Doc 4.10

Comparability of English and Spanish/English Mathematics Tests: Measurement Invariance

Oregon Department of Education
October 2010

Introduction

In international testing, the goal is to develop translated and adapted tests for use in different languages that have comparable factor structures (see Hambleton, 1994). Similarly, a reasonable requirement for any side-by-side translation used as a test accommodation is that the scores produced by the translated test have a comparable factor structure and demonstrate measurement equivalence with the test scores produced by the standard administration. If the ultimate goal is to make meaningful and comparable decisions with the scores produced by the side-by-side accommodation and the scores obtained from the standardized administration, one needs to examine both the structural and metric equivalence of the scores produced by the different administrative practices.

Oregon utilizes the Rasch or one-parameter model to score and scale its tests (Wright & Stone, 1979). Rasch models a latent variable that predicts a score based on the items that comprise the test and the number of correct item responses. Since the Oregon Knowledge and Skills Tests are adaptive, the items that comprise the test are often very different from one person to the next. When the test is unidimensional and the local independence assumption holds, items are combined additively to operationalize the underlying construct. Two research questions are investigated: Will different conditions for administering and scoring mathematics achievement yield an invariant factor structure measuring a single attribute or construct? And, relatedly, will the metric of latent scores be invariant across administrative conditions?

The added challenge of fairly presenting items in different languages to produce comparable scores is not made without evaluating the structural equivalence of the test given the decisions being made with the scores. Using a common factor framework, confirmatory factor analysis is used to test the measurement equivalence or invariance of the side-by-side scores of the test compared to scores produced using standardized administrations of the Oregon test (Sireci & Allalouf, 2003; Wu, Li, & Zumbo, 2007). The common factor model for the two-group comparisons is represented below:

$$
\begin{aligned}
y_{1i} &= \tau_{1i} + \lambda_{1i}\,\xi + \delta_{1i} \\
&\ldots \\
y_{2j} &= \tau_{2j} + \lambda_{2j}\,\xi + \delta_{2j} \qquad (1)
\end{aligned}
$$

A set of five strand scores or subscales summarize mathematical categories of content taught to students, including: 1. computation and estimation, 2. measurement, 3. statistics and probability, 4. algebraic relations, 5. geometry. The side-by-side and standard-administrative groups each received scores for the y_{1i} to y_{2j} subscales shown above. These subscores were employed as manifest or observed variables in the multigroup framework of the common factor model. Each λ represents a subscale's factor loading for each group (1 or 2) on the latent variable, ξ, that is found in the vector of regression slopes. Each τ symbolizes individual regression intercepts for each group on each manifest variable that is placed within the vector of regression intercepts. Finally, an error term, δ, refers to a vector of unique variances whose correlation is assumed to be 0 with ξ, the latent variable. Because a

Rasch model for dichotomous variables assumes unidimensionality, one hypothesizes a single latent or hypothetical construct of interest, ξ, whose variation signals concomitant variation in the five manifest or observed scores received by individuals within each group.

Joreskog's fundamental covariance equation shown below is next used to estimate the implied variance/covariance matrices for each group (see Joreskog and Sorbom, 1993).

$$\Sigma^{1} = \Lambda_{y}^{1} \Phi^{1} \Lambda_{y}^{1} + \theta_{\delta}^{1}$$

$$\ldots$$

$$\Sigma^{2} = \Lambda_{y}^{2} \Phi^{2} \Lambda_{y}^{2} + \theta_{\delta}^{2} \qquad (2)$$

Where Σ is the implied variance/covariance matrix for each group, and Λ is the matrix of subscale factor loadings for each group. While the Φ matrix summarizes the variances and covariances of the latent variable, ξ, for each group, the θ matrix represents the unique variances associated with each subscore and group's respective error term. These equations are primarily employed to produce the fit statistics and indices used in the analysis.

Comparing the factor structure and measurement invariance of the observed scores may be defined with varying degrees of stringency by constraining different parameters at various stages of the analysis (see Vandenberg and Lance, 2000). Employing the above two-group model, several hypotheses are tested within an hierarchical framework in an effort to answer the research questions. First, a separate unidimensional model is separately fit to each group's observed scores to examine their dimensionality and structure. Second, an omnibus test of the equality of implied variance/covariances , Σ , across groups determines whether the latent trait measures the same conceptual framework or factor structure. Third, by constraining the λ's associated with each observed score to be equal across strand scores and groups, one may test whether the regression slopes linking manifest variables, y_{1i} to y_{2j}, to the latent variable are metric invariant. Here, metric invariance implies that the scales are equivalently applied and comparable. Fourth, by constraining the τ's associated for each observed score to be equal across strand scores and groups, one may test the scalar equivalence of the measurement scale. Here, scalar equivalence implies that latent scores are not unit consistent across test administration conditions and this difference in the latent means favors lower performers (see Wu, Li, and Zumbo, 2007). Finally, fifth, by constraining the δ's to be equal across strand scores and groups, one may test whether the unique variances are equal across groups. This final test would imply that the latent trait is being measured with the same level of reliability.

