DOCUMENT RESUME

ED 468 489 TM 034 399

AUTHOR Sireci, Stephen G.; Khaliq, Shameem Nyla

TITLE An Analysis of the Psychometric Properties of Dual Language Test Forms.

INSTITUTION Massachusetts Univ., Amherst. School of Education.
REPORT NO RR-458
PUB DATE 2002-04-00
NOTE 44p.; A version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS *Item Bias; *Language Proficiency; Limited English Speaking; Multidimensional Scaling; Psychometrics; Spanish; *Structural Equation Models; Test Construction; *Test Format; Translation
IDENTIFIERS *Dual Language Text; SIBTEST (Computer Program)

ABSTRACT

Many students in the United States who are required to take educational tests are not fully proficient in English. To address this problem, a state-mandated testing program created dual language English-Spanish versions of some of their tests. In this study, the psychometric properties of the English and dual language versions of a fourth-grade mathematics test were explored. Analyses of the consistency of test structure across the two forms were conducted using structural equation modeling and multidimensional scaling. Analyses of differential item functioning (DIF) were conducted using Poly-SIBTEST. The results suggest slight structural differences across the two versions of the test. Part of this difference was attributed to overall proficiency differences across the two studied groups, and part was attributed to DIF. The implications of the findings for future research in this area are discussed. (Contains 4 figures, 11 tables, and 28 references.) (Author/SLD)

An Analysis of the Psychometric Properties of Dual Language Test Forms[1]

Stephen G. Sireci and Shameem Nyla Khaliq

University of Massachusetts Amherst

[1] Center for Educational Assessment Research Report No 458. Amherst, MA: School of Education, University of Massachusetts Amherst. An earlier version of this paper was presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April 2, 2002.

Abstract

Many students in the United States who are required to take educational tests are not fully proficient in English. To address this problem, a state-mandated testing program created dual language English-Spanish versions of some of their tests. In this paper we explore the psychometric properties of the English and dual language versions of a 4th-grade math test. Analyses of the consistency of test structure across the two forms were conducted using structural equation modeling and multidimensional scaling. Analyses of differential item functioning (DIF) were conducted using Poly-SIBTEST. The results suggest slight structural differences across the two versions of the test. Part of this difference was attributed to overall proficiency differences across the two studied groups and part was attributed to DIF. The implications of the findings for future research in this area are discussed.

Comparing the Psychometric Properties of Monolingual and Dual Language Test Forms

Mandated assessments are a common component in contemporary educational reform movements. In these movements, standardized tests are used to ensure uniform content, test administration, and scoring procedures. However, when linguistic diversity is present in an educational system, a test administered in the dominant language of the system may not provide a level playing field for students who are not fully proficient in that language. For this reason, different language versions of a test are sometimes offered. In other cases, dual language test booklets, where two different language versions of each test item appear side-by-side in the same booklet, are offered. Both strategies aim toward reducing threats to the validity of test score interpretations due to limited proficiency in a dominant language.

Research has shown that the use of different language versions of a test may not provide test scores that are comparable across languages (e.g., Gierl & Khaliq, 2001; Hambleton, 1994, 2001; Sireci, 1997; van der Vijver & Tanzer, 1998). Therefore, the use of dual language test booklets is intriguing for balancing the measurement goals of equity and score comparability. Dual language test booklets may enhance assessment equity by reducing construct-irrelevant variance in students' test scores due to proficiency in a specific language. The use of dual language booklets also explicitly respects the linguistic diversity within a school system and empowers students to choose the language medium in which they think they can best demonstrate their knowledge and skills. However, the hypothesis that dual language test booklets have similar psychometric properties to the original, monolingual version of the test is rarely empirically studied. The degree to which dual language test forms are comparable to their monolingual counterparts is an important validity issue, since scores from these different tests are typically interpreted as if they are equivalent.

Very little research has been conducted on the benefits and limitations of dual language test booklets. Garcia et al. (2000) randomly assigned Spanish-speaking limited English proficient (LEP) students to an experimental dual language or to an English language 8th grade math test from the National Assessment of Educational Progress to evaluate the "psychometric equivalence" of the dual language and monolingual versions of the test (p. 6). They found that Spanish-speaking LEP students appreciated the dual language booklets, with 85% of the students finding the booklets "useful" or "very useful." However, they also found that students who had three years or less instruction in English predominantly read and responded only to the Spanish versions of the test items. They also found that LEP students with high levels of English proficiency scored slightly lower when taking the dual language version of the test. These results suggest that the use of dual language booklets is promising, but more research is needed to determine its utility.

As Garcia et al. (2000) noted, a limitation of their study was that they were unable to statistically evaluate the psychometric comparability of the dual language and English booklets due to the small numbers of students who took the dual language version. A central focus of our analysis is comparing the psychometric properties of the original and dual language test booklets. Specifically, we compared the structure underlying the data from both booklets as well as the functioning of the items.

The *Guidelines for Adapting Educational and Psychological Tests* developed by the International Test Commission (Hambleton, 1994, 2001) and the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) underscore the need for statistical procedures to evaluate test comparability across cultures and languages. For example

the *Standards* state "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)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 factor structure 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 der Vijver & Tanzer, 1998).

Multidimensional scaling and structural equation modeling can be used to compare the structure of an assessment across varying language and cultural groups. Recently, these methods have been used to evaluate the structural equivalence of different language versions of a test (Gierl, 1999; Reise, Widaman, & Pugh, 1993; Robin, Sireci, & Hambleton, 2000; Sireci & Allalouf, in press). Gierl used principal components analysis and structural equation modeling, Reise et al. used structural equation modeling and item response theory, and Robin et al. and Sireci and Allalouf used multidimensional scaling to evaluate equivalence across language groups. These and other studies support the contention that several statistical techniques are useful for evaluating the psychometric comparability of tests across different language and cultural groups.

The purpose of the present study is to compare data from English and English-Spanish versions of a statewide mathematics test and to evaluate the comparability of scores across the two versions. Specifically, we evaluate and compare the dimensional structure of the test and evaluate the items for differential functioning across the English and English-Spanish versions.

Method

The data analyzed here come from a state-mandated testing program in the United States that tests students in several grades and subject areas. All exams associated with this testing program are developed in English. However, given the large numbers of Spanish-speaking LEP students, dual language English-Spanish test booklets were created for the History/Social Sciences, Mathematics, and Science and Technology tests. Approximately 60% of the 12,000 LEP students in this state are Spanish-language dominant. To be eligible to take the English-Spanish version of a test, LEP students must have been enrolled in U.S. schools for less than three years, be slated for Spanish-language instruction the following year, and read and write at or near grade level in Spanish. Only those students who were not eligible to sit for the English or English-Spanish administrations (e.g., Polish-speaking LEP students in U.S. schools for less than three years) were not required to sit for the test. However, they are permitted to do so at their discretion.

