DOCUMENT RESUME

ED 428 120 TM 029 591

AUTHOR Sireci, Stephen G.; Bastari, B.
TITLE Evaluating Construct Equivalence across Adapted Tests.
PUB DATE 1998-08-14
NOTE 35p.; Paper presented at the Annual Meeting of the American Psychological Association (106th, San Francisco, CA, August 14-18, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Construct Validity; *Cross Cultural Studies; Evaluation Methods; Multidimensional Scaling; *Second Languages; Simulation; *Test Format

ABSTRACT

In many cross-cultural research studies, assessment instruments are translated or adapted for use in multiple languages. However, it cannot be assumed that different language versions of an assessment are equivalent across languages. A fundamental issue to be addressed is the comparability or equivalence of the construct measured by each language version of the assessment. This paper presents and critiques several methods for evaluating structural equivalence across different language versions of a test or questionnaire. Applications of these techniques to large-scale, cross-lingual tests are presented and discussed. Simulated data are also used to evaluate the methods. It is concluded that weighted multidimensional scaling and confirmatory factor analysis are effective for helping evaluate construct equivalence across groups. Qualifications for using these procedures to evaluate construct equivalence are provided. (Contains 2 figures, 6 tables, and 42 references.) (Author/SLD)

Evaluating Construct Equivalence Across Adapted Tests[1,2]

Stephen G. Sireci and B. Bastari

University of Massachusetts-Amherst

Avi Allalouf

National Institute for Testing and Evaluation, Jerusalem, Israel

[1]Paper presented at the annual meeting of the American Psychological Association, August 14, 1998, San Francisco, CA.

[2]Laboratory of Psychometric and Evaluative Research Report No. 340. School of Education, University of Massachusetts-Amherst.

Abstract

In many cross-cultural research studies, assessment instruments are translated or adapted for use in multiple languages. However, it cannot be assumed that the different language versions of an assessment are equivalent across languages. A fundamental issue to be addressed is the comparability or equivalence of the construct measured by each language version of the assessment. This paper presents and critiques several methods for evaluating structural equivalence across different language versions of a test or questionnaire. Applications of these techniques to large-scale, cross-lingual tests are presented and discussed. Simulated data are also used to evaluate the methods. It is concluded that weighted multidimensional scaling and confirmatory factor analysis are effective for helping evaluate construct equivalence across groups. Qualifications for using these procedures to evaluate construct equivalence are provided.

Evaluating Construct Equivalence Across Translated or Adapted Tests

Tests, questionnaires, and other types of assessments are commonly used to compare individuals from different cultural groups. In many cross-cultural comparisons, especially international studies, "culture" is confounded with "language." In these cases, assessment instruments are typically translated or adapted across languages so that comparisons among groups and individuals from different cultures can be made. However, it cannot be assumed different language versions of an assessment are equivalent across languages. As stated in the *Standards for Educational and Psychological Testing*, "when it is intended that two versions of dual-language tests be comparable, evidence of test comparability should be reported" (AERA, APA, & NCME, 1985, p.75).

Many other test specialists and cross-cultural researchers have also stressed the need to ensure that instruments translated or adapted across languages are measuring the same construct (e.g., Geisinger, 1994; Hambleton, 1993, 1994; Sireci, 1997, in press; van der Vijver & Poortinga, 1997; van der Vijver & Tanzer, 1998). For example the *Guidelines for Adapting Educational and Psychological Tests* developed recently by the International Test Commission stipulate:

> Instrument developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the instrument, and (2) identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations. (Hambleton, 1994, p. 232)

The first requirement relates to construct equivalence, while the second requirement relates to item bias or differential item functioning. Both lack of construct comparability and item bias can lead to test bias, which implies that inferences derived from test scores are not equivalent across groups.

Although it is clear that evaluating construct equivalence across different language versions of a test is imperative, performing such evaluations is not straightforward. The methods available are complex and tend to require a great deal of technical expertise. Furthermore, the research is equivocal with respect to which methods are preferred.

The purposes of this paper are to describe popular methods for evaluating construct equivalence, present the results of some recent applications of these methods, and provide some suggestions for conducting such studies in the future. The suggestions provided are based on research using both real and simulated data. This paper is limited to the evaluation of construct equivalence across different language versions of an assessment. Issues of item bias or differential item functioning due to translation are discussed elsewhere (e.g., Budgell, Raju, & Quartetti, 1995; Ellis, 1989; Sireci & Berberoglu, 1997).

Methods for Evaluating Construct Equivalence Across Translated/Adapted Tests

Construct equivalence is a very general term that states the same psychological construct is measured across all studied groups. Evaluation of construct equivalence begins by ascertaining the theoretical legitimacy of a construct for all cultural groups of interest and involves full investigation of the nomological network of variables relevant to the construct. The present study explores only one specific aspect of construct equivalence that is particularly important in cross-lingual research: the "structural equivalence" of an assessment instrument across its different language versions. If the dimensional structure of an assessment is found to be consistent across its different language versions, then evidence that the assessment is measuring the same construct in these languages is provided. Test development and test adaptation processes are designed to promote structural equivalence. Thus, evaluating translated or adapted tests for this particular aspect of construct equivalence is an important component of evaluating the validity of inferences drawn from such tests, especially when such inferences are cross-cultural in nature.

Exploratory factor analysis, multidimensional scaling, and confirmatory factor analysis have all been used to evaluate construct equivalence of assessments across different cultural groups. These procedures are described briefly in this section, and applications of these procedures to real and simulated data are presented in subsequent sections.

Exploratory Factor Analysis

Exploratory factor analysis, including common factor analysis and principal components analysis, has been widely used in studying the structure of assessment instruments, including ascertaining structural equivalence across different language versions of an assessment (Anastasi, 1992; Geisinger, 1994; Paunonem, Jackson, Trzebinski, & Forsterling, 1992; van der Vijver & Poortinga, 1997). These studies typically involve performing separate factor analyses for each language group, and then comparing the results. If similar patterns of factor loadings are observed across groups, then evidence of construct equivalence (i.e., structural equivalence) is obtained.

There are at least two limitations of the use of exploratory factor analysis for evaluating construct equivalence. First, the analyses are conducted separately for each group, which makes evaluation of a "common" factor structure difficult. Although items and subscores comprising an assessment may exhibit similar loadings on the first (dominant) factor, interpretations of the secondary factors are difficult because they are not constrained to be in the same order across language groups. A second limitation of exploratory factor analysis is that there are no statistical tests or numerical indices to help determine the degree to which structural equivalence holds across groups.

Weighted multidimensional scaling

Multidimensional scaling (MDS) provides an alternative to exploratory factor analysis for

discovering the structure of assessment instruments Some researchers have argued that MDS is preferred over factor analysis for such investigations (e.g., Davison, 1985) Weighted MDS models, also called "individual differences" models are particularly relevant to the multi-group situation because a common stimulus space (dimensional representation of test structure) can be derived simultaneously for all groups. Furthermore, differences among the groups with respect to dimensional structure are reported using "subject weights." These weights are used to adjust the stimulus space so that it can be "stretched" or "shrunk" to best fit the data for one or more groups. For example, the INDSCAL model proposed by Carroll and Chang (1970) uses a weighted Euclidean distance formula to scale stimuli:

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

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

$$x_{kia} = \sqrt{w_{ka}}\, x_{ia}$$

where x_{kia} represent the coordinate for stimulus 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.

Although weighted MDS models can evaluate test structure simultaneously across all groups, most MDS models do not provide statistical tests of structural equivalence (cf. Ramsay, 1982). Rather, descriptive fit indices are used to evaluate data-model fit. The STRESS index represents the square root of the normalized residual variance of the monotonic regression of the MDS distances on the transformed proximities. Thus, lower values of STRESS indicate better fit. The R^2 index reflects proportion of variance of the transformed proximities accounted for by the MDS distances. Thus, higher values of R^2 indicate better fit. Recent applications of weighted MDS have illustrated its advantages for evaluating structural equivalence across cultural groups (Day & Rounds, 1998; Day, Rounds, & Swaney, 1998) and across different language versions of a test (Sireci, Fitzgerald, & Xing, 1998).