Sample

A sample of 750 student records is randomly selected from both the side-by-side examinee file and the standard examinee file for each of the seven grade levels (i.e., 3, 4, 5, 6, 7, 8, and 10) tested. Since there was one sample from each of the 7 grades studied, there were 7 samples from the standard examinee

files and 7 samples from the side-by-side examinee files. Student records were only selected when they completed all 40 items presented during the test. Since students who take the Spanish Side-by-Side test in mathematics are largely self selected, it is difficult to evaluate their comparability in an even handed fashion. For this reason, successive tests of equivalence are applied to evaluate varying degrees of scale comparability.

Significance Tests and Fit Indices

An array of chi square significance testing is first performed test for group differences in test scores generated using the standard test administrative procedures and the side-by-side accommodation. The first step is separately fit a unidimensional model to the data sampled from the examinee file. Applying Confirmatory Factor Analysis (CFA) to single groups, the fit of the one factor model is evaluated by examining the overall chi-square tests for both the standard administration and the Spanish accommodated mathematics tests. The test of the null maintains that each model is a one factor model, yielding an obtained χ^2 value for each test and evaluating each result using the obtained significance levels (P-value > 0.05). When a probability value of greater than 0.05 is obtained for both overall chi square tests, one fails to reject the null designating that the model sufficiently reproduces the sample variance/covariance matrix.

A Goodness of Fit (GFI) index and the Root Mean Square Approximation (RMSEA) are next applied to test data to model fit. The GFI predicts the percentage of the variance/covariance in the sample variance/covariance (S) that is reproduced by the predicted variance/covariance matrix (Σ), given the one factor model. A GFI that is greater than or equal to 0.90 is typically accepted as displaying good fit and this index works better with parsimonious models having few parameters. The RMSEA is a "badness of fit" index that employs a non-central chi square distribution and produces a confidence interval. RMSEA takes into account errors in approximation by asking, "How well would a model, with unknown but optimally chosen parameters values, fit the population variance/covariance matrix if it were available? "(see Browne and Cudek, 1993). A noncentrality parameter is estimated from a an approximate noncentral chi-square distribution, designated as δ. The value of δ increases as the null becomes more false. Using Cheung and Resvold's (2002) recommendations, reasonable errors in approximation would be equal to or less than 0.05, but these recommendations are considered conservative. Others suggest that RMSEA values between 0.05 and 0.08 are moderately acceptable (MacCallum, Browne and Sugawara, 1996). A confidence band of 90% is presented and there is a greater potential for misfit at the higher end of the scale.

The multigroup strategy utilizing a hierarchical strategy is next employed to test the null hypotheses (H_0) that $\Sigma_1 = \Sigma_2$, where Σ is the population variance-covariance matrix for the two groups. Rejection of the null thus provides evidence of the nonequivalence of the factor structure attributed to each

group's subscores. Acceptance of the null hypothesis provides evidence that the factor structure produced by both models possess structural or "configural" equivalence (see Wu, Li, & Zumbo, 2007).

Subsequent chi-square difference tests are next performed to test the increasingly more restrictive assumptions associated with multigroup equivalence tests that were previously discussed. Statistical Tests are applied by constraining parameters in a restrictively increasing fashion and by observing changes in the chi-square likelihood test (χ^2 **).** Each time one tests the differences in the likelihood chi-square statistics ($\Delta\chi^2$) in the augmented model minus the more compact model. A p-value is computed for each change in the chi-square statistic with respect to the appropriate degrees of freedom. A good fitting comparative model will have a $\Delta\chi^2$ p-value greater than 0.05, demonstrating that the differences in the likelihood square statistic are negligible.