Description of the Mathematics Test

The mathematics test consisted of 39 items. Twenty-nine of the items were multiple-choice, 5 were short constructed-response items that were scored dichotomously, and 5 were extended constructed-response items, four of which were scored 0-4 and the remaining item scored 0-2. There were four mathematics content areas covered on the test: (a) number sense, (b) patterns, relations, and functions, (c) geometry and measurement, and (d) statistics and

probability. The dual language version of the test contained the English version of the test questions on the right-facing pages with the Spanish translation of the same test questions on the left-facing page. Any extra testing material was provided in both languages and test administrators were provided with test instructions in both languages.

Examinees

At Grade 4, there were 76,783 examinees who completed the English form of the mathematics test and 585 students who completed the dual language form of the test. When groups of such very different sizes are compared statistically, this difference can conceal meaningful group differences or exaggerate trivial differences. For this reason, three independent randomly selected groups of 585 examinees were extracted from the total English group (called English-1, English-2, and English-3). These groups served as comparison groups for evaluating differences among random groups due only to the fluctuations expected from sampling small groups of students. Three additional "matched" English groups were also selected from the original English sample. These three matched English groups were selected such that the distribution of total test score was made identical to the total test score distribution of the dual language sample (i.e., if there were 25 students with a score of 15 on the Dual language form, then 25 students with a score of 15 were independently and randomly selected for one of the three English groups). These three matched English groups are called English-M1, English-M2, and English-M3.

Analyses

Four different statistical procedures were conducted for this study: principal components analysis (PCA), multidimensional scaling (MDS), structural equations modeling, and DIF analyses. These analyses were conducted twice. The first set of analyses compared the dual language group to the three random English samples and the second set of analyses compared the dual language group to the three matched English samples.

Principal components analysis

PCA was used as a preliminary step in determining the dimensionality of the data from the English and dual language test booklets. We were primarily interested in discovering the number of dimensions (components) that were required to account for the substantive variation in the data. The number of dimensions underlying the data for each group was determined using the Kaiser-Guttman rule for the lower bound, and inspecting the scree plots. The Kaiser-Guttman rule suggests that the number of factors is equal to the number of components with an eigenvalue equal to or greater than one, when using the full-correlation matrix (Gorsuch, 1983). The scree plots are obtained by plotting the eigenvalues against the number of factors associated with each solution. If an "elbow" appears in the plot, it indicates a tapering off of fit to the data (i.e., over-factoring) and the solution that precedes the elbow reflects the number of factors needed to represent the data. For the PCAs, polychoric correlations were computed among the test items separately for each group of examinees. All polychorics were computed using PRELIS 2.20 (Jöreskog & Sörbom, 1998).

Multidimensional scaling

Three sets of MDS analyses were conducted. The first set of analyses focused on the total English booklet group. Classical MDS analyses (involving a single matrix of dissimilarity

data) were conducted fitting one- through six-dimensional solutions to the data. The second set of analyses fit the weighted MDS (WMDS) model to the data from the three random English samples and the dual language booklet sample. The third set of analyses fit the WMDS model to the data from the three matched English and dual language groups.

A limitation of PCA is that different samples of examinees cannot be simultaneously analyzed and compared within a single analysis. Weighted multidimensional (WMDS) retains the exploratory nature of PCA, but allows for simultaneous comparison of dimensionality across groups. WMDS analyzes several matrices of dissimilarity data to derive both a common structure that best represents the data for all groups considered together as well as individual group weights for adjusting this common structure to arrive at separate structures that best fit the data for each group. The weights for each group can be used to compare the relevance of the dimensional structure across groups.

For the MDS analyses, squared Euclidean distances were computed among the items separately for each group of examinees. This process provided a symmetric inter-item distance matrix for each group. Scores on the polytomous items were divided by the maximum possible score on the item to place all items within a 0-1 metric.

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} represent 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 10.0 (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 dimensionality of the data structure was not known; therefore, for the classical (i.e., one-matrix) MDS, the data were fit to one- through six-dimensional models. The maximum number of dimensions fitted by the SPSS version of ALSCAL is six. A one-dimensional model

is not relevant to WMDS since it implies a single, consistent dimension for all groups. Thus, two- through six-dimensional solutions were conducted for the WMDS analyses. The most appropriate dimensional solution was determined using the criteria of data-model fit and interpretability of the solution. 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.

Structural equations modeling

The third procedure applied to the investigation of test structure across groups was structural equations modeling (SEM). SEM allows for analysis of test structure across multiple groups, but it requires a priori specification of the hypothesized structure. Since the specifications for the test posited a unidimensional construct, a one-factor model was specified. Three different models were tested. The first model tested whether a single factor existed across the groups. (If this model did not display adequate fit, multidimensional models based on the PCA and MDS results were to be fit.) The second model tested whether the factor loadings for the items from each group were invariant. The third model constrained the errors associated with the factor loadings to be invariant across groups.

Four fit indices were used to assess each model: the chi-square test, root mean square error of approximation (RMSEA), root mean square residual (RMR) and goodness-of-fit index (GFI). A non-significant chi-square test indicates adequate data-model fit, but the chi-square test is sensitive to sample size and so it should not be used as the only test to determine data-model fit (Hayduk, 1987). Conventions for using the other indices to evaluate a particular data-model fit have been offered (e.g., Browne & Cudek, 1993; Byrne, 1998; MacCallum & Browne, 1993, Reise, et al., 1993; Sireci et al., 1998), and although there has been much debate about the utility of these indices, non-significant chi-squares, RMSEA and RMR values below .05, and GFI/AGFI values of .90 or above, are generally considered to indicate reasonable data-model fit. All confirmatory analyses were conducted with LISREL 8.20 (Jöreskog & Sörbom, 1998) using maximum likelihood estimation and the asymptotic item covariance matrix.

DIF analyses

In addition to the extensive structural analyses, the data were also evaluated for differences across the test forms at the item level using DIF analyses. DIF occurs when two groups of examinees have a different probability of answering an item correctly, after controlling for overall proficiency. The Simultaneous Item Bias Test (SIBTEST) is a popular method for evaluating DIF and the Poly-SIBTEST computer program (Chang, Mazzeo, & Roussos, 1996) can be used for both dichotomous and polytomous items. In Poly-SIBTEST, examinees are typically matched on the total test score, to control for proficiency differences across groups. We chose Poly-SIBTEST for the DIF analyses in this study because it is appropriate for the mix of dichotomous and polytomous items on the test and it has been shown to successfully identify DIF items when they are present (Chang et al., 1996).