Confirmatory factor analysis

Confirmatory factor analysis (CFA) is becoming an increasingly popular technique for evaluating structural equivalence across different language versions of an assessment (Brown & Marcoulides, 1996; Reise, Widaman, & Pugh, 1993; Robie & Ryan, 1996; Sireci, Fitzgerald, & Xing, 1998). CFA is attractive in this situation because it can handle multiple groups

simultaneously, statistical tests of model fit are available, and descriptive indices of model fit are provided. In multi-group CFA analyses, the hypothesized structure of an assessment is incorporated into a structural equation model, and the structure is constrained to be equal across all groups. A typical hypothesis tested using CFA is whether the factor loading matrix is equivalent across all groups. The structure of the factor loadings is usually an "independent clusters structure" (MacDonald, 1985), which specifies that: 1) each measured variable has a nonzero loading on only the factor it was designated to measure, 2) correlations among the factors (i.e., lower diagonal of the phi matrix) are freely estimated, and 3) the errors associated with the factor loadings (i.e., theta delta matrix) are uncorrelated (Marsh, 1994). In this paper, we use the term "invariant independent clusters structure" to refer to a model that constrains this structure to be equal across two or more groups.

Evaluating Construct Equivalence: Examples

The results from two recent studies of construct equivalence are presented in this section. The first study evaluated the structural equivalence of a subset of items (developed in Hebrew and translated into Russian) taken from the Verbal Reasoning section of the Psychometric Entrance Test (PET). The PET is a high-stakes test used for admissions decisions in universities and colleges in Israel (see Beller, 1994, 1995 for further description of the PET). The second study evaluated the structural equivalence of a high-stakes information technology certification exam (one of Microsoft's Certified Professional exams) across four language groups: English, French, German, and Japanese.

Hebrew and Russian versions of the PET

Allalouf, Bastari, Hambleton, & Sireci (1997) used exploratory factor analysis, MDS, and CFA to evaluate the structural equivalence of two different language versions of the PET. The PET is developed in Hebrew and translated/adapted into five other languages: Arabic, Russian, English, French, and Spanish. The data analyzed came from four of the five content areas composing the Verbal Reasoning subtest: analogies, logic, reading comprehension, and sentence completion. The fifth content area, antonyms, was excluded from analysis because these items are not considered to be equivalent across languages and are not used to equate the different language versions of the PET. A total of 41 items were included in the analysis; all were dichotomously scored multiple-choice items. Only data from the Hebrew and Russian versions of this test were analyzed. The sample sizes were 7,149 and 2,604 for the Hebrew and Russian samples, respectively.

Previous research demonstrated that the Verbal Reasoning subtest of the PET is multidimensional and that the dimensionality corresponded closely to the content areas composing the test (Budescu, 1985; Beller, 1994; Kaplan-Shefer et. al., 1992; Rokas & Melamed, 1996). Allalouf et al. sought to confirm this specific test structure using both exploratory and confirmatory analyses.

Exploratory factor analyses

PCA and non-linear factor analyses of the PET data were conducted. The PCA was conducted using SPSS (version 6.0) and the non-linear factor analyses were conducted using NOHARM (Fraser & MacDonald, 1988) and TESTFACT (Wilson, Wood, & Gibbons, 1991). Tetrachoric correlations among the items were computed for all analyses, and separate analyses were conducted for the Hebrew and Russian data.

The results of the PCA indicated the data were multidimensional. Six components had eigenvalues greater than one for both language versions. For the Hebrew data, the first component accounted for about 16% of the variance in the item data, the second component accounted for about 4% of the variance. For the Russian data, the first two components accounted for about 13% and 4% of the variance, respectively. The cumulative variance accounted for by the six components was 31% and 29% for the Hebrew and Russian data, respectively. The results for the non-linear factor analyses were similar. Both NOHARM and TESTFACT indicated five factors were required to represent the Hebrew version of the PET and six factors were required to represent the Russian version. Although the results were similar across three sets of exploratory factor analysis, interpretation of the factor loadings was clearest for the obliquely-rotated TESTFACT solution. These factor loadings are summarized in Table 1. For both the Hebrew and Russian data, separate factors corresponding to each of the four content areas were observed. For the Hebrew data, five factors were required because the two sets of reading comprehension items (that corresponded to different reading passages) loaded on separate factors. Six factors were required for the Russian data because in addition to the reading comprehension passages, two separate sets of analogy items (one set related to vocabulary analogies, the other to "logic-type" analogies) also loaded on separate factors. Allalouf and Sireci (1998) found that the analogy items from the PET exhibited a large degree of differential item functioning across Hebrew and Russian, which may partly explain this difference in dimensionality.

[Insert Table 1 Here]

Multidimensional scaling analyses

To evaluate the structure of the PET across the Hebrew and Russian versions simultaneously, Allalouf et al. used weighted MDS. As mentioned earlier, weighted MDS models derive a common set of stimulus (item) coordinates for all groups entered into an analysis and a vector of dimension weights for each group. These dimension weights reflect the differential weighting of the dimensions necessary to best account for the correlations among the items for each group. Thus, differences in dimensional weights across language groups would suggest differences in the dimensional structure across language versions of the test. The Russian data were split randomly and separate inter-item tetrachoric correlation matrices were computed for each sample. There were 1,302 Russian examinees in each sample. For the Hebrew data, two

random samples (without replacement) of 1,302 examinees were selected and separate inter-item tetrachoric matrices were also derived Two tetrachoric correlation matrices were derived for each language group so that variation among the weights within each language could be compared with variation among the weights across language groups.

MDS models fit distances to dissimilarity data, not to similarity data. Therefore, the tetrachoric correlations were transformed to dissimilarities using the transformation suggested by Davison (1985):

$$\delta_{ij} = \sqrt{2 - 2r_{ij}}$$

where δ_{ij}=the dissimilarity between item *i* and *j*, and r_{ij}= the tetrachoric correlation between items *i* and *j*.

A five-dimensional MDS solution was accepted as the best dimensional representation of the data. This solution accounted for 48% of the variation in the (transformed) item tetrachoric correlations (about 17% more than the PCA). The first dimension accounted for about 14% of the variance, the second and third dimensions each accounted for about 10% of the variance, and the third and fourth dimensions each accounted for about 7%. The STRESS value for the five-dimensional solution was .19, which is relatively large for a MDS solution. This relatively poor fit is probably reflective of a high level of error (i.e., random variation) in the data. However, all five dimensions were interpretable. The first three dimensions essentially recovered the content areas. The first dimension separated the reading comprehension items from the other items, the second dimension polarized the analogy and logic items, and the third dimension separated the sentence completion items from the other items. The fourth dimension tended to segregate the logic and sentence completion items from one another. The fifth dimension separated the two sets of logic items from one another, and polarized some of the analogy items. The first three dimensions from this solution are displayed in Figure 1. Clusters of items corresponding closely to their content area designations are evident in the figure.

[Insert Figure 1 Here]

The weights for each of the four groups, together with their projections in a two-dimensional subspace, are presented in Figure 2. The weights are very similar for the Hebrew and Russian groups, suggesting that the structure of these dichotomous item response data is similar across the Hebrew and Russian versions of the PET.

[Insert Figure 2 Here]

Confirmatory factor analyses

Separate inter-item tetrachoric and asymptotic covariance matrices were derived from the Hebrew and Russian data using PRELIS-2 (Jöreskog & Sorbom, 1993b) for the confirmatory

factor analyses Four four-factor models were fit to the data using the weighted least squares estimation procedure. In each case, the factor loadings for the items were specified according to their content specifications (i.e., all the analogy items were specified to load on one factor, all the logic items were specified to load on a second factor, etc.). The first model specified a common four-factor structure underlying the data for both groups. The second model constrained the factor loadings to be equal across the Hebrew and Russian data (i.e., independent clusters structure). The third model added the constraint that the errors associated with the factor loadings (i.e., theta delta matrices) were also equivalent across the two groups. The fourth model added the constraint that the correlations among the latent variables were equivalent across the groups (i.e., invariant phi matrix). All models had goodness of fit (GFI) indices above .96, and root mean square errors ranging from .05-.08, suggesting reasonable fit to the data. The results of these analyses are summarized in Table 2.

[Insert Table 2 Here]

Summary of PET analyses

The analyses performed on the Hebrew and Russian PET items tended to compliment one another. In general, the hypothesized content structure of the PET was confirmed and it was found to be similar across the two language versions of the subset of items. Although some minor differences were observed in the exploratory analyses, the CFAs suggest that these differences are not large enough to reject the hypothesis of structural equivalence.

Microsoft's Network Technology Exam

Sireci et al. (1998) analyzed data from a version of Microsoft's Networking Technology Server (NTS) exam, which is one of the four operating systems exams required to become a Microsoft Certified Systems Engineer. Random samples of candidates from four of the most popular language versions of the exam were selected: English (n=2,000), French (n=1,329), German (n=1,576), and Japanese (n=2,000). The NTS exam comprises 55 items measuring six global content areas: planning (5 items), installation and configuration (14 items), managing resources (10 items), connectivity (8 items), monitoring and optimization (8 items), and troubleshooting (10 items).

