Score Comparability of the Oregon Mathematics Assessment between English and Dual language (English-Spanish) Forms

Oregon Department of Education

Abstract

Multi-group confirmatory factor analyses (CFA) were used to evaluate whether construct invariance could be established between the English only and dual language (English-Spanish) versions of the Oregon Statewide Mathematics knowledge and skills tests. The evaluation incorporated a rigorous evaluation by constraining factor loadings, means, and residual variances to be equal across both groups. This methodology provides the strongest possible evaluation of score comparability. Results of the analyses indicate there is strong evidence to support a determination of “Strict Invariance” between the English and dual language forms for grades 6-8 and 10 and evidence for invariance of most model parameters in grades 3-5. There was marginal evidence for differences in strand score means between the two forms in grades 3-5. These differences may well be due to construct relevant differences between form groups including sample size, local testing decisions, demographics, level of English proficiency, opportunity to learn and numerous other concomitant variables. Given the non-random assignment to group, results of the analyses suggest a high degree of score comparability across forms.

The state of Oregon offers mathematics multiple-choice statewide assessments in grades 3-8 and 10 in English and English-Spanish dual language versions. The intent of providing the alternate language forms is to reduce the likelihood that the results of the mathematics assessment are a description of students' proficiency in English rather than their ability to demonstrate mastery of the state mathematics content standards.

However, for large scale assessments, we want to ensure that the accountability designations are derived from comparable scores. That is, we want to ensure that the construct we are using to evaluate school performance is equivalent regardless of the method or mode that was used to assess the construct.

Optimally, we would want to create an experimental design in which students are randomly allocated to versions of the forms. However, given that the assessments are used to evaluate student instructional needs and used for high stakes accountability, there would be political, legal and ethical implications of such a study. Alternatively, a version of the form that isn't high stakes could be used. However, there would questions regarding student motivation and the equivalence of the newly created form that would compromise any inferences one might make about the results of such a study.

Therefore the purpose of this analysis is to establish whether the hypothesis of form equivalence among the language forms is demonstrated by the empirical data.

Sample

The highest score for all Oregon students who took a standard administration multiple-choice Mathematics test via Oregon's online assessment in 2005-06 were included in the model. The students were grouped based on whether they were assessed on an English or a dual language version of the assessment. The allocation of the groups is not random and is based on a local decision regarding which form will best allow the student's ability to demonstrate his or her mastery of the content. Across all grades, a total of 264,938 students taking the English forms and 5,142 students taking the dual language form were included in the analysis. For students taking the 10^{th} grade test, the data were restricted to only those students who took the test as a 10^{th} grader in 2005-06. Students are able to take the 10^{th} grade test as early as 9^{th} grade and as late as 12^{th} grade, but adding this additional non-random effect would likely reduce the clarity of evaluation. The N, means, standard deviations and correlations of the data are provided in Table 1 (v1-v5 are the strands of Calculations and Estimations, Measurement, Statistics and Probability, Algebraic Relationships, and Geometry respectively).

Table 1. Correlations and Descriptive Statistics for students taking English or Spanish forms