Using Poly-SIBTEST, at each test score level the probability of obtaining a specific response is determined for the reference and focal groups separately and then compared. The examinees in the reference and focal groups were matched on total test score. The test statistic used by Poly-SIBTEST is:

$$\hat{B} = \frac{\hat{B}_U}{\hat{\sigma}(\hat{B}_U)} \qquad [3]$$

where $\hat{B}_U = \sum_{k=0}^{n} \hat{P}_k(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ and

$$\hat{\sigma}(\hat{B}_U) = (\sum_{k=0}^{n} p_k^2 (\frac{1}{J_{Rk}} \hat{\sigma}^2 (Y \mid k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2 (Y \mid k, F)))^{\frac{1}{2}}$$ (Shealy & Stout, 1993).

In these formulas, $\hat{P}_k$ is the proportion of examinees from the Focal group obtaining a score of k on the valid subtest, $\hat{\sigma}^2 (Y \mid k, F)$ and $\hat{\sigma}^2 (Y \mid k, R)$ are the variances of the Focal and Reference groups for those examinees with a score of k, J_{Fk} and J_{Rk} are the number of examinees in the Focal and Reference groups with a score of k.

SIBTEST tests the null hypothesis that $B = 0$. However, given large sample sizes, the magnitude of DIF is more important than its statistical significance; therefore, the $\hat{B}$ (Beta-uni) statistic was used as an effect size for gauging the magnitude of DIF. In this study, we used this effect size to classify items into one of three categories (no DIF, moderate DIF, or large DIF). Items with Beta-unis between 0 and .059 were classified as not displaying DIF, items with Beta-unis between .059 and .088 were classified as exhibiting moderate DIF, and items with Beta-unis greater than .088 were classified as large DIF. These classifications are based on guidelines

provided by Roussos and Stout (1996), and are consistent with the DIF effect size classifications used by Educational Testing Service (Dorans & Holland, 1993).

If DIF were present in a test, the use of total test score as a matching criterion may be problematic, since it is affected by the presence of items that function differentially. Therefore, a two-stage procedure was conducted in this study, in which we replicated the DIF analyses after removing items that displayed large DIF from the total test score matching variable.

Results

Overall Descriptive Statistics

Total test score and item-level descriptive statistics, including means, standard deviations, coefficient alphas, and standard errors of measurement for the full English, dual language, and reduced size random English groups are presented in Table 1. The dual language test booklet group scored about 1.5 standard deviations lower than the English groups and exhibited much less variability. The test score distributions for these four groups are presented in Figure 1. Very few students who took the dual language version of the test scored above the mean of the students who took the English-only version, which explains the smaller mean and variability of the dual language group. The coefficient alpha and standard error of measurement were also lower for the dual language booklet group. The range of test scores, item difficulties, and item discriminations were also lower for this group. These differences hold up across both the multiple-choice and polytomous items. It is interesting to note that one of the items on the dual language form exhibited a negative corrected item-total correlation with the total test score (item 32), none of the items on the English form exhibited a negative item-total correlation.

[Insert Table 1 Here]

[Insert Figure 1 Here]

Descriptive statistics for matched samples

Given the vast differences between the English-only and the dual language booklet samples, it was important to replicate all analyses using English samples that better matched the proficiency distribution of the dual language booklet group. These "matched analyses" attempted to disentangle differences due to overall group proficiency from those due to psychometric qualities of the tests. The descriptive statistics for the three matched English groups are presented alongside the descriptive statistics for the dual language group in Table 2. The mean and standard deviations of the multiple-choice and polytomous items are fairly similar across the English and dual language samples and the differences in coefficient alpha and conditional standard error of measurement essentially disappear.

[Insert Table 2 Here]

Principal Components Results

For the dual language test form, the PCA results reported 14 eigenvalues greater than one. The first eigenvalue was 5.1, which accounted for 13.2% of the total variance. This eigenvalue was three times larger than the second eigenvalue (1.6 for 4% of the total variance). For English samples, ten eigenvalues were greater than one in each analysis. For each sample, the first eigenvalue was around 8.0, accounted for over 20% of the total variance, and was about 6 times larger than the second eigenvalue. The scree plots for all four groups are presented in Figure 2. A large and consistent dominant dimension is evident for the English booklet samples. The dominance of the first dimension for the dual language booklet is less pronounced and the secondary dimensions appear larger, relative to those from the English samples. Thus, these PCA results indicate that a dominant dimension underlies the data for both test versions, but that it is much more salient for the English version.

[Insert Figure 2 Here]

PCA for matched samples

After matching English and dual language booklet examinees, the PCA results for the matched samples mimicked the results for the dual language sample. For each sample, the first 14 components had eigenvalues greater than 1.0, with the first eigenvalue for each sample being around 5.1 and accounting for approximately 13% of the total variance. In general, the first eigenvalue was about three times larger than the second eigenvalue, which accounted for approximately 3.5% of the total variance. The scree plots for the dual language and matched English groups are presented in Figure 3. The eigenvalue plots are nearly coincident, which suggests that the relatively weak first factor and slight multidimensionality noticed for the dual language sample is more likely due to an interaction between proficiency and dimensionality (e.g., lower-achieving students using a different strategy to solve difficult items) than due to item translation problems or other differences between the test forms.

MDS Results

CMDS on total English group

The fit statistics for the MDS analysis on the total English booklet group suggested that the data were multidimensional. The STRESS and R^2 fit indices for the one- through six-dimensional solutions are presented in Table 3. The STRESS was large for the one-dimensional solution and dropped substantially for the two- and three-dimensional solutions. These fit statistics suggest that at least two dimensions are needed to account for the similarities among the items. This finding is different from the PCA, but MDS is known to pick up on smaller dimensions that may not be identified by PCA or factor analysis (Davison, 1985; Davison & Skay, 1991).

[Insert Table 3 Here]

To interpret these MDS solutions and select the one that best represented the structure of the test, statistical and qualitative attributes about the items were compared with the solutions by correlating these attributes (or dummy coded values of the attributes) with the MDS item coordinates. For each item, the following attributes were available: item format (multiple-choice or constructed response), mathematics content area, proportion correct difficulty index, and standard deviation. Only correlations that were statistically significant at $p < .01$ were considered meaningful (note that the sample size for all correlations was 39, which is the number of items).

The two- and three-dimensional solutions were the only ones in which all dimensions exhibited statistically significant correlations with at least one item attribute. Both solutions contained dimensions that correlated with item difficulty and one content area. Since these attributes were captured in the two-dimensional solution, it was taken as the best representation of the structure of the exam. The first dimension correlated .91 with the proportion correct difficulty index, which suggests this dimension accounted for the general difficulty of the items. The second dimension correlated moderately with the "geometry and measurement" content area of the test (r=.45). The two-dimensional solution accounted for 74% of the variation of the (transformed) item dissimilarities.