Similar to Allalouf et al. (1997), PCA, MDS, and CFA were used to evaluate the structure of the examination data across the four language versions of the test. To account for the presence of a high level of error in the item-level data, the PCA analyses were conducted using both item-level data and item parcel data. Thirteen parcels of items were created based on the content specifications of the test and an attempt to balance the difficulty and variability of parcel scores. The thirteen parcels comprised between three to six items. Pearson correlations were computed among the thirteen parcels. Separate matrices were derived for each language version of the test, and the PCA were conducted separately for each version. The Pearson correlations were transformed to dissimialrities using equation 3 for use in the MDS analyses.

For the MDS analyses, the data for each language group were split into two random samples, and separate inter-parcel Pearson correlation matrices were computed for each sample. This procedure provided a total of eight correlation matrices for the analysis: two matrices for each language group.

PCA results

The item-level PCA results exhibited low percentages of variance accounted for by the first factor across all four groups. The variance in the item-level data accounted for by the first factor ranged from 10.4% (French) to 13.0% (German). The number of eigenvalues greater than one ranged from 16 to19 across the four groups. These results are not particularly revealing, except for confirming the expectation of a large amount of error variance present in these item-level data.

The PCA results for the thirteen item parcels were similar across the four language groups in terms of eigenvalues and percentages of variance accounted for by the first factor. In all cases, a one-factor solution fit the data well. The first factor accounted for between 31.2% (French) and 36.4% (German) of the variance among the item parcels. The eigenvalues for the first component were between 4.1 (English and French) and 4.7 (German), and the eigenvalues for the second component were all close to one. The largest proportion of variance accounted for by the second component was 9.3% for the English language sample.

Although the separate-group PCA analyses were similar in terms of variance accounted for, there were some notable differences among the factor loadings across groups. Table 3 gives the factor loading matrix for each language group. For seven of the parcels (parcels 1, 2, 4, 6, 7,12, and 13) the loadings were similar across all four groups. However, for the English data, parcels 3, 5, 8, and 10 exhibited small loadings (i.e., < .30) on the first factor relative to the loadings for the other groups. For the Japanese data, parcels 10 and 11 exhibited different loadings in comparison to the other groups. For the French data, parcel 9 had a loading of zero on the first factor, which was small relative to the other groups. Across the four groups, the factor loadings for the French and German data appear most similar. Given these findings, it appears possible that more than one dimension is necessary to account for the variation among the parcels across the four language groups.

[Insert Table 3 Here]

MDS Results

To evaluate the factor structure among the groups simultaneously, the data for each language group were split randomly (without replacement) and separate inter-parcel correlation matrices were derived for each sample. A three-dimensional solution was selected as the best representation of the data. This solution accounted for 75% of the variation among the

transformed parcel dissimilarities The percentages of variance accounted for by the first through third dimensions were 35%, 23%, and 17%, respectively

The three-dimensional MDS solution is portrayed visually in Figure 3. The first dimension was interpreted as distinguishing between the proactive and reactive aspects of network technology. The second dimension seemed to be related to wiring issues such as installation and connectivity, and the third dimension was related to managing resources.

[Insert Figure 3 Here]

The subject weights for each sample matrix are displayed in Figure 4. The German samples have the largest weights on the first dimension, while the English samples have the smallest weights. The reverse pattern occurs on the second dimension. Comparing the variation among the weights within each language group with the variation among the averaged weights across language groups indicates that the dimensional structures for the German and English versions of the test appear most different. Dimension 1 best accounted for the variation (about 61%) in the German data, while dimension 2 best accounted for the variation (56%) in the English data. More equal weighting of the dimensions was required to account for the variance in the French and Japanese data.

[Insert Figure 4 Here]

CFA Results

Three one-factor CFA models were fit to the data for all four language groups. Polychoric correlations were computed among the item parcels and the maximum likelihood estimation procedure was used for all analyses. The first model specified a single factor underlying the data for all four groups. The second model specified invariant factor loadings across groups, and the third model specified invariant uniquenesses (theta deltas) across the groups. These results are summarized in Table 4. Inspection of the fit values indicates that all three unidimensional models fit the data well. The GFI indices for all models are above .97, and the RMRs are all below .04. Because these one-factor models exhibited reasonable fit, no two-factor models were fit to the data.

[Insert Table 4 Here]

At first glance, the CFA results seem to contradict the PCA and MDS results. Fit of a common unidimensional model to the data was not expected given the differences observed among the PCA factor loadings and the MDS dimension weights. However, the PCA and MDS results are primarily descriptive, and therefore useful for discovering differences among the groups. Although differences among the groups do seem to exist, the results of the CFA suggest these differences are not large enough to warrant different factor structures for one or more groups.

Summary of NTS analyses The structural equivalence of the Microsoft NTS exam over the four studied language versions was evaluated using both exploratory and confirmatory methods. The exploratory methods revealed a dominant factor for all groups, but differences in the factor loadings for some item parcels were observed among the different language versions of the test. However, when the structure of the original English language exam was imposed on the data for the other language versions, adequate fit to the data was observed. These seemingly contradictory findings seem to indicate that although some structural differences can be observed across the four different language versions of the NTS exam, these differences are relatively minor, supporting the view that the same construct is being measured by all four versions of the exam.

Simulation Study

The preceding analyses using real test data are illuminating regarding the types of structural information provided by PCA, MDS, and CFA. However, given the differences observed between the exploratory and confirmatory analyses of the Microsoft data, it is unclear whether CFA provides the final judgment regarding structural equivalence, or whether CFA is not powerful enough to detect departure from structural equivalence. To investigate this issue, data were simulated to model situations of structural equivalence and structural non-equivalence across groups. These data were analyzed using CFA and weighted MDS to discover whether these procedures would correctly identify the conditions of structural equivalence and non-equivalence.

Data generation

Because the Microsoft data provided somewhat equivocal results, the simulated data were generated to have similar characteristics to those data. Data for thirteen measured variables were simulated for four groups. The data were generated to fit one of two structural models. In the first data generation condition (Condition I), an invariant independent clusters structure was simulated. This model posited two correlated factors underlying the data for all four groups, and the factor loadings for the measured variables were equal across all four groups. The first nine measured variables were specified to load on the first factor, the other four measured variables were specified to load on the second factor. This model mimicked the PCA factor loadings for the English Microsoft data.

Condition II simulated a specific situation of structural non-equivalence. In this condition, the factor loadings for three of the four groups were identical (i.e., same factor loadings specified in condition I). The factor loadings for the fourth group had two differences from the other groups. First, the tenth measured variable was specified to load on the first factor instead of the second factor. Second, the magnitude of the factor loadings were specified using the results from the PCA analysis of the French Microsoft data. Thus, Condition II contained two departures from an invariant independent clusters structure: the factor loadings were not equal across all four groups, and one of the measured variables loaded on a different factor in one group than in

the other three Figures 5 and 6 illustrate the pattern of factor loadings for Conditions I and II, respectively.

Two variants of each condition were simulated. In the first condition, the true correlation between the latent variables was .10. In the second condition, the correlation between the latent variables was .60. These two conditions were simulated so that the effect of the degree of latent variable correlation on the ability of CFA and MDS to recover true dimensionality could be observed. The .10 condition represents a very low correlation condition, in which the multidimensionality should be relatively easier to detect. The .60 condition represents a high correlation condition, where the multidimensionality may be more difficult to detect. These two correlation conditions have been widely used to evaluate methods for discovering test structure (e.g., Hambleton & Rovenelli, 1986).

[Insert Figures 5 and 6 Here]

The data for all conditions were generated using MGRPGEN (Rogers, 1996), which is a Fortran program for generating data according to pre-specified structural equation models. The samples of simulated "examinees" ranged from 1,329 to 2,000 for each "language" group, which mimicked the Microsoft NTS sample sizes. The errors associated with the factor loadings (i.e., the theta delta matrix) were specified to be equal to the uniqueness values from the PCA analysis of the English Microsoft data for Condition I, and to the English and French uniquenesses for Condition II. Ten replications for each condition were generated. All 40 data sets (4 conditions X 10 replications) were analyzed using LISREL version 8.0 (Jöreskog & Sorbom, 1993a). In each analysis, an invariant independent clusters model was fit to the data. This was the correct model for the Condition I analyses, but an incorrect model for the Condition II analyses. In addition, another ten replications of a one-factor model were applied to the Condition II data. This model mimicked the model fit to the Microsoft data in the Sireci et al. (1998) study.

