theory. *Multivariate Behavioral Research*, 26,765-792.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304-312.

Gonzalez, E.J., & Miles, J.A. (2001). TIMSS 1999 User Guide for the Benchmarking Database, Chestnut Hill, MA: Boston College

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. European Journal of Psychological Assessment, 17, 164-172.

MacCallum, R. (1981). Evaluating goodness of fit in nonmetric multidimensional scaling by ALSCAL. *Applied Psychological Measurement, 5*, 377-382.

Meara, K. P., Robin, F., & Sireci, S. G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral Research, 35 (2),* 229-259.

Martin, M. O., Mullis, I.V. S., Gonzalez, E., O'Connor, K. M., Chrostowski, S. J., Gregory, K. D., Smith, T. A., & Garden, R. A. (2000). *Science benchmarking report: TIMSS 1999—eighth grade.* Chestnut Hill, MA: TIMSS International Study Center, Lynch School of Education, Boston College.

O'Connor, K., & Malak, B. (2000) Translation and Cultural Adaptation of the TIMSS Instrument, in M.O. Martin, K.D. Gregory, K.D., S.E. Stemler, (Eds.) *TIMSS 1999 Technical Report*. Chestnut Hill, MA: Boston College.

Organization for Economic Co-operation and Development (2000). *OECD Program for International Student Assessment: National Project Manager's Manual.* Available at http://www.oecd.org//els/PISA/Docs/Downlaod/npmmanual110200.doc.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16(1)*, 12-19, 29.

Sireci, S. G. (in press). Evaluating cross-lingual test comparability using bilingual research designs. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Erlbaum.

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

Sireci, S.G. & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 35 (2)*, 229-259.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *Eurpoean Journal of Psychological Assessment, 13*, 29-37.

van de Vijver, F. & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*, 263-279.

Yamamoto, K., & Kulick, E. (2000) Scaling methodology and procedures for the TIMSS mathematics and science scales, in M.O. Martin, K.D. Gregory, K.D., S.E. Stemler, (Eds.) *TIMSS 1999 Technical Report.* Chestnut Hill, MA: Boston College.

Young, F.W., & Harris, D.F. (1993). Multidimensional scaling. In M.J. Noursis (Ed.). *SPSS for windows: Professional statistics* (computer manual, version 6.0) (pp. 155-222). Chicago, IL: SPSS, Inc.

Zumbo, B. D., Sireci, S. G., & Hambleton, R. K. (2003, April). Revisiting exploratory methods for construct comparability and measurement invariance: Is there something to be gained from the ways of old? Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Footnotes

[2] This is standard 9.9 in the current (1999) version of the *Standards* and standard 13.6 in the previous (1985) version.

[3] However, it should be noted that all countries participating in TIMSS are allowed to review items with respect to their curricular relevance and identify those that are typically not taught at the relevant grade level. Reanalysis of TIMSS data after excluding such "irrelevant" items typically does not change the rank orderings of countries, which also suggests construct comparability across countries.

[4] Tests in Hong Kong were administered in both languages, Chinese and English side by side. Each student chose to answer the test in the language of his or her preference.

[5] The inter-item distances were standardized to range from 0 to 1 to account for differences in sample size across item pairings.

Table 1

Rank-Order of 13-year Old Students in Science in TIMSS 1999

Country	Average Score	Standard Error
Chinese Taipei	569	4.4
Singapore	568	8.0
Hungary	552	3.7
Japan	550	2.2
Korea, Rep. of	549	2.6
Netherlands	545	6.9
Australia	540	4.4
Czech Republic	539	4.2
England	538	4.8
Finland	535	3.5
Slovak Republic	535	3.3
Belgium (Flemish)	535	3.1
Slovenia	533	3.2
Canada	533	2.1
Hong Kong, SAR	530	3.7
Russian Federation	529	6.4
Bulgaria	518	5.4
United States	515	4.6
New Zealand	510	4.9
Latvia (LSS)	503	4.8
Italy	493	3.9
Malaysia	492	4.4
Lithuania	488	4.1
Thailand	482	4.0
Romania	472	5.8
Israel	468	4.9
Cyprus	460	2.4
Moldova	459	4.0
Macedonia, Rep. of	458	5.2
Jordan	450	3.8
Iran, Islamic Rep.	448	3.8
Indonesia	435	4.5
Turkey	433	4.3
Tunisia	430	3.4
Chile	420	3.7
Philippines	345	7.5
Morocco	323	4.3
South Africa	243	7.8

Source: TIMSS 1999 International Science Report (Martin et al., 2000), p. 32.