WMDS on random English and dual language groups

The fit statistics for the WMDS analysis that included data from the random English and dual language booklet groups are presented in Table 4. In general, an additional dimension was needed to achieve the level of fit observed for the analysis of the total English group, with the three-dimensional solution exhibiting fit to the data similar to that observed for the two-

dimensional classical MDS solution. Using MacCallum's (1981) formulae for evaluating STRESS in ALSCAL, the STRESS value associated with the three-dimensional solution (.24) was below that expected from random data (.33). This finding is consistent with a good fit in three dimensions for data containing a moderate degree of measurement error.

[Insert Table 4 Here]

The correlations among the item attributes and the three-dimensional WMDS item coordinates exhibited statistically significant correlations with the proportion correct item difficulty indices (p-values) for both the English and dual language groups; however, these group-specific p-values had different patterns of correlations with the dimensions. The p-values for the English group correlated .89 with the item coordinates on the first dimension, while the p-values for the dual language booklet group correlated .73 with this dimension. Interestingly, the dual language group p-values also correlated .75 with the item coordinates on the *second* dimension, while the English group p-values correlated .55 with the coordinates on this dimension. These findings suggest that the general difficulty dimension derived from analysis of the total English sample does not perfectly generalize to the dual language booklet group. The third dimension did not correlate significantly with any of the known item attributes. None of the higher-dimensional solutions exhibited significant correlations with item attributes beyond those observed in the three-dimensional solution, and no content area designations correlated with the MDS coordinates.

The WMDS group weights are presented in Table 5. The weights for the three random English samples were similar across all three dimensions. However, the dual language booklet group had a noticeably smaller weight on dimension 1 and a noticeably larger weight on dimension 2. These results are consistent with the correlational analyses that suggested

dimension 1 was most relevant to the English sample p-values and dimension 2 was most relevant to the dual language booklet p-values. Therefore, the WMDS results using the random English groups do not support the conclusion of structural equivalence across the English and dual language versions of the test.

[Insert Table 5 Here]

WMDS on dual language and matched samples

The fit statistics for the WMDS based on matched English and dual language groups are presented in Table 6. These fit statistics lie between those from the classical MDS conducted on the total English sample and the WMDS results based on the random English and dual language groups. The improvement in fit tapered off after the three-dimensional solution, which accounted for 76% of the variation in the item dissimilarity data. When the MDS coordinates from the three-dimensional solution were correlated with the item attributes, only the item difficulty statistics for the English and dual language samples exhibited significant correlations. For the two-dimensional solution, the p-values for both groups correlated strongly with the coordinates on the first dimension (r=.74 for both groups) and did not correlate with the coordinates on the second dimension.

[Insert Table 6 Here]

For the three-dimensional solution, the p-value/coordinate correlations for both groups were lower on the first dimension (r=.70 for the English group and .64 for the dual language group), but significant correlations were also observed for both sets of p-values on the second dimension (r=.61 and .72 for the English and dual language groups, respectively). The dual language group p-values also exhibited a slight correlation with the item coordinates on the third dimension (r=.40), whereas the p-values for the English group exhibited a lower correlation with

this dimension (r=.30). These results suggest that the dimensionality of the test data across the dual language and matched English groups is more similar than before matching, but that some subtle differences still exist across the English and dual language test booklets.

The finding of reduced, but still noticeable structural differences across test versions was also borne out by the WMDS group weights, which are presented in Table 7. For both the two- and three-dimensional solutions, a slightly different pattern of weights was observed for the dual language booklet group, relative to the matched English groups. In the two-dimensional solution, all groups had the highest weight on the first dimension. In the three-dimensional solution, the largest weight for the matched English samples was on the first dimension and the largest weight for the dual language sample was on the second dimension. These findings indicate that the increased dimensionality is tuning into subtle structural differences between the English and dual language versions of the exam.

[Insert Table 7 Here]

Structural Equation Modeling Results

Unidimensional SEM models were fit separately to the data for each group (random English, matched English, and dual language) to ascertain whether the unidimensional model was viable for all groups. The fit statistics for these models are presented in Table 8. Although the chi-square statistics were statistically significant, the other fit indices suggested reasonable data-model fit. Therefore, simultaneous fit of this unidimensional model across all groups was investigated. First, we fit the unidimensional models to the data for the random English and dual language groups. Then, we fit the models to the data from the matched English and dual language groups.

[Insert Table 8 Here]

For the random English/dual language SEMs, the model that posited a single dimension for each group exhibited reasonable fit to the data, but when the factor loadings were constrained to be equal across groups, the GFI index dropped below .90. When the errors associated with the factor loadings were constrained to be equal across groups, both the RMR and GFI fell below the levels where adequate data-model fit could be concluded. These findings, which are summarized in Table 9, suggest that the factor loadings and the errors associated with these loadings may not be equivalent across groups.

[Insert Table 9 Here]

For the matched English/dual language SEMs, the lack of structural equivalence seemed to disappear, suggesting that the relatively poor fit noted in the previous models was due to group proficiency differences, rather than due to differences in test characteristics—a finding similar to the PCA results. In the matched analyses, which are summarized in Table 10, the CFA models with equivalent factor loadings and equivalent errors associated with those loadings, all exhibited reasonable fit to the data.

[Insert Table 10 Here]

DIF Results

In conducting the DIF analyses, the dual language group was compared to the three random English groups as well as to the three matched English groups. Using the Poly-SIBTEST Beta-uni effect size criteria, items were classified as "no DIF," "moderate DIF," and "large DIF." The three random and matched samples were used to evaluate the reliability of the DIF results. Only those items that were flagged for moderate or large DIF in all three comparisons were considered to truly function differentially across groups.

For the dual language/random English group comparisons, most of the items (27 of 39) were classified as no DIF, one item was classified as moderate DIF, and eleven were classified as large DIF. For the matched English/dual language DIF analyses, five items were classified as moderate DIF, but the number of large DIF items dropped to seven. The direction of DIF was not consistently against the English or dual language group. For example, of the 12 moderate or large DIF items identified in the matched analyses, six favored the English group and six favored the dual language group.