		V1	V2	V3	V4	V5
Grade 3 English						
MEAN		211.111	210.560	210.828	209.971	212.431
STDDEV		12.695	12.437	15.100	13.057	12.735
N		37245	37245	37245	37245	37245
CORR	V1	1.000	0.498	0.453	0.535	0.414
CORR	V2	0.498	1.000	0.383	0.458	0.376
CORR	V3	0.453	0.383	1.000	0.417	0.332
CORR	V4	0.535	0.458	0.417	1.000	0.396
CORR	V5	0.414	0.376	0.332	0.396	1.000
Grade 4 English						
MEAN		221.331	217.945	219.552	216.591	219.544
STDDEV		13.981	13.057	14.178	12.775	12.938
N		37951	37951	37951	37951	37951
CORR	V1	1.000	0.535	0.488	0.517	0.464
CORR	V2	0.535	1.000	0.446	0.459	0.449
CORR	V3	0.488	0.446	1.000	0.432	0.434
CORR	V4	0.517	0.459	0.432	1.000	0.421
CORR	V5	0.464	0.449	0.434	0.421	1.000
Grade 5 English						
MEAN		222.420	223.891	223.485	222.635	223.720
STDDEV		12.366	12.817	13.916	13.130	12.162
N		38309	38309	38309	38309	38309
CORR	V1	1.000	0.445	0.458	0.489	0.414
CORR	V2	0.445	1.000	0.431	0.449	0.423
CORR	V3	0.458	0.431	1.000	0.473	0.408
CORR	V4	0.489	0.449	0.473	1.000	0.422
CORR	V5	0.414	0.423	0.408	0.422	1.000
Grade 6 English						
MEAN		224.565	225.745	226.565	226.685	226.807
STDDEV		14.436	12.532	14.458	13.732	12.933
N		38139	38139	38139	38139	38139
CORR	V1	1.000	0.451	0.533	0.515	0.415
CORR	V2	0.451	1.000	0.510	0.498	0.441
CORR	V3	0.533	0.510	1.000	0.584	0.487
CORR	V4	0.515	0.498	0.584	1.000	0.504
CORR	V5	0.415	0.441	0.487	0.504	1.000
Grade 7 English						
MEAN		231.487	231.216	233.082	232.017	231.145
STDDEV		15.603	15.997	14.133	13.427	13.421
N		39300	39300	39300	39300	39300
CORR	V1	1	0.41062	0.49064	0.5678	0.46109
CORR	V2	0.411	1.000	0.474	0.514	0.462
CORR	V3	0.491	0.474	1.000	0.638	0.514
CORR	V4	0.568	0.514	0.638	1.000	0.585
CORR	V5	0.461	0.462	0.514	0.585	1.000

		V1	V2	V3	V4	V5
Grade 3 Spanish						
MEAN		202.999	202.432	200.959	203.459	205.633
STDDEV		10.767	11.153	11.890	11.570	11.381
N		1440	1440	1440	1440	1440
CORR	V1	1.000	0.480	0.474	0.515	0.408
CORR	V2	0.480	1.000	0.439	0.458	0.402
CORR	V3	0.474	0.439	1.000	0.469	0.409
CORR	V4	0.515	0.458	0.469	1.000	0.401
CORR	V5	0.408	0.402	0.409	0.401	1.000
Grade 4 Spanish						
MEAN		208.806	208.899	210.079	210.259	209.455
STDDEV		11.371	11.096	10.872	11.169	10.647
N		986	986	986	986	986
CORR	V1	1.000	0.543	0.451	0.545	0.480
CORR	V2	0.543	1.000	0.450	0.475	0.456
CORR	V3	0.451	0.450	1.000	0.443	0.390
CORR	V4	0.545	0.475	0.443	1.000	0.455
CORR	V5	0.480	0.456	0.390	0.455	1.000
Grade 5 Spanish						
MEAN		214.940	214.226	212.431	215.930	217.648
STDDEV		12.103	11.722	12.111	11.597	10.235
N		937	937	937	937	937
CORR	V1	1.000	0.501	0.499	0.529	0.438
CORR	V2	0.501	1.000	0.500	0.529	0.444
CORR	V3	0.499	0.500	1.000	0.502	0.391
CORR	V4	0.529	0.529	0.502	1.000	0.456
CORR	V5	0.438	0.444	0.391	0.456	1.000
Grade 6 Spanish						
MEAN		216.319	218.091	215.545	216.955	219.656
STDDEV		11.341	10.266	10.850	10.530	10.899
N		679	679	679	679	679
CORR	V1	1.000	0.376	0.292	0.413	0.326
CORR	V2	0.376	1.000	0.370	0.400	0.405
CORR	V3	0.292	0.370	1.000	0.380	0.447
CORR	V4	0.413	0.400	0.380	1.000	0.430
CORR	V5	0.326	0.405	0.447	0.430	1.000
Grade 7 Spanish						
MEAN		218.999	224.540	225.647	221.712	225.986
STDDEV		11.089	8.755	9.921	7.473	8.515
N		339	339	339	339	339
CORR	V1	1.000	0.299	0.324	0.310	0.388
CORR	V2	0.299	1.000	0.289	0.278	0.368
CORR	V3	0.324	0.289	1.000	0.353	0.382
CORR	V4	0.310	0.278	0.353	1.000	0.375
CORR	V5	0.388	0.368	0.382	0.375	1.000