The purpose of this simulation study was to discover whether any of the goodness of fit indices and other descriptive statistics associated with the CFA and MDS analyses were sensitive to the lack of structural equivalence in Condition II. This simulation study is not a comprehensive analysis of the utility of CFA and MDS for evaluating structural equivalence. Only one type of structural non-equivalence was simulated, and the degree of non-equivalence was restricted to only one of four groups. The relatively small number of replications also limits the generalizability of the results. However, the main purpose of the simulation was to discover whether these two analytic tools would uncover a type of structural non-equivalence that was suggested in the analysis of real test data. If the LISREL goodness of fit indices were sensitive to the departure from structural equivalence simulated in Condition II, then it is likely that the LISREL results for the Microsoft data really do reflect structural equivalence across the four studied language versions of the exam. On the other hand, if the LISREL fit indices suggest the invariant independent clusters model fits Condition II, then CFA may not be powerful enough for detecting this type of structural non-equivalence.

Simulation study results

CFA results

In the Sireci et al. (1998) study, only the goodness of fit (GFI) and root mean square residual (RMR) indices were reported. However, LISREL version 8 reports over 20 different indices of overall data-model fit. To discover which fit indices were sensitive to the lack of structural equivalence in Condition II, the performances of nine of these fit indices were evaluated. Selection of these nine indices was motivated by a desire to: 1) save time by avoiding those indices requiring comparing several nested models, 2) focus on indices that should be sensitive to misfit in the measurement model (as opposed to the structural model), and 3) focus on indices that are commonly reported in the literature. After reviewing recent research in this area (e.g., Browne & Cudek, 1993; Byrne, 1998; Marsh, 1994; Mulaik, et al., 1989), the nine measures selected were: chi-square, root mean square error of approximation (RMSEA), RMR, GFI, normed fit index (NFI), non-normed fit index (NNFI), parsimony normed fit index (PNFI), comparative fit index (CFI), and incremental fit index (IFI). Although testing hierarchical models is preferred when analyzing goodness of fit in CFA (e.g., Jöreskog & Sorbom, 1996; Reise et al., 1993), 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). There has been much debate about the utility of these indices, however, non-significant chi-squares, RMSEA and RMR values below .10, and values for the other indices of .90 or above, have been proposed as indicators of reasonable data-model fit.

The minimum, maximum, and mean fit statistics across the ten replications for both scenarios of Conditions I and II are presented in 5. All fit indices suggest good data-model fit under Condition I ("true" structural equivalence), regardless of the level of correlation among the factors. For the Condition I analyses, it may seem surprising at first to see non-significant chi-squares with such large sample sizes, but this finding can be explained by the fact that the data were generated using the same model that was fit to these data.

For the Condition II analyses (i.e., structural non-equivalence), only the chi-square test, RMR, GFI, and PNFI were sensitive to the lack of structural equivalence under the low correlation condition. Although the values for the other fit indices moved in the direction of relatively poorer fit under condition II, all provided values that would probably be interpreted as reflecting reasonable data-model fit, which in this case is the wrong conclusion.

For the high correlation Condition II analyses, only two fit indices correctly suggested rejection of the structural equivalence model: the chi-square test and the RMR. Given the fact that most researchers seeking to obtain data-model fit ignore the chi-square test when large sample sizes are used, the RMR appears most useful for evaluating structural equivalence. Other popular measures, such as the CFI, GFI, and RMSEA provided values that would probably be interpreted as suggesting reasonable data-model fit, which again is the wrong conclusion.

[Insert Table 5 Here]

Given the fact that a one-factor CFA model was fit to the Microsoft data, it is also of interest to evaluate the performance of the LISREL fit indices when fitting a one-factor model to the Condition II data. This one-factor model (that restricted the factor loadings to be equal across the groups) is inappropriate for these data on two accounts: 1) two factors underlie the data for all groups, and 2) the factor loadings are not equivalent across all groups. The results of applying this one-factor structural equivalence model to the Condition II data are reported in Table 6. Under the low correlation condition, all fit indices suggest poor data-model fit. However, for the high correlation condition, the results were more equivocal for several indices, including the GFI. The RMR and six other indices (NFI, NNFI, PNFI, CFI, IFI, and RFI) correctly suggested poor data-model fit under this condition. Although it is encouraging these indices were on-target, it is important to note the relatively poor performance of the widely used GFI under this condition.

[Insert Table 6 Here]

It is interesting to note the relatively small ranges for all of the fit indices across all replicated conditions. Although only ten replications of each condition were simulated, the very minor changes in the fit statistics suggest that the results probably will not change dramatically if the number of replications were increased.

Given these findings, a natural question is "What were the values of the other fit indices for the structural equivalence model fit to the NTS data?" As previously reported, the RMR was very low (.032) suggesting good data-model fit. This is an important finding because this index performed well under all simulation conditions. All other indices also suggested good fit (e.g., the lowest fit index was the PNFI, which equaled .91). Thus, the simulation results support the conclusion that the structure of the Microsoft NTS exam is equivalent across the four language versions studied.

MDS results

The INDSCAL MDS model was also fit to all four data simulation conditions. For the low and high correlation scenarios under Condition I, and for the low correlation scenario under Condition 2, MDS performed very well. The two-dimensional solutions fit like a glove for these three scenarios (the largest STRESS was .08 and the lowest R^2 was .96 across these 30 replications). Under condition I, the stimulus coordinates captured the specified factor loadings on the first dimension. Those variables specified to load on the first factor had large positive coordinates on this dimension, and the variables specified to load on the second factor had large negative coordinates.

Under the low correlation scenario for Condition I, the first dimension accounted for almost all (at least 98%) of the variance in the data and no "groups" exhibited weights larger than

.02 on the second dimension, for eight of the ten replications For the other two replications, the second dimension accounted for about 22% of the variance and all four groups had similar weights on both dimensions For the high correlation condition, the subject weights were also similar in every replication, but the variance accounted for by the second dimension was relatively larger (ranging from 9% to 38% across the 10 replications). All interpretations of the MDS analyses under Condition I (both low and high correlation) correctly led to the conclusion of structural equivalence across groups.

For the low correlation scenario under Condition II, the two-dimensional MDS model recovered the factor loadings for the first three groups on the first dimension, and recovered the factor loadings for the fourth group on the second dimension. Across the ten replications, the dimension weights were essentially 1.0, 0.0 for the first three groups and 0.0, 1.0 for the fourth group—exactly what the model would "predict." The R^2 was 1.00 for all analyses (the first dimension accounted for 75% of the variance and the second dimension accounted for 25%). Thus, MDS had no trouble discovering structural non-equivalence, and specifying the cause of the non-equivalence, under the low correlation condition.

MDS did not perform as well under the high correlation scenario of Condition II. Three dimensions were needed to fit the data properly, and even then, the STRESS values were "borderline" (median STRESS was .11). However, the R^2 values were consistently high (median=.97). Inspection of the dimension coordinates generally provided the following interpretation: one dimension accounted for the pattern of factor loadings generated for the first three groups, a second factor accounted for the pattern of factor loadings generated for the fourth group, and a third factor polarized two or three variables that were specified to load on the same factor for the first three groups. In all replications, the first three groups had large weights on two of the factors (that together accounted for about 73% of the variance) and a weight on the other factor near zero. The inverse pattern of weights appeared for the fourth group (a weight near 1.0 on one dimension and weights near zero on the other two). Thus, under this condition, MDS was able to detect the lack of structural equivalence (via the group weights), but the coordinates for the measured variables were difficult to interpret for one of the three factors, even though the factor typically accounted for about 20% of the variance.

The results from the MDS analysis of the simulated data are informative with respect to MDS analysis of the Microsoft NTS data. Although differences in the dimension weights were observed across the four language groups, none of the groups exhibited weights near zero on any of the dimensions. Thus, if any structural differences do exist across the groups, they are probably very minor and reflect subtle differences across dimensions that are highly correlated.

Discussion

The extensive set of analyses conducted using real and simulated data provided a great deal of information regarding the factorial structure of the studied tests, as well as information regarding the utility of the data analytic procedures for illuminating test structure. With respect to

the PET and NTS exams, the sum of the analyses supports the view that the structure of these tests is equivalent across the different language groups studied This finding is encouraging with respect to the PET because there is a large body of research on the structure of the PET that has informed the test development process (e.g., Budescu, 1985; Kaplan-Shefer et. al., 1992; Rokas & Melamed, 1996). In the present study, the four content areas studied were "recovered" and this structure appeared appropriate for both the Hebrew and Russian data. Of course, the analyses should be replicated on other test forms and across the other language versions of the PET.