The Comparative Fit Index (CFI) provides additional quality information about fit when evaluating two competing models. Values of the CFI range from 0 to 1 and are derived from a comparison made between the hypothesized or implied model and an independence model. The CFI is not sensitive to large sample sizes and is derived by a comparison of the hypothesized model to an independence model – a model in which the variables are assumed to be uncorrelated. In short, the CFI represents a ratio of the discrepancy of the implied model to the independence model. Again using Cheung and Resvold's (2002) recommendations, the change in CFI (ΔCFI) should be less than or equal to -0.01, which is a more robust test than the commonly used standard of accepting model fit for any CFI greater than 0.95. However, this recommendation respresents a "gold" standard that is more readily attained in large samples with well specified models that have very precise parameter estimates. RMSEA fit statistics are finally applied to note how well these more restrictively fitting models exhibit "bad" fit to the hypothesized model.

Single Group Results

Because there were seven grade levels studied, a single group analysis fit a unidimensional model to all 7 samples from the standard test and all 7 samples from the side-by-side test. The grade level results of each separate analysis for each single group are presented in Tables 1 through 7. A case can be made that each group's scores fit a unidimensional model, but the side-by-side scores generally fit a unidimensional model better than scores produced by the standard administration. The overall chi-squares for the side-by-side scores suggests that a unidimensional model fits the data at every grade level, while the standard test scores strictly fit a unidimensional model at grades 4, 6, 7, and 10. Test scores analyzed at grades 3, 5, and 8 come close to fitting, but the geometry and algebra strands appears to produce a larger residual for these samples. The goodness of fit index is over 0.99 each time it was employed, and the GFI close to 1 predicts that the fit the sample to the modeled variance/covariance is good. The RMSEA's fit the data in all cases but two models – the grades 3 and 5 standard or regular administration. However, even in these two cases, the RMSEA came within

Single Group Analysis Model Fit

Table 1

Model Fit	χ^2	df	P-Value	RMSEA	GFI
Side-by-Side Mathematics	1.218	5	.943	0.000 (0.00, 0.009)	0.999
Mathematics	15.11	5	.001	0.052 (0.023, 0.083)	.992

N=750 Grade 3

Single Group Analysis Model Fit

Table 2

Model Fit	χ^2	df	P-Value	RMSEA	GFI
Side-by-Side Mathematics	5.548	5	.353	0.012 (0.00, 0.053)	0.997
Mathematics	5.67	5	.345	0.013 (0.00, 0.054)	0.997

N=750 Grade 4

Single Group Analysis Model Fit

Table 3

Model Fit	χ^2	df	P-Value	RMSEA	GFI
Side-by-Side Mathematics	0.962	5	.996	0.00 (0.000, 0.000)	0.999
Mathematics	14.753	5	.001	0.051 (0.022, 0.082)	0.993

N=750 Grade 5

Single Group Analysis Model Fit

Table 4

Model Fit	χ^2	df	P-Value	RMSEA	GFI
Side-by-Side Mathematics	3.33	5	.649	0.000 (0.000, 0.041)	0.998
Mathematics	5.943	5	.312	0.016 (0.000, 0.055)	0.997

N=750 Grade 6

Single Group Analysis Model Fit

Table 5

Model Fit	χ^2	df	P-Value	RMSEA	GFI
Side-by-Side Mathematics	5.961	5	.310	0.016 (0.00, 0.055)	0.997
Mathematics	5.17	5	.395	0.007 (0.00, 0.052)	0.997

N=750 Grade 7

Single Group Analysis Model Fit

Table 6

Model Fit	χ^2	df	P-Value	RMSEA	GFI
Side-by-Side Mathematics	4.63	5	.463	0.00 (0.00, 0.049)	0.998
Mathematics	13.147	5	.022	0.047 (0.016, 0.078)	0.993

N=750 Grade 8

Single Group Analysis Model Fit

Table 7

Model Fit	χ^2	df	P-Value	RMSEA	GFI
Side-by-Side Mathematics	5.279	5	.383	0.009 (0.00, 0.052)	0.997
Mathematics	10.834	5	.055	0.039 (0.000, 0.072)	0.994

N=750 Grade 10

thousandths of "closely" fitting the model, and these values are still considered a "fair" fit by some experts (see MacCallum, Browne and Sugawara, 1996). The RMSEA further suggested that much of the misfit occurred at the higher end of the scale, where the .90 confidence interval often included values greater than 0.05. This pattern of results may be explained by larger numbers of students taking the more difficult questions in the standard pools – particularly for content associated with the strands algebraic relations and geometry. Patterns of misfit in the residuals were more likely observed in the algebra and geometry subscales compared to other subscales. The more challenging items used to score these subscales are more likely to be administered to the higher ability populations of test takers in the standard pools. Success on these more complex items may be relatively more influenced by a small second factor not measured by the other subscores -- e.g., spatial or abstract reasoning.