Table 2

Content and Performance Expectation Dimensions of the TIMSS 1999 Science Assessment

Content	Performance Expectation
Earth Science	Understanding
Life Sciences	Theorizing, Analyzing, and Solving Problems
Physical Science	Using Tools, Routine Procedures and Science Processes
History of Science and Technology	Investigating the Natural World
Environmental and Resource Issues	Communicating
Nature of Science	
Science and Other Disciplines	

Source: Gonzalez & Miles (2001)

Table 3

Number of Test Items and Score Points By Item Type and Science Reporting Category

Reporting Category	Item Type				
	Multiple-Choice	Short-Answer	Extended-Response	Number of Items	Score Points
Earth Science	17	4	1	22	23
Life Science	28	7	5	40	42
Physics	28	11	-	39	39
Chemistry	15	2	3	20	22
Environmental and Resource Issues	7	2	4	13	14
Scientific Inquiry and the Nature of Science	9	2	1	12	13
Total	104	28	14	146	153

Source: Gonzalez & Miles (2001)

Table 4

Summary of PCA Results for Eight Major Content Clusters

Component/Factor	λ	% VAF	Cum % VAF
1	5.58	12.122	12.122
2	1.40	3.050	15.172
3	1.30	2.821	17.993
4	1.20	2.605	20.597
5	1.16	2.511	23.108
6	1.14	2.486	25.594
7	1.12	2.434	28.028
8	1.07	2.334	30.362
9	1.05	2.272	32.634
10	1.02	2.226	34.860
11	1.00	2.189	37.049

Table 5

Fit Indices for WMDS Solution for Eight Major Content Clusters

Dimensional Solution	STRESS	Improvement	Expected Stress[a]	Expected - Observed	R^2
1[b]	.471				.529
2	.351	.120	.482	.131	.577
3	.287	.064	.377	.090	.630
4	.242	.045	.316	.074	.677
5	.221	.021	.280	.059	.730
6	.195	.026			.777

[a] Based on MacCallum's (1981) formula 5. Estimates are only available for 2D through 5D solutions.

[b] A one-dimensional solution is not a weighted solution. It is fit by via replicated MDS (Young & Harris, 1993), which allows for separate transformations of each data matrix.

Table 6

Group Weights and Item Difficulty/Coordinate Correlations for 2D Solution

Group	Dimension			
	1		2	
	W	$r_{b,c}$	W	$r_{b,c}$
United States	.88	.91	.17	.40
Canadian English	.77	.87	.40	.53
England	.71	.87	.39	.49
Italy	.67	.79	.29	.52
Russia	.58	.69	.26	.44
Canadian French	.56	.78	.48	.54
Belgium	.49	.68	.54	.67
Hong Kong	.49	.72	.42	.47
Korea	.47	.66	.52	.53
Japan	.27	.47	.75	.78
% VAF	.37		.20	

W=weight for group on dimension.

$r_{b,c}$=correlation between IRT difficulty estimates and MDS coordinates.

VAF= % of variance in item dissimilarities accounted for by coordinates.