Table 1. Correlations and Descriptive Statistics for students taking English or Spanish forms (cont.)

Grade 8 English						
MEAN		234.028	234.368	235.467	234.722	233.905
STDDEV		16.180	16.421	13.234	13.063	13.485
N		40954	40954	40954	40954	40954
CORR	V1	1.000	0.535	0.508	0.589	0.493
CORR	V2	0.535	1.000	0.541	0.602	0.517
CORR	V3	0.508	0.541	1.000	0.595	0.493
CORR	V4	0.589	0.602	0.595	1.000	0.573
CORR	V5	0.493	0.517	0.493	0.573	1.000
Grade 10 English						
MEAN		232.553	233.490	233.182	234.556	234.284
STDDEV		16.133	19.296	13.376	11.393	12.932
N		33040	33040	33040	33040	33040
CORR	V1	1.000	0.257	0.379	0.461	0.435
CORR	V2	0.257	1.000	0.299	0.338	0.359
CORR	V3	0.379	0.299	1.000	0.550	0.523
CORR	V4	0.461	0.338	0.550	1.000	0.619
CORR	V5	0.435	0.359	0.523	0.619	1.000

Grade 8 Spanish						
MEAN		222.955	223.200	223.742	224.662	224.513
STDDEV		10.945	12.786	7.702	8.592	9.864
N		330	330	330	330	330
CORR	V1	1.000	0.341	0.179	0.377	0.339
CORR	V2	0.341	1.000	0.298	0.313	0.342
CORR	V3	0.179	0.298	1.000	0.213	0.255
CORR	V4	0.377	0.313	0.213	1.000	0.315
CORR	V5	0.339	0.342	0.255	0.315	1.000
Grade 10 Spanish						
MEAN		226.604	228.541	223.864	227.552	226.470
STDDEV		10.453	11.289	7.402	5.733	7.890
N		431	431	431	431	431
CORR	V1	1.000	0.136	0.117	0.092	0.039
CORR	V2	0.136	1.000	0.149	0.054	0.039
CORR	V3	0.117	0.149	1.000	0.055	0.023
CORR	V4	0.092	0.054	0.055	1.000	0.055
CORR	V5	0.039	0.039	0.023	0.055	1.000

Method

We used a multi-group confirmatory factor analysis using AMOS 7.0 (Arbuckle, 2006). The analyses are conducted separately by grade level with students grouped according to whether their mastery of mathematics content was assessed via an English or Spanish-English side by side form. The multiple-group analysis is comprised only of an analysis of a single factor (i.e. mathematics) representing the 5 strands/traits of Calculations and Estimations, Measurement, Statistics and Probability, Algebraic Relationships, and Geometry (v1-v5 respectively as described in Figure 1.) For each grade-level, a nested hierarchy of invariance tests was conducted that sequentially constrained factor loadings, means, intercepts and residual variances to be equal between the two groups of forms.

As outlined by (Wu et al. , 2007) we progressively tested for configural invariance (i.e. unconstrained factor structure), weak invariance (i.e. factor loadings constrained to be equal), strong invariance (i.e. indicant means constrained to be equal) and finally strict invariance (i.e. error variances constrained to be equal).