Multigroup Analysis Results

Using CFA, one evaluates measurement equivalence using a hierarchical procedure that compares a number of increasingly more restrictive models using a likelihood ratio goodness of fit difference test along with a comparative fit index. Since most models are either slightly misspecified or do not account for all measurement error, when sample sizes are large, a nonsignificant chi-square test is rarely obtained. Because a researcher's model is so frequently rejected in large samples, other measures of fit have been developed to assess the congruence of model fit to the data. A better fitting model does not always mean a more correct model.

The Rasch model attempts to estimate person and item points of estimation along a line using joint maximum likelihood. When constructing comparable measures, one attempts to isolate the trait of interest and build an additive measure that is invariant across language groups. CFA employs alternative forms of maximum likelihood estimation to produce parameter estimates describing the relationship between scores of the two samples shown in Tables 8 through 14 to follow. A multigroup strategy is next employed to test the null hypotheses (H_0) that $\Sigma_1 = \Sigma_2$, where Σ is the population variance-covariance matrix for the two groups. Rejection of the null thus provides evidence of the nonequivalence of the factor structure of the two sets of subscores produced by these administration groups at grade 3. For example, Table 8 on page 9 provides a chi-square value of 17.854 for the test of configural invariance. This result fails to reject the null that the factor structure for the two groups is equal. This result suggests that the side-by-side and standard tests measure the same conceptual framework. Tests of configural invariance demonstrate little difference in the factor structure of the two models.

In Table 8 on page 9, the weak invariance tests of equal factor loadings are next displayed for the standard and adapted or side-by-side tests in mathematics. Statistical Tests are applied by constraining parameters in a restrictively increasing fashion and by observing changes in the chi-square likelihood test (χ^2). Constraining the factor loadings implies that they are equal both within and between language groups, meaning that the underlying metric in the latent variable is similar for both the

standard and side-by-side assessments at grade 3. To apply the tests of weak invariance, one examines the chi-square difference tests ($\Delta\chi^2$ in the Tables). Each time one tests the differences in the likelihood chi-square statistics ($\Delta\chi^2$) in the augmented model minus the more compact model. A p-value is computed for each change in the chi-square statistic. A good fitting comparative model will have a $\Delta\chi^2$ p-value greater than 0.05, demonstrating that the differences in the likelihood square statistic are negligible. This result suggests that the latent scale's metric is "invariant" across the different administration conditions.

Changes in the comparative fit indices (ΔCFI) are appreciably small when subtracting the CFI of the configural model from the CFI of the weak invariance model. In addition, the RMSEA provides additional support that the restrictions on the factor loadings are having little "bad effect" on model fit. A change in the CFI (ΔCFI) ranging between 0 than -0.01 provides alternative evidence of little or no differences attributed to the factor loadings, while a small RMSEA of .015 with the 90% confidence interval of 0.00 and 0.03 demonstrates that the null hypothesis of close approximation is not rejected.

Tests of strong and strict invariance are similarly applied in Table 8. To test for strong invariance, changes in the chi-square value ($\Delta\chi^2$ **)** and the comparative fit index (ΔCFI) is similarly evaluated after constraining the intercepts to be equal ($\tau_{1i} = \tau_{2i} \ldots \tau_{1j} = \tau_{2j}$). Likewise, to test the strict variance assumptions ($\theta^1_\delta = \theta^2_\delta$), changes in the chi-square values and the comparative fit index are evaluated after constraining the unique variances associated with each manifest variable to be equal. Neither strong nor strict invariance are not observed in Table 8. The regression intercepts or subscale means linking the observed variables (y_{1i} to y_{2j} in equation 1) to the latent construct, ξ , are very different across groups, suggesting that the ability distributions of the two groups are very different. Moreover, the unique variances are not evaluated since the strong invariance model does fit the data.

The results observed in Tables 8 are similar in Tables 9 and 10 on page 9 and Table 11 on page 10, suggesting that the factored subscales possess both structural and metric equivalence across the conditions of the assessment. However, Tables 12 through 14 on pages 10 and 11 present some small aberrations that need more clarification. As one moves up grade levels into middle school, the mathematics items become increasingly challenging to more and more students. One begins to observe that more and more students are challenged to perform more complex mathematics that demands abstract thinking – particularly when the more difficult items are classified as assessing algebraic relations and geometry content. Moreover, by blue print specification and adaptively applied content constraints, there are proportionately more of these challenging items presented at the higher grade levels. As the adaptive engine adjusts to any increasing mathematics ability, it begins to select items that demand more abstract thinking. These differences in the algebraic and geometry items may be introducing secondary factors not measured by a unidimensional, latent scale, and these differences are believed to be more at the top of the scale.