Table 7 WMDS Weight Matrix for 5D Solution

Group	Dimension									
	1		2		3		4		5	
	W	$r_{b,c}$	W	$r_{b,c}$	W	$r_{b,c}$	W	$r_{b,c}$	W	$r_{b,c}$
United States	.62	-.77	.41	-.66	.23	-.42	.27	.58	.16	.50
England	.61	-.74	.47	-.70	.23	-.45	.15	.55	.24	.55
Canadian English	.55	-.71	.41	-.66	.34	-.50	.27	.61	.30	.58
Italy	.56	-.69	.20	-.50	.43	-.60	.29	.66	.09	.50
Canadian French	.38	-.61	.33	-.61	.40	-.50	.29	.61	.30	.56
Korea	.32	-.48	.40	-.63	.41	-.41	.27	.56	.23	.53
Belgium	.23	-.48	.29	-.58	.68	-.72	.27	.57	.23	.62
Russia	.16	-.42	.05	-.44	.12	-.46	.89	.89	.09	.47
Hong Kong	.11	-.44	.93	-.89	.05	-.33	.10	.41	.10	.50
Japan	.07	-.34	.12	-.48	.11	-.44	.08	.36	.91	.86
% VAF	.17		.18		.12		.13		.12	

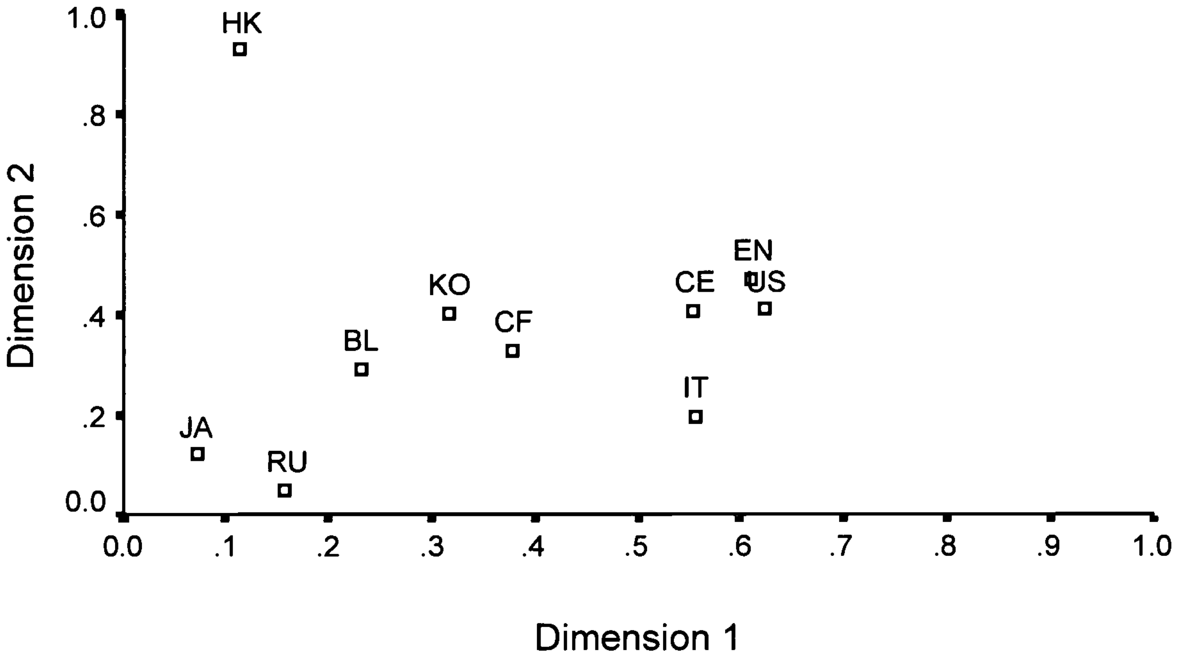
W=weight for group on dimension.

$r_{b,c}$=correlation between IRT difficulty estimates and MDS coordinates.

VAF= % of variance in item dissimilarities accounted for by coordinates.

Figure 1

2D Weight Space From 5D WMDS



BL=Belgium, CE=Can.Eng., CF=Can. Fr., EN=England, HK=Hong Kong,

IT=Italy, JA=Japan, KO=Korea, RU=Russia, US=USA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035268

I. DOCUMENT IDENTIFICATION:

Title:	Evaluating the Structural Equivalence of Tests Used in International Comparisons of Educational Achievement
Author(s):	Stephen G. Sireci and Eugenio J. Gonzalez
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☑	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please →

Signature:	Printed Name/Position/Title: Stephen G. Sireci	
Organization/Address:	Telephone:	FAX:
	E-Mail Address:	Date:

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org