The finding of structural equivalence for the NTS exam is also encouraging because for this exam, it is clear the construct that is intended to be measured is legitimate for each language group and is the same in each language group. That is, all language groups are tested on their proficiency with the same computerized software. If large structural differences were observed for these data, clearly, they would point to translation problems, not to construct non-equivalence. Replications of the analyses conducted in this study on other samples of NTS examinees should shed light on the meaningfulness of some of the minor differences noticed in the exploratory analyses.

It is interesting to note that previous research on the same PET data (Allalouf & Sireci, 1998) and NTS data (Sireci, Fitzgerald, & Xing, 1998) did identify some items as functioning differentially across languages. Although these differentially functioning items may affect mean differences across the groups, they do not appear to be large enough, or numerous enough, to cause structural differences.

With respect to the performance of the different methods for evaluating structural equivalence across groups, CFA and MDS appear to be appropriate, but qualifications are necessary for each method. When using CFA, the GFI index should not be relied upon for evaluating data-model fit. This index, and many others, suggested reasonable model fit to the data when the factor loadings were not equivalent across all groups. In the present study, the RMR index performed well, and we suggest its use in future studies. However, our simulation study was extremely limited. Only one specific type of structural non-equivalence was simulated and the number of replications was small. Future research should evaluate the performance of the RMR and other fit indices under varying conditions of structural non-equivalence, number of measured variables, factor loading patterns, and number of groups studied. Also, it is important to note that the present study focused on overall measures of model fit. A more detailed inspection of model fit, for example, evaluating residuals, modification indices, and other elements of the solution, is recommended when evaluating structural equivalence across groups (e.g., Reise et al., 1993). Although not reported above, inspection of the residuals and other elements of the solutions did not signify poor fit for any of the analyses, with one exception. For the simulated data under Condition II (where the fourth group had a different pattern of factor loadings), the estimated correlation between the factors was consistently higher for the "misfit" (fourth) group in comparison to the other three groups. This finding should be researched further to discover if it could be indicative of lack of structural equivalence.

With respect to using weighted MDS to evaluate structural equivalence, two findings cause us to offer qualifications before endorsing MDS for this purpose First, the MDS solution for the NTS exam suggested differences in structure among the language groups that were not supported by the CFA results. Second, under the high correlation/non-structural equivalence condition, MDS analysis of the simulated data yielded dimensions that were not directly interpretable, and may have been spurious. A comparison of the MDS group weights, however, was accurate under all conditions. Thus, when using MDS to evaluate structural equivalence across groups, we recommend focusing on the group weights and interpretable dimensions only. For example, if a dimension is interpretable and large differences are noted in the group weights associated with that dimension, the solution may represent lack of structural equivalence. Based on the simulation results, the weight differences should be large (e.g., one or more groups have weights near zero on a dimension, and one or more other groups have large weights on the dimension) before concluding such differences represent departure from structural equivalence.

In summary, the results of this study suggest that exploratory and confirmatory procedures can be used in complimentary fashion to help evaluate construct equivalence across different language versions of an assessment. If the focus is on discovering subtle structural differences across groups, exploratory procedures are recommended. Weighted MDS is particularly suited to this focus because all language groups can be included in a single analysis. If the focus is on discovering whether a specific factor structure is appropriate for all groups, CFA is recommended. As stated earlier, statistical evaluation of test structure is only one aspect of evaluating construct equivalence across different language versions of an assessment. Evidence of structural equivalence should be taken together with other evidence of the legitimacy of the construct in each language group, and evidence of the similarity of the construct's nomological network across language groups, before concluding construct equivalence holds across different language versions of an assessment.

References

Allalouf, A., Bastari, B., Hambleton, R. K., & Sireci, S. G. (1997). Comparing the dimensionality of a test administered in two languages. Laboratory of Psychometric and Evaluative research report no. 319. Amherst, MA: University of Massachusetts, School of Education.

Allalouf, A., & Sireci, S.G. (1998, April). Detecting the causes of differential item functioning in translated verbal items. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Anastasi, A. (1992). Introductory remarks. In K.F. Geisinger (Ed.) Psychological Testing of Hispanics (pp. 1-7). Washington, DC: American Psychological Association.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. Educational Measurement: Issues and Practice, 13, 12-21.

Beller, M. (1995). Translated versions of Israel's Interuniversity Psychometric Entrance Test (PET). In T. Oakland & R. K. Hambleton (Eds.) International perspectives on academic assessment, (pp. 207-217). Norwell, MA: Kluwer Academic.

Brown, R., & Marcoulides, G. A. (1996). A cross-cultural comparison of the Brown Locus of Control Scale. Educational and Psychological Measurement, 56, 858-863.

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.

Budescu, D. (1985). Factor analysis of the 1995 inter-universities entrance exams (Report No. 21). National Institute for Testing and Evaluation, Jerusalem.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19, 309-321.

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.

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

Day, S. X., & Rounds, J. (1998). Universality of vocational interest structure among racial and ethnic minorities. American Psychologist, 53, 728-736.

Day, S. X., Rounds, J., & Swaney, K. (1998). The structure of vocational interests for diverse racial-ethnic groups. Psychological Science, 9, 40-44.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74, 912-920.

Fraser, C., & McDonald, R.P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269

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

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

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., & Rovenelli, R. J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Jöreskog & Sorbom (1993a). LISREL-8 [computer program]. Mooresville, IN: Scientific Software.

Jöreskog & Sorbom (1993b). PRELIS-2 [computer program]. Mooresville, IN: Scientific Software.

Jöreskog & Sorbom (1996). LISREL 8: Structural equation modeling with the SIMPLIS command language. Mooresville, IN: Scientific Software.

Kaplan-Shefer, Ben Simon, & Cohen (1992). Assessing the dimensions of the verbal item bank (Report No. 165). National Institute for Testing and Evaluation, Jerusalem.

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

McDonald, R. P. (1985). Factor analysis and related methods. Hillsdale, NJ: Lawrence Erlbaum.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifacted approach. Structural Equation Modeling, 1, 5-34.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness of fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Paunonen, S. V., Jackson, D. N., Trzebinski, J., & Forsterling, F. (1992). Personality structures across cultures: A multimethod evaluation. Journal of Personality and Social Psychology, 62, 447-456.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. Journal of the Royal Statistical Society, 145, 285-312.

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.

Robie, C. & Ryan, A. M. (1996). Structural equivalence of a measure of cross-cultural adjustment. Educational and psychological measurement, 56, 514-521.

Rokas, S. & Melamed, E. (1996). The structure of the psychometric entrance exam: analysis in the item level (Report No.). National Institute for Testing and Evaluation, Jerusalem.

Rogers, H. J. (1996). MGRPGEN: A FORTRAN program to generate data for multiple group LISREL analyses. Amherst, MA: University of Massachusetts, School of Education.

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

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: Lawrence Erlbaum.

Sireci, S. G., & Berberoglu, G. (1997, March). Evaluating translation DIF using bilinguals. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). Adapting credentialing examinations for international uses. Laboratory of Psychometric and Evaluative Research Report No. 329. Amherst, MA: University of Massachusetts, School of Education.

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

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.

Wilson, D, Wood, R., & Gibbons, R. D. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [computer program]. Mooresville, IN: Scientific Software.

Table 1 Rotated TESTFACT Factor Loadings for Hebrew and Russian PET Data

Item	Content Area	Heb. F1	Heb F2	Heb. F3	Heb. F4	Heb. F5	Rus. F1	Rus. F2	Rus. F3	Rus. F4	Rus. F5	Rus F6
1	AL								45			
2	AL			30					42			
3	AL			49					26			
4	AL			46								
5	AL			60					25			
6	SC					47			30			
7	SC					40					25	
8	SC					39						
9	SC					32					26	
10	SC					43					45	
11	LO	33					28		29			
12	LO	33					24					
13	LO	41					29					
14	LO	53					30		30			
15	LO	29										
16	LO	52					22					
17	RC		70					66				
18	RC		52					51				
19	RC		63					69				
20	RC		71					57				
21	RC		61					51				
22	AL											
23	AL			52								43
24	AL			37								36
25	AL			43								30
26	SC					51						
27	SC					34					43	
28	SC					34						
29	SC					42					70	
30	SC					30					27	
31	LO	57					51					
32	LO	45					49					
33	LO	49					52					
34	LO	53					47					
35	LO	54					52					
36	LO	41					45					
37	RC				52					58		
38	RC				53					72		
39	RC				31					30		
40	RC				63					41		
41	RC				43					31		

Notes: Decimals and loadings less than .25 are omitted.