Due to the complexity of the data, the analysis of the models must be multi-faceted and should include an evaluation of the overall fit of the model as defined by the levels of invariance as well as the progressive decrement in fit associated with the additional constraints imposed at each step of the nested hierarchy of invariance tests (Little, 1997).

Evaluation of Model Fit

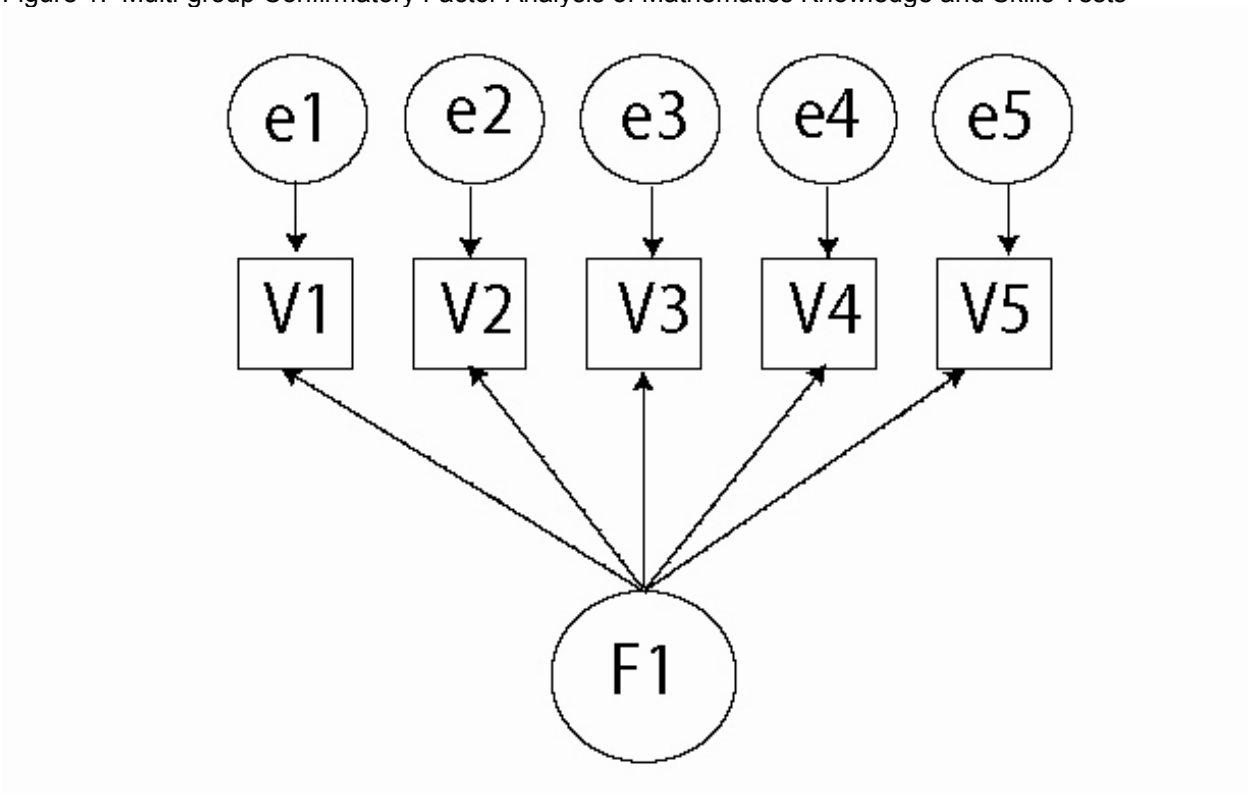
It is well documented that when used in the analysis of large samples, the chi-square goodness of fit statistic will detect non-material differences between hypothesized models and covariance structures. For this reason, indexes such as the CFI (Bentler, 1990) and RMSEA (Sreiger, 1989) can be used to evaluate model fit in lieu of only using the chi-square statistic. Commonly used rules of thumb are that approximate model fit is established when the CFI is ≥ .95 and RMSEA is ≤ 05 (Hu & Bentler, 1999)

Evaluation of progressive constraints

Because each of the models described by (Wu et al., 2007) are nested, we can determine whether additional constraints decrease the fit enough to off-set the increase in degrees of freedom by virtue of

the additional constraints. The criterion suggested by Chueng & Rensvold (2002) is that a difference of .02 in the CFI between the models indicates a substantive difference between groups and a lack of measurement invariance.