Multiple Group Analysis Model Fit

Table 8

Model Fit	χ^2	df	P-Value	RMSEA	CFI	$\Delta\chi^2$	P-Value $\Delta\chi^2$	ΔCFI
Configural Invariance	17.8524	11	0.085	0.021 (0.00, 0.038)	0.998	--	--	--
Weak Invariance	19.8883	15	0.176	0.015 (0.00, 0.030)	0.998	1.7832	0.87826	0.00
Strong Invariance	112.57	20	0.00	0.056 (0.046, 0.066)	0.970	92.682	0.00	-0.028
Strict Invariance	124.646	25	0.00	0.052 (0.043, 0.061)	0.968	12.076	0.00	-.002

N=750 per group tested Grade 3

Multiple Group Analysis Model Fit

Table 9

Model Fit	χ^2	df	P-Value	RMSEA	CFI	$\Delta\chi^2$	P-Value $\Delta\chi^2$	ΔCFI
Configural Invariance	11.78	11	0.38	0.007 (0.000, 0.028)	1.00	--		--
Weak Invariance	13.49	15	0.564	0.000 (0.000, 0.022)	1.00	1.71	0.88764	0.00
Strong Invariance	134.59	20	0.00	0.062 (0.052, 0.072)	0.959	121.1	0.00	-0.041
Strict Invariance	146.28	25	0.00	0.057 (0.048, 0.066)	0.956	11.69	0.00	-0.003

N=750 per group tested Grade 4

Multiple Group Analysis Model Fit

Table 10

Model Fit	χ^2	df	P-Value	RMSEA	CFI	$\Delta\chi^2$	P-Value $\Delta\chi^2$	ΔCFI
Configural Invariance	21.82	11	0.026	0.026 (0.009, 0.041)	0.996	--		--
Weak Invariance	27.858	15	0.022	0.028 (0.015, 0.041)	0.995	6.08	0.2985	-0.001
Strong Invariance	106.54	20	0.000	0.054 (0.044, 0.064)	0.967	78.682	0.00	-0.028
Strict Invariance	110.56	25	0.00	0.048 (0.039, 0.057)	0.967	4.02	0.00	0.00

N=750 per group tested Grade 5

Multiple Group Analysis Model Fit

Table 11

Model Fit	χ^2	df	P-Value	RMSEA	CFI	$\Delta\chi^2$	P-Value $\Delta\chi^2$	ΔCFI
Configural Invariance	10.745	11	0.465	0.00 (0.00, 0.027)	1.00	--		--
Weak Invariance	13.83	15	0.539	0.00 (0.000, 0.023)	1.00	3.085	0.6869	0.00
Strong Invariance	82.67	20	0.00	0.046 (0.036, 0.056)	0.977	68.84	0.00	-0.023
Strict Invariance	85.62	25	0.00	0.041 (0.031, 0.05)	0.978	2.95	0.00	-0.001

N=750 per group tested Grade 6

Multiple Group Analysis Model Fit

Table 12

Model Fit	χ^2	df	P-Value	RMSEA	CFI	$\Delta\chi^2$	P-Value $\Delta\chi^2$	ΔCFI
Configural Invariance	12.126	11	0.354	0.008 (0.00, 0.029)	0.999	--		--
Weak Invariance	48.119	15	0.00	0.038 (0.027, 0.051)	0.985	35.992	0.00	-0.014
Strong Invariance	232.35	20	0.00	0.085 (0.076, 0.095)	0.903	184.23	0.00	-0.082
Strict Invariance	237.57	25	0.00	0.075 (0.067, 0.084)	0.903	5.217	0.00	0.000

N=750 per group tested Grade 7

Multiple Group Analysis Model Fit

Table 13

Model Fit	χ^2	df	P-Value	RMSEA	CFI	$\Delta\chi^2$	P-Value $\Delta\chi^2$	ΔCFI
Configural Invariance	22.63	11	0.02	0.023 (0.003, 0.039)	0.996	--		--
Weak Invariance	34.706	15	0.003	0.03 (0.017, 0.043)	0.992	12.076	0.0338	-0.004
Strong Invariance	211.43	20	0.00	0.081 (0.071, 0.091)	0.926	176.72	0.00	-0.066
Strict Invariance	215.28	25	0.00	0.071 (0.063, 0.080)	0.926	3.846	0.00	0.00