To "purify" the DIF matching criterion, we excluded the seven items that exhibited large DIF in the matched analysis, and one item that exhibited large DIF in two of the three matched replications (and had a borderline-large effect size of .8 in the third replication), from the total test score and reran all the Poly-SIBTEST analyses. The items that were flagged for large DIF in both the random and matched analyses using all 39 items as the matching criterion were once again flagged as large DIF using the purified 31-item criterion. Items that were identified as moderate DIF when using all 39 items in the matching criterion tended not to be classified as such in all three replications, while some other items that were previously classified as no DIF, were classified as moderate DIF in one or more replication. These results suggest that the 39-item matching criterion was appropriate (which is not surprising given that the direction of DIF varied) and that the large DIF classifications were reliable. The moderate DIF classifications do not appear to be reliable and so it seems appropriate to focus on large DIF items when looking for translation problems or other reasons why some items functioned differentially across the English and dual language test versions.

As a first step in interpreting the DIF results, we examined whether the eight large DIF items related to specific content areas or item format. A summary of this analysis is presented in

Table 11. The two DIF items from the number sense content area both favored the dual language group, while the three DIF items from the statistics and probability content area favored the English language group, which suggests that there may be some relationship between content area and DIF. The two short constructed-response items that were flagged for DIF both favored the dual language group. These results should be followed up by bilingual instructors and math curriculum specialists who are familiar with both groups of students.

[Insert Table 11 Here]

Comparing the Dimensionality and DIF Results

The DIF results indicated that some items functioned differentially, even after accounting for overall group differences in proficiency by selecting an English language comparison group that performed similarly to the dual language group. The only dimensionality analysis that provided evidence of structural differences was the weighted MDS analysis. To examine whether the presence of DIF affected the consistency of dimensionality across the two test versions, the eight large DIF items were removed from the data and the WMDS analysis was repeated using the matched English samples. The WMDS weight spaces from the three-dimensional solutions for the 39-item (DIF items included) and 31-item (DIF items removed) analyses are presented in Figures 4a and 4b, respectively. A comparison of the figures reveals that after the DIF items are removed, the weight vector for the dual language group is much more similar to those of the matched English groups. This finding suggests that DIF accounted for some of the structural differences noted between the dual language and English language groups.

[Insert Figure 4 Here]

Discussion

Several statistical methods were applied to the problem of evaluating the equivalence of English and English-Spanish versions of a statewide mathematics assessment. From a methodological perspective, several features of this study are important. First, multiple methods were used to evaluate structural equivalence, which provides greater information for evaluating equivalence than does the use of only a single method. Second, replications were used at each stage of analysis to evaluate the stability of the conclusions over random samples from the English language group. Third, matched samples of test takers from the majority group were used to disentangle group proficiency effects on structure and DIF from those due to characteristics of the items. Fourth, the structural and DIF analyses were done sequentially and interpreted together. Although these steps increased the complexity of the study, they allow for greater confidence in the conclusions drawn from the results.

With respect to structural equivalence, the results suggest that there are slight differences in the structure of the test across the two versions. This finding was supported by all structural analyses. However, the PCA and CFA results attributed this difference to overall group proficiency differences, rather than due to real structural differences that may be due to translation problems or other inconsistencies. This finding was evident when the analyses were repeated using matched samples and compared to the results using random samples. The WMDS results using matched samples also reduced the magnitude of structural differences across the groups, but some differences were still noted, as was evident in the group weights. It is possible that MDS is picking up at aberrancies at the item level better than PCA or CFA. The fact that the structure became more similar after removing the DIF items provides partial support for this hypothesis.

The DIF results identified several items that are functioning differentially across the two test forms, even when matched groups of students were used. Obviously, the items flagged for DIF should be inspected by bilingual math experts to determine why they function differentially across the two forms (Allalouf, Hambleton, & Sireci, 1999). Although substantive DIF was found, it was comforting to see that the degree of DIF was relatively balanced across the English and dual language groups.

The use of dual language test booklets for measuring the academic proficiencies of LEP students deserves further study. On an untimed test like the one presented here, this option may provide more benefits than limitations. However, Garcia et al. (2000) noted that dual language booklets and tests translated from English to Spanish, may introduce an unintended speed factor into the test administration.

The results of this research are obviously limited to the specific test form studied and so more research on dual language test forms is needed. In addition to appraising the psychometric properties of such tests, future research in this area should focus on students' impressions of dual language test booklets, and their experiences with them, to better gauge the practical benefits and limitations of this test administration option.

Although this research focused on the psychometric issues involved in the use of dual language test booklets, the policy issues associated with this practice, and other practices in testing LEP students, deserve much more study. Wherever possible, test development and administration practices that can improve measurement of LEP students' knowledge and skills, and that can result in improved instructional practices for these students, should be identified and encouraged.

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.

Browne, M. W. & Cudek, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.) Testing structural equation models (pp. 445-455). Newbury Park, CA: Sage.

Byrne, B. M. (1998). Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Hillsdale, NJ: Lawrence Erlbaum.

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.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. Journal of Educational Measurement, 33, 333-353.

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.

Dorans, N. J., & Holland. P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.) Differential item functioning (pp. 35-66). Hillsdale, New Jersey: Lawrence Erlbaum.

Garcia, T., del Rio Paraent, L., Chen, L., Ferrara, S., Garavaglia, D, Johnson, E., Liang, J., Oppler, S., Searcy, C., Shieh, Y., & Ye, Y. (2000, November). Study of a dual language test booklet in 8th grade mathematics: Final report. Washington, DC: American Institutes for Research.

Gierl, M. J. (1999). Construct equivalence on translated achievement tests. Unpublished manuscript, University of Alberta.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. Journal of Educational Measurement, 38, 164-187.

Gorsuch, R. L. (1983). Factor Analysis 2nd Edition. Hillsdale, NJ: Lawrence Erlbaum and Associates.

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.

Hayduk, L. A. (1987). Structural equation modeling with LISREL: Essentials and advances. Baltimore: John Hopkins University Press.

Joreskog, K., & Sorbom, D. (1998). LISREL 8.20 and PRELIS 2.20 for Windows. Software. Mooresville, IN: Scientific Software, International.

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

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. .Psychological Bulletin, 114, 533-541.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114, 552-566.

Robin, F., Sireci, G. S., & Hambleton, R. K. (2000). Evaluating the equivalence of different language versions of a credentialing exam. Laboratory of Psychometric and Evaluative Research Report No. 359. Amherst, MA: School of Education, University of Massachusetts.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33, 215-230.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true biass/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58, 159-194.

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., & Allalouf, A. (in press). Appraising item equivalence across multiple languages and cultures. Language Testing.

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.

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

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.