Figure 1. Multi-group Confirmatory Factor Analysis of Mathematics Knowledge and Skills Tests



Results

Evaluation of Model Fit

As expected given the sample sizes, the chi-square statistic is significant at each grade in each of the models. Results for each grade can be found in Tables 2, 3 and 4.

Based on the fit indices, the model can be interpreted as fitting well under each set of constraints. For the unconstrained model the CFI ranged from .998 to 1.000 for grades 3 – 8 and 10. The RMSEA ranged from .008 to .021 also indicating a good model fit for each grade. As expected given the sample sizes, the chi-square statistic is significant at each grade. Results for each grade can be found in Table 2. These results demonstrate that the configural model is the same for the standard English form and side by side form groups.

The model also fit well when the factor loadings were constrained (i.e. Weak Invariance). With these constraints imposed between the two groups, the CFI ranged from .989 to .999 and the RMSEA ranged from .010 to .032 (See Table 3).

Similarly the model fit well when the intercepts (i.e. means) were constrained to be equal (i.e. Strong Invariance) with the CFI ranging from .977 to .994 and RMSEA ranging from .026 to .040 and when the error variances (i.e. residuals) were constrained to be equal with the CFI ranging from .973 to .993 (see Table 4).

Evaluation of progressive constraints

The comparison of the relative fit of models yields a more complex picture of the data. For all grades, the difference in the CFI between the unconstrained and the "weak invariance" model is less than .01 suggesting that the additional constraints do not create a material reduction in model fit (see Table 2).

For the test of Strong Invariance, the results differ by grade. For grades 3-5 the difference in the CFI ranged from .016 to 022 suggesting a marginal reduction in model fit when the means are constrained to be equal. In contrast, grades 6 – 8 and 10 have differences that range from .006 to .008 and suggest that constraining the means to be equal does not materially reduce the model fit (see Table 3).

For the test of Strict Invariance (i.e. constraining variances to be equal), grades 6-8 and 10 do not show material reductions in fit. If we were to assume that the test of Strong Invariance was met for grades 3-5, we would also subsequently conclude that there was not a reduction in fit when constraining the variances to be equal (see Table 4).

Table 2. Model Fit for the Multiple Groups Confirmatory Factor Analysis Comparison of Unconstrained and Weak Invariance

Grade		Unconstrained (*df*=10)			Weak Invariance (*df*=15)			
	N	**X^2**	**CFI**	**RMSEA**	**X^2**	**CFI**	**RMSEA**	ΔCFI
3	38685	32.858	1.000	.008	78.115	.999	.010	0.001
4	38937	145.762	.998	.019	207.520	.997	.018	0.001
5	39246	97.265	.998	.015	119.097	.998	.013	0
6	38818	163.149	.998	.020	294.302	.996	.022	0.002
7	39639	185.969	.998	.021	372.830	.995	.025	0.003
8	41284	62.421	.999	.011	231.856	.997	.019	0.002
10	33471	82.128	.998	.015	524.681	.989	.032	0.009

Table 3 Model Fit for the Multiple Groups Confirmatory Factor Analysis Comparison of Weak and Strong Invariance

Grade		Weak Invariance	Strong Invariance(*df*=20)			
	N	**CFI**	**X^2**	**CFI**	**RMSEA**	ΔCFI
3	38685	.999	1142.172	.977	.038	0.022
4	38937	.997	1275.568	.978	.040	0.019
5	39246	.998	955.197	.982	.034	0.016
6	38818	.996	794.941	.988	.032	0.008
7	39639	.995	689.650	.991	.029	0.004
8	41284	.997	561.616	.994	.026	0.003
10	33471	.989	816.185	.983	.034	0.006