N=750 per group tested Grade 8

Multiple Group Analysis Model Fit

Table 14

Model Fit	χ^2	**df**	**P-Value**	**RMSEA**	**CFI**	$\Delta\chi^2$	**P-Value** $\Delta\chi^2$	Δ**CFI**
Configural Invariance	16.801	11	0.114	0.019 (0.000, 0.036)	0.996	--		--
Weak Invariance	58.582	15	0.00	0.044 (0.033, 0.056)	0.971	41.78	0.00	-0.025
Strong Invariance	340.93	20	0.00	0.103 (0.094, 0.113)	0.784	282.35	0.00	-0.187
Strict Invariance	346.05	25	0.00	0.111 (0.084, 0.101)	0.784	5.12	0.00	0.00

N=750 per group tested Grade 10

In Tables 12 through 14, the structural results suggested by the configural model demonstrates that the model fits the data, providing empirical evidence for the structural equivalence of the conceptual framework being used. But the metric equivalence of the latent variable is relatively less invariant at higher grade levels. For example, when examining the weak invariance tests, the chi-square difference tests ($\Delta\chi^2$ **)** and the change in the comparative fit index (Δ**CFI)** is slightly larger for the higher grades (see Tables 12 through 14) when compared to the lower grades (see Tables 8 through 11). All three chi-square tests associated with the weak invariance tests suggest that the measurement equivalent assumption does not hold, while the comparative fit indices associated with each test range from highly acceptable to moderately acceptable fit. However, these differences are borderline and small, and most experts would quarrel about whether they are meaningful. The RMSEA for each test of weak invariance or scale equivalence indicate superior fit at all grade levels, with confidence values only slightly exceeding good ft at the higher end.

As in the example presented in Table 8, the strict invariance conditions of measurement equivalence are not met for all of Tables 9 through 14. This result implies that there is no consistency in the scaling units of the calibrated latent scale. In fact, the differences in the τ regression intercepts are a product of the higher average scores earned by the higher achieving groups taking the standard assessment. There is little chance that these scalar differences in the latent metric can be remedied without subsequent increases in achievement from the lower achieving group taking the side-by-side accommodation.

Conclusions

The establishment of measurement invariance across accommodated conditions of a test's administration is a logical prerequisite for establishing the comparability of the scale. This study attempts a hierarchical analysis to answer the following two research questions regarding the comparability of the latent scale: Will different conditions for administering and scoring mathematics achievement yield an invariant factor structure measuring a single attribute or construct? And, relatedly, will the metric of latent scores be invariant across administrative conditions?

In all the cases studied, both in the single group and multiple groups analyses, the factorial structure of the latent variable fit the unidimensional model and the conceptual equivalence of the underlying theoretical variable was comparable under both administrative conditions. In both single group and multiple group analyses, the precision of the omnibus chi square test was sometimes affected by larger residuals observed for the algebraic and geometric strands in the standard test, but the fit indices always confirmed a properly fitting model.

Likewise, when scale equivalence was evaluated using a weak invariance test previously described, chi-square difference tests rejected the null hypothesis that state that the slopes of the factor loadings were invariant across administrative conditions for grades 3 through 6. The chi-square tests for grades 7 through 10 suggested that model no longer fit, but the CFI and the RMSEA's suggested otherwise. In large samples, it is well understood that chi-square difference tests are more likely to demonstrate significant differences. For this reason, most experts suggest using the fit indices to supplement the decisions during this hierarchical analysis.

References

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

Chueng, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing MI. *Structural Equation Modeling*, 9, 235-255.

Hambleton, R.K. (1994). Guidelines for adapting educational or psychological tests: A progress report. *Bulletin of the International Test Commission, 10*(3), 229-244.

Joreskog, K. G., and Sorbom, D. (1993). *LISREL8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Erlbaum.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130-149.

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

Vandenberg, R.J. & Lance, C.E. (2000). A review and synthesis of the MI literature: Suggestions, practices, recommendations for organizational research. *Organizational Research Methods*, 3, 4-69.

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

Wu, A. D., Li, Z. & Zumbo, B. D. (2007). Decoding the meaning of factor invariance and updating the practice of multi-group confirmatory factor analysis. *Practical Assessment, Research and Evaluation*, *12* (3), 1-26.