Table 2

Goodness of Fit Results for CFA of PET Data

Model	GFI	RMR
One Common Factor for all Groups	.97	.057
Equivalent Factor Loadings (Λ_x) for all Groups	.96	.060
Equivalent Errors of Factor Loadings (Θ_δ) for all Groups	.96	.066
Equivalent Correlations Among Factors	.96	.076

Table 3

PCA Rotated Factor Loading Matrix for Microsoft NTS Data

		1st Factor				2nd Factor			
Content Area[a]	Parcel	English	French	German	Japanese	English	French	German	Japanese
1	1	.56	.52	.47	.51	.34	.43	.41	.32
2	2	.41	.54	.46	.56	.55	.33	.49	.30
2	3	.10	.56	.48	.58	.72	.23	.32	.29
2	4	.67	.56	.61	.53	.23	.35	.37	.47
3	5	.22	.58	.59	.46	.62	.14	.18	.33
3	6	.66	.64	.62	.51	.19	.09	.22	.29
4	7	.61	.55	.44	.44	.13	.21	.43	.49
4	8	.00	.37	.28	.48	.73	.24	.54	.22
5	9	.61	.00	.28	.56	.17	.75	.64	.16
5	10	.28	.16	-.08	.72	.39	.63	.78	-.18
6	11	.58	.59	.65	.00	-.01	-.14	.08	.79
6	12	.65	.55	.51	.58	.20	.34	.51	.36
6	13	.36	.28	.49	.32	.32	.48	.35	.50

[a]1=planning, 2=installation and configuration, 3=managing resources, 4=connectivity, 5=monitoring and optimization, and 6=troubleshooting.

Table 4

Goodness of Fit Results for CFA of Microsoft NTS Data

Model	GFI	RMR
One Common Factor for all Groups	.99	.021
Equivalent Factor Loadings (Λ_x) for all Groups	.99	.031
Equivalent Errors of Factor Loadings (Θ_δ) for all Groups	.98	.032

Table 5

Fit Indices for Simulated Data Fit to Two-Factor Independent Clusters Model

	Condition I						Condition II					
	r=.10			r=.60			r=.10			r=.60		
Index	Mean	Min.	Max.	Mean	Min.	Max.	Mean	Min.	Max.	Mean	Min.	Max.
Chi-square	275[a]	255	304	289[a]	264	329	2240[b]	2206	2315	1443[b]	1270	1543
RMR	.018	.016	.021	.016	.014	.021	.171	.168	.176	.125	.118	.133
PNFI	.938	.936	.938	.937	.937	.938	.875	.872	.877	.903	.900	.908
GFI	.995	.994	.995	.995	.994	.996	.883	.879	.888	.918	.910	.927
NFI	.992	.990	.992	.991	.991	.992	.925	.922	.928	.955	.952	.960
NNFI	1.000	1.00	1.001	1.000	.999	1.001	.930	.928	.933	.962	.959	.966
CFI	1.000	1.00	1.000	1.000	.999	1.000	.934	.932	.937	.964	.961	.969
IFI	1.000	1.00	1.001	1.000	1.00	1.001	.934	.932	.937	.964	.961	.969
RMSEA	0.000	0.00	.002	.001	0.00	.004	.031	.030	.032	.024	.022	.025

[a]None of the chi-squares were statistically significant (at $p < .10$) under these conditions.

[b]All of the chi-squares were statistically significant (at $p < .001$) under these conditions.

Table 6

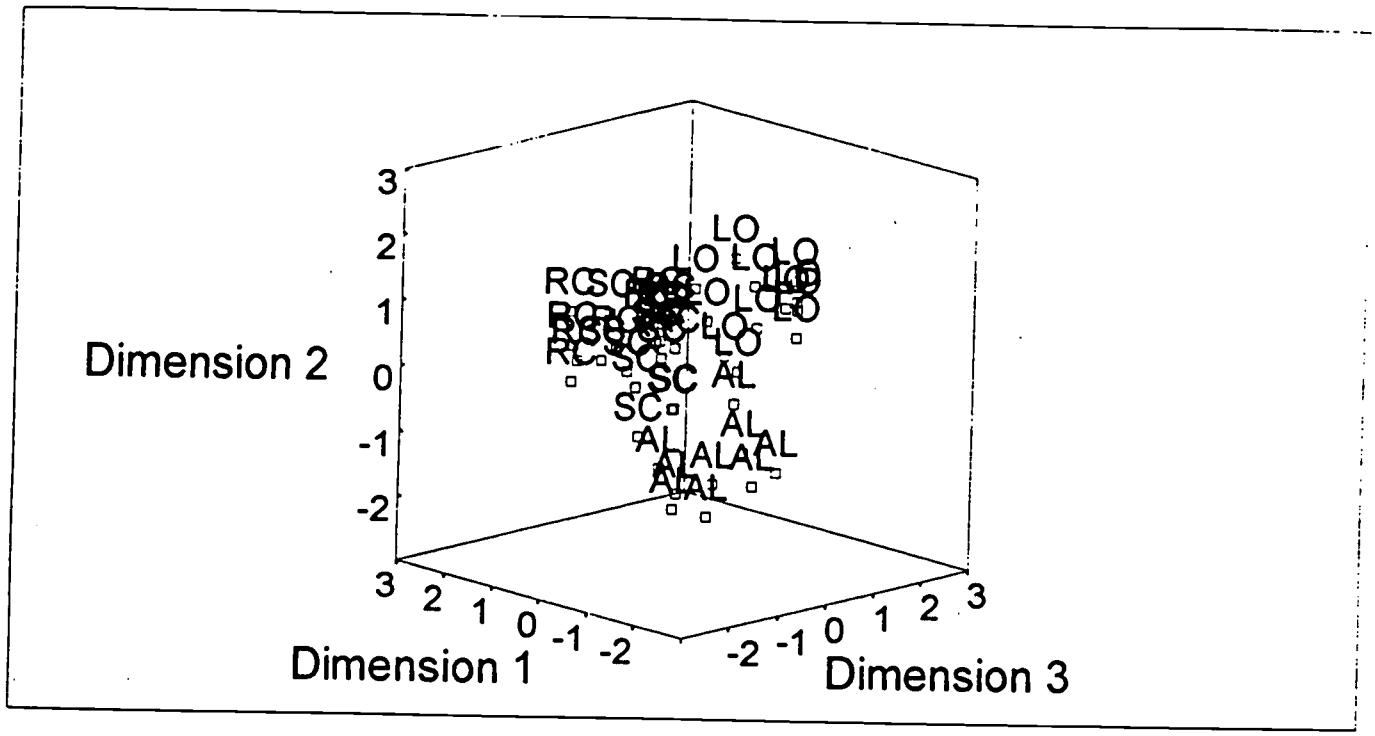
Fit Indices for Simulated Data Fit to One-Factor Model

	r=.10			r=.60		
Index	Mean	Min.	Max.	Mean	Min.	Max.
Chi-square	9292 a	8927	9371	5515 a	5165	5705
RMR	.142	.140	.147	.109	.102	.117
PNFI	.661	.658	.667	.793	.786	.801
NFI	.690	.687	.692	.828	.821	.836
NNFI	.684	.679	.685	.828	.821	.837
CFI	.697	.692	.703	.835	.828	.844
IFI	.697	.692	.703	.835	.828	.844
GFI	.856	.850	.859	.905	.893	.916
RMSEA	.066	.065	.066	.051	.049	.051

[a]All of the chi-squares were statistically significant (at $p < .001$) under these conditions.