Table 1

Descriptive Characteristics for Dual language and English Groups

Characteristic	Dual Language	English 1	English 2	English 3	English Total
No. Examinees	585	585	585	585	76,784
Mean Total Test Score	15.95	31.15	30.77	30.14	30.92
St. Dev. Total Test Score	7.47	10.90	11.10	10.57	10.81
Coefficient Alpha	.81	.89	.89	.88	.89
St. Error of Measurement	3.28	3.62	3.61	3.63	3.63
Mean Multiple-Choice Items	10.59	18.67	18.33	18.14	18.50
St. Dev. Multiple-Choice Items	4.18	5.93	6.05	5.74	5.87
Mean Polytomous Items	5.36	12.49	12.44	11.99	12.43
St. Dev. Polytomous Items	4.00	5.52	5.65	5.43	5.53

Table 2

Descriptive Characteristics for Dual language and Matched English Groups

Characteristic	Dual Language	English M1	English M2	English M3
No. Examinees	585	585	585	585
Mean Total Test Score	15.95	15.95	15.95	15.95
St. Dev. Total Test Score	7.47	7.47	7.47	7.47
Coefficient Alpha	.81	.81	.81	.81
St. Error of Measurement	3.28	3.23	3.24	3.27
Mean Multiple-Choice Items	10.59	10.59	10.59	10.53
St. Dev. Multiple-Choice Items	4.18	4.11	4.35	4.35
Mean Polytomous Items	5.36	5.36	5.26	5.42
St. Dev. Polytomous Items	4.00	3.78	3.87	3.87

Table 3

MDS Fit Indices for Total English Booklet Group

# Dimensions	STRESS	R^2
1	.36	.66
2	.25	.74
3	.20	.78
4	.17	.82
5	.14	.84
6	.12	.87

Table 4

WMDS Fit Indices for Random English and Dual language Groups

# Dimensions	STRESS	R^2
2	.31	.58
3	.24	.70
4	.20	.73
5	.17	.76
6	.15	.79

Table 5

WMDS Group Weights for the Random English and Dual language Groups

	Dimension 1	Dimension 2	Dimension 3
Dual language	.19	.74	.23
English-1	.80	.15	.27
English-2	.77	.13	.30
English-3	.78	.17	.33
Proportion of Variance Accounted for by Dimension	.47	.15	.08

Table 6

WMDS Fit Indices for the Dual language and Matched English Groups

# Dimensions	STRESS	R^2
2	.29	.68
3	.22	.76
4	.19	.78
5	.16	.81
6	.14	.83

Table 7

WMDS Group Weights for the Dual language and Matched English Groups

Group	Two-dimensional Solution		Three-dimensional Solution		
	Dimen. 1	Dimen. 2	Dimen. 1	Dimen. 2	Dimen. 3
Dual language	.62	.45	.36	.62	.46
English-M1	.81	.22	.75	.34	.29
English-M2	.82	.20	.78	.30	.26
English-M3	.84	.16	.78	.36	.20
Proportion of Variance Accounted for by dimension	.60	.08	.47	.18	.10

Table 8

SEM Fit Statistics for Separate (Single Group) Analyses

Group	χ^2	df	RMSEA	RMR	GFI	AGFI
Random English 1	890.78*	702	.021	.012	.93	.92
Random English 2	1003.12*	702	.027	.014	.92	.91
Random English 3	944.00*	702	.024	.013	.92	.92
Matched English 1	740.35*	702	.006	.010	.94	.93
Matched English 2	914.71*	702	.021	.010	.93	.92
Matched English 3	873.09*	702	.021	.011	.93	.92
Bilingual	937.71*	702	.025	.011	.92	.91

* $p < 0.00$

Table 9

SEM Fit Statistics for Multi-Group Analyses: Random English and Dual Language Groups

Model	χ^2	df	RMSEA	RMR	GFI
Equated Factors	4128.84*	2922	.027	.022	.90
Equated Factors Equated Factor Loadings	5043.63*	3039	.034	.037	.86
Equated Factors Equated Factor Loadings Equated Error Variances	5148.09*	3042	.034	.071	.86
Model Comparison					
Model 1 vs Model 2	905.79*	117			
Model 2 vs Model 3	104.46*	3			

* $p < 0.00$

Table 10

SEM Fit Statistics for Multi-Group Analyses: Matched English and Dual Language Groups

Model	χ^2	df	RMSEA	RMR	GFI
Equated Factors	3690.53*	2922	.021	.016	.92
Equated Factors	3875.80*	3039	.022	.018	.91
Equated Factor Loadings					
Equated Factors	3875.92*	3042	.022	.018	.91
Equated Factor Loadings					
Equated Error Variances					
Model Comparison					
Model 1 vs Model 2	185.27*	117			
Model 2 vs Model 3	0.12	3			

* $p < 0.00$

Table 11

Characteristics of Eight Items Flagged For Large DIF

Item Format	Content Area			
	Number Sense	Patterns	Geometry& Measurement	Statistics
Multiple-choice		D		E, E
Short CR	D, D			
Extended CR			D, E	E

Note: "D" indicates an item that favored the dual language group and "E" indicates an item that favored the English language group.