Figure 1

3-D MDS Subspace of PET Data

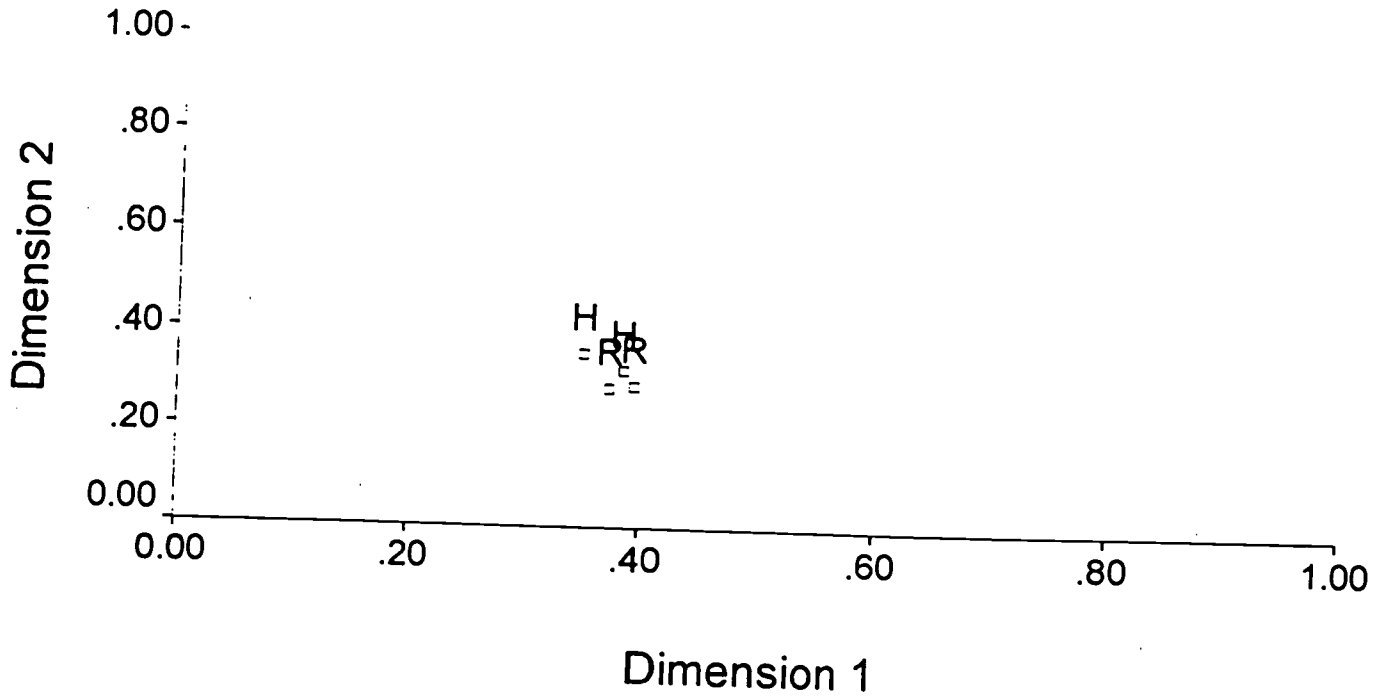


Note: AL=Analogy, LO=Logic,

RC=Reading Compr., SC=Sentence Compl.

Figure 2

Group Weights for PET Data



Note: H=Hebrew, R=Russian

Group Weights

Dimension

Group	1	2	3	4	5
Hebrew	.39	.32	.28	.25	.23
Hebrew	.35	.36	.32	.28	.23
Russian	.39	.29	.30	.28	.29
Russian	.37	.29	.31	.28	.33

Figure 3

3-D MDS Stimulus Space for NTS Data

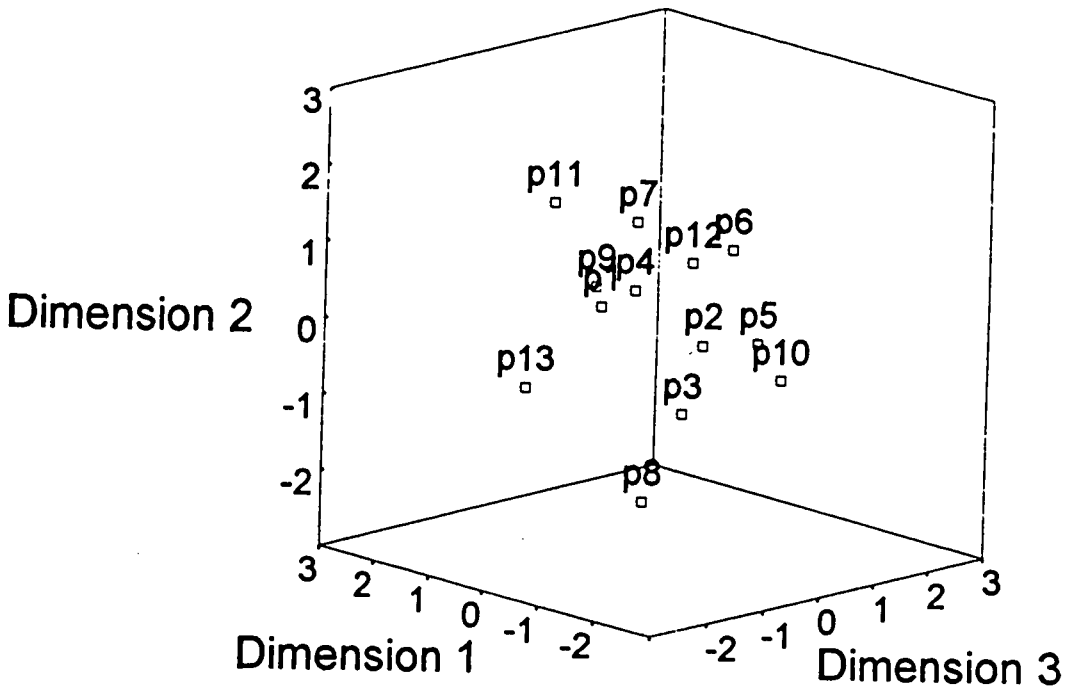
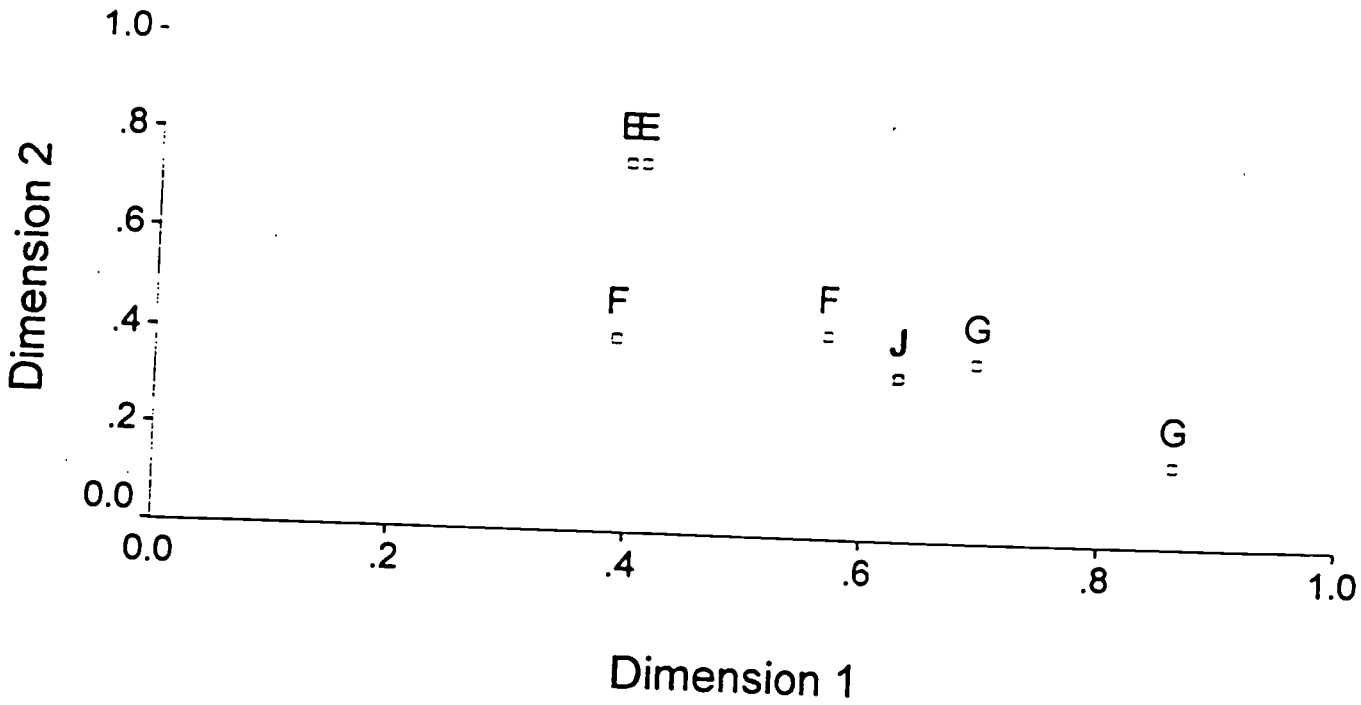