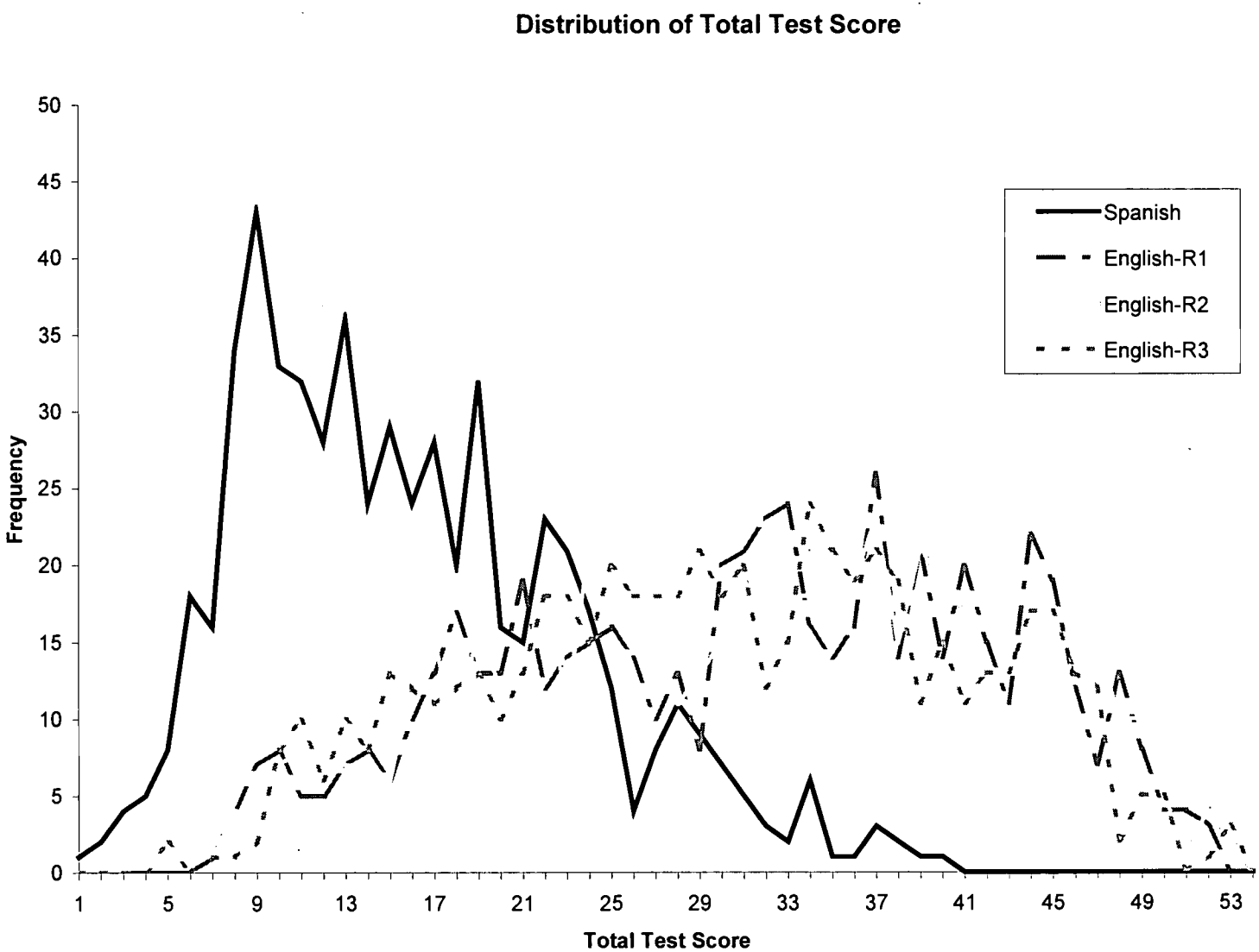
Figure Captions

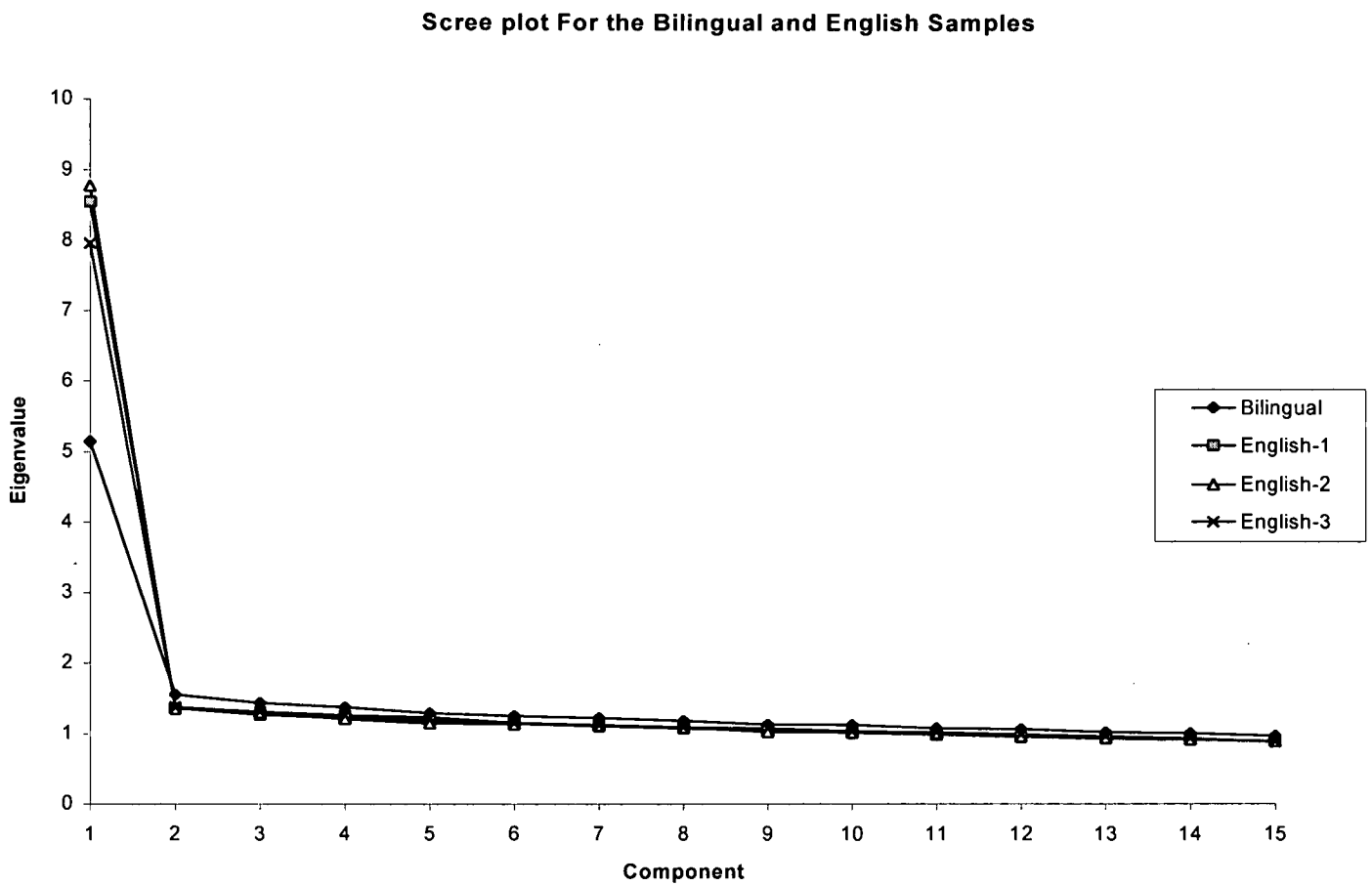
Figure 1. Distribution of the Dual language and English Samples.