Figure 4

Group Weights for NTS Data



Note: E=English, F=French, G=German, J=Japanese

Group Weights

Group	Dimension 1	2	3
English	.41	.75	.29
English	.40	.75	.34
French	.57	.42	.45
French	.39	.40	.63
German	.70	.37	.32
German	.86	.17	.22
Japanese	.63	.33	.43
Japanese	.63	.34	.43

Figure 5

Structure of Condtion I Simulated Data

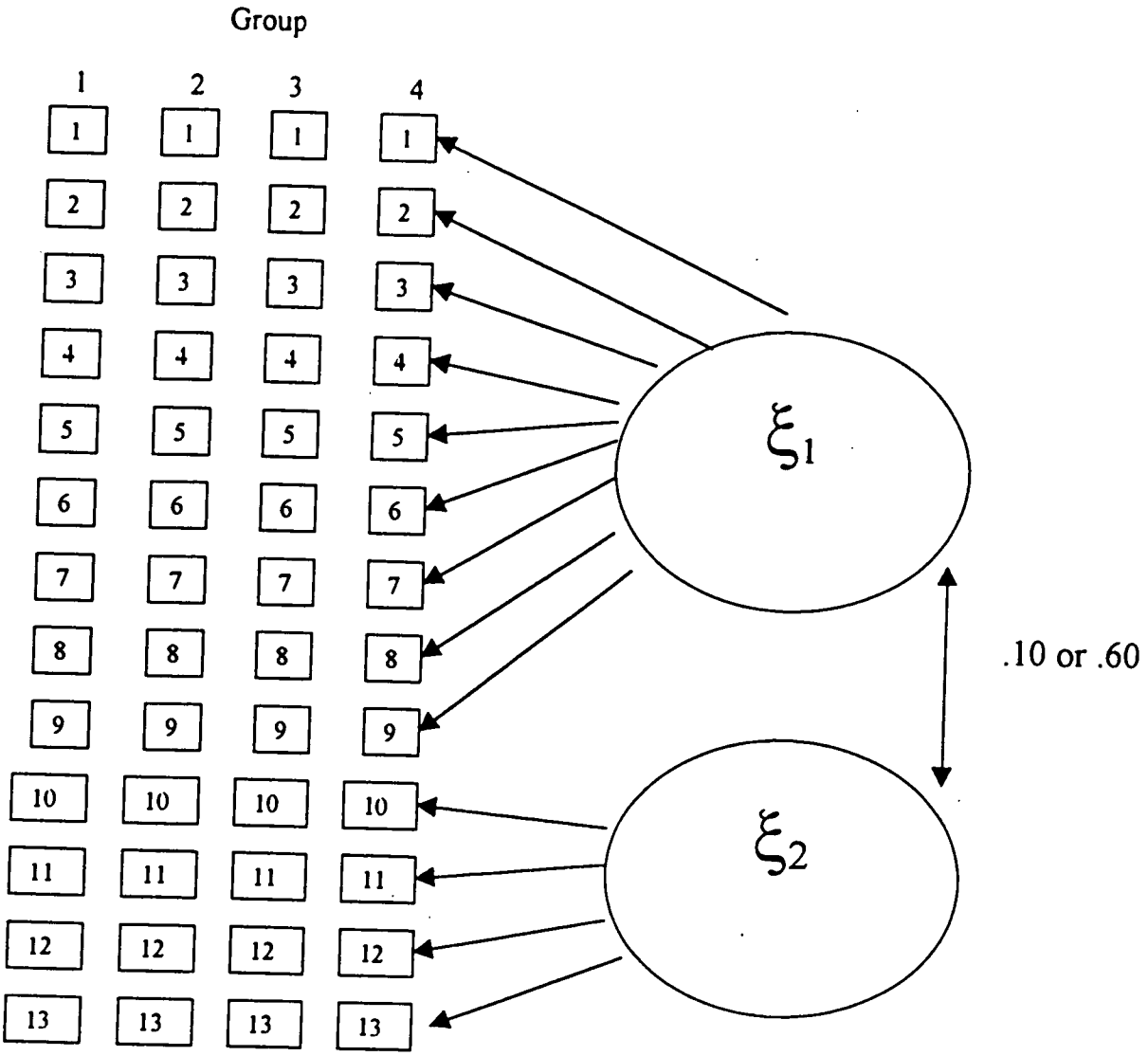
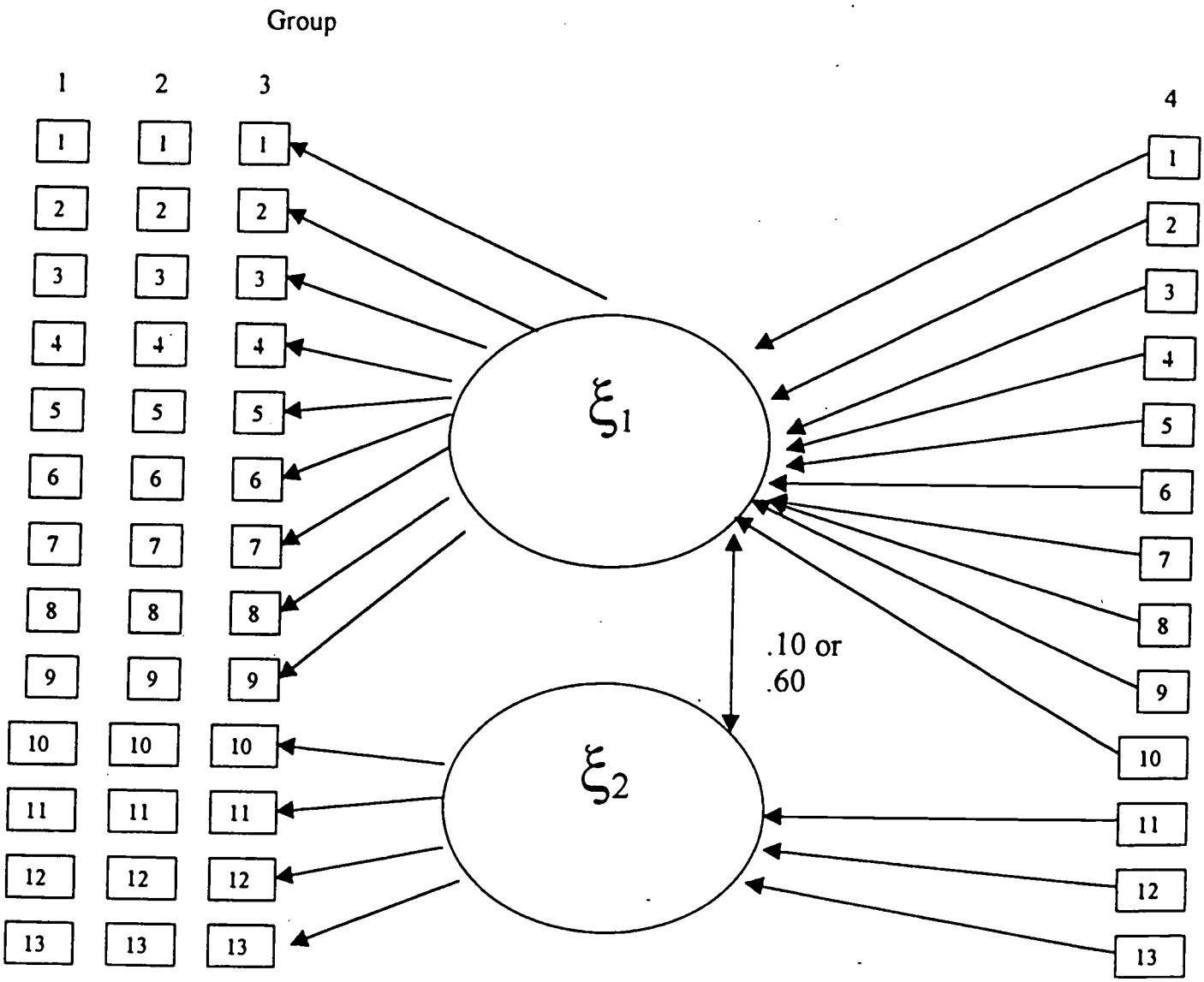


Figure 6

Structure of Condtion II Simulated Data



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

ERIC

TM029591

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:
Evaluating Construct Equivalence across Adapted Tests

Author(s): Sireci, S.G. ; Bastari, B. ; Allalouf, A.

Corporate Source:
University of Mass. - Amherst

Publication Date:
1998

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

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

Level 2A ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ☐

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:
University of Massachusetts
School of Education
Hills South
Amherst, MA 01003

Telephone: 413 545 0564
FAX:

E-Mail Address: sireci@acad.umass.edu
Date: 2/5/99

(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:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
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
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com