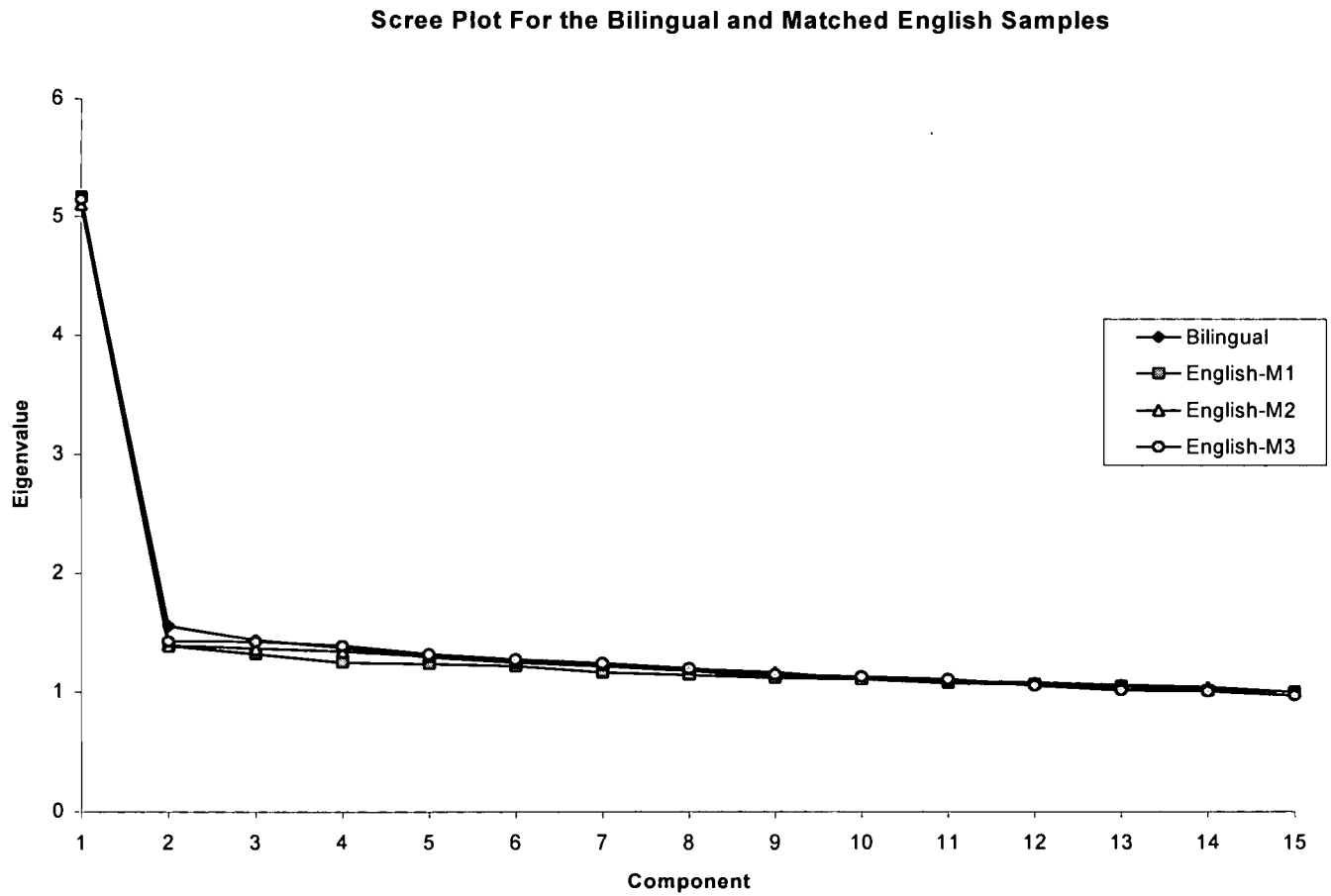
Figure 2. Scree plot of the Dual language and English Samples.

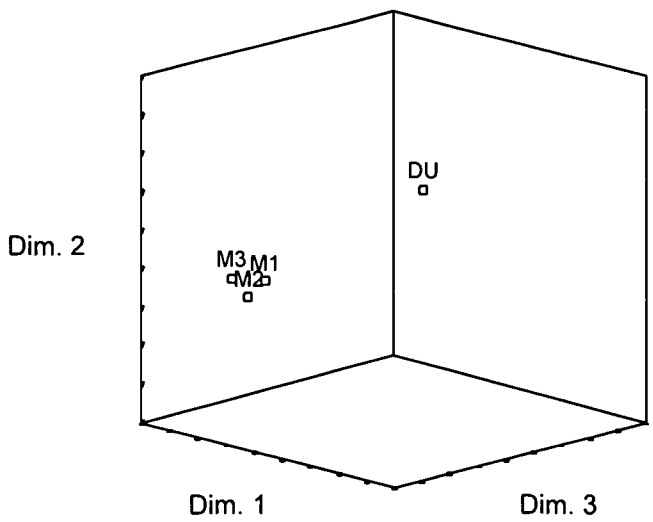
Figure 3. Scree plot of the Dual language and Matched English Samples.

Figure 4. Weight Space for 3-D WMDS Solution: (a) All Items and (b) Large DIF Items Removed.

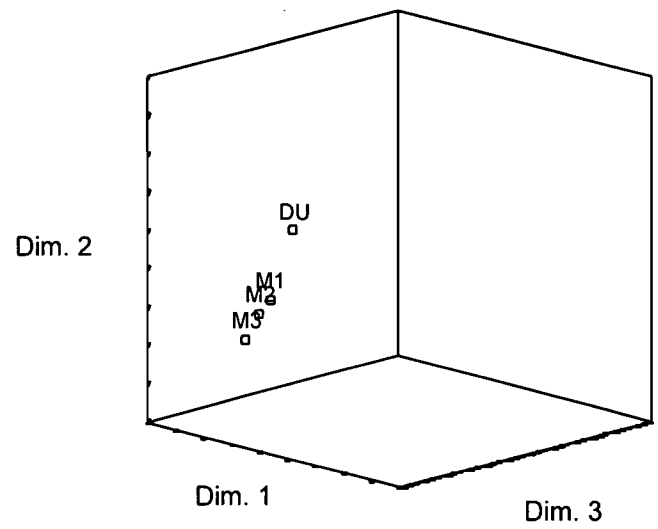

Distribution of Total Test Score
Frequency
Total Test Score
Spanish
English-R1
English-R2
English-R3
0
5
10
15
20
25
30
35
40
45
50
1
5
9
13
17
21
25
29
33
37
41
45
49
53


Scree plot For the Bilingual and English Samples
Eigenvalue
Component
Bilingual
English-1
English-2
English-3
0
1
2
3
4
5
6
7
8
9
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


Scree Plot For the Bilingual and Matched English Samples
Eigenvalue
Component
Bilingual
English-M1
English-M2
English-M3
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



DU=Dual Language, M=Matched English



DU=Dual Language, M=Matched English

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

ERIC

TM034399

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Analysis of the Psychometric Properties of Dual Language Test Forms	
Author(s): Stephen G. Sireci and Shameem Nyla Khaliq	
Corporate Source: University of Massachusetts Amherst	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 / Professor / Univ. of Mass	
Organization/Address:	Telephone: 4135450564	FAX: 4135454181
	E-Mail Address: Sireci@ACAD.UMASS.EDU	Date: 8/25/02

(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:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
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://ericfac.piccard.csc.com