Table 4 Model Fit for the Multiple Groups Confirmatory Factor Analysis Comparison of Strong and Strict Invariance

Grade		Strong Invariance	Strict Invariance (*df*=25)			
	N	**CFI**	**X^2**	**CFI**	**RMSEA**	ΔCFI
3	38685	.977	1365.702	.973	.037	0.004
4	38937	.978	1470.823	.975	.039	0.003
5	39246	.982	1061.877	.980	.032	0.002
6	38818	.988	869.737	.987	.029	0.001
7	39639	.991	880.900	.988	.029	0.003
8	41284	.994	615.174	.993	.024	0.001
10	33471	.983	1057.603	.978	.035	0.005

Evaluation of Parameters

For brevity, we display only the parameters for the fully constrained (i.e. strict invariance) model. Though we may make different inferences about this model given results of the invariance evaluation, the model fit indices would suggest it is still appropriate to review the parameters of the fully constrained model (see Table 5).

Table 5. Parameter estimates for the Confirmatory Factor Analysis –Strict Invariance

		3		4		5	
	Parameter	Effect	t-value	Effect	Effect	Effect	t-value
Factor Loadings							
	V1,F1	9.673	156.661	10.714	160.673	8.566	139.706
	V2,F1	8.267	131.971	9.195	144.682	8.496	132.137
	V3,F1	9.100	117.604	9.286	132.365	9.408	135.425
	V4,F1	9.227	142.951	8.577	136.592	9.229	142.851
	V5,F1	7.216	109.175	8.320	128.883	7.507	121.860
Error Variances							
	E1,E1	68.257	91.645	82.849	96.282	80.689	108.075
	E2,E2	87.568	112.213	86.749	109.143	93.624	113.035
	E3,E3	145.455	119.675	114.882	116.295	106.868	110.986
	E4,E4	85.488	104.496	89.635	114.062	87.375	105.745
	E5,E5	110.544	123.098	99.288	117.986	91.390	118.558
Means							
	V1	210.809	3259.311	221.013	3102.042	222.242	3546.961
	V2	210.257	3311.788	217.716	3282.406	223.660	3440.902
	V3	210.460	2739.711	219.312	3051.495	223.221	3163.680
	V4	209.729	3157.891	216.430	3342.929	222.475	3355.211
	V5	212.178	3272.448	219.288	3333.203	223.575	3643.881
Factor Variance							
	F1,F1	1.0	N/A	1.0	N/A	1.0	N/A

Table 5 Parameter estimates for the Confirmatory Factor Analysis (cont.)

		6		7		8		10	
	Parameter	Effect	t-value	Effect	t-value	Effect	t-value	Effect	t-value
Factor Loadings									
	V1,F1	9.781	139.970	10.398	140.136	11.570	157.743	9.125	103.441
	V2,F1	8.355	136.952	9.983	129.412	12.199	165.966	8.492	77.657
	V3,F1	11.155	165.621	10.585	163.786	9.556	159.652	9.093	129.395
	V4,F1	10.492	163.575	11.317	192.623	10.729	191.023	9.114	159.041
	V5,F1	8.350	131.517	9.341	148.748	9.361	151.551	9.982	151.853
Error Variances									
	E1,E1	112.481	116.011	135.613	121.806	127.745	119.312	175.511	116.450
	E2,E2	87.344	117.451	155.099	125.604	120.965	114.927	297.363	122.841
	E3,E3	85.084	98.777	87.284	109.452	83.971	118.364	95.745	105.326
	E4,E4	78.747	100.553	52.044	81.815	55.561	95.223	46.116	78.166
	E5,E5	97.575	119.825	92.167	118.081	94.231	122.133	67.012	86.575
Means									
	V1	224.421	3064.635	231.380	2950.661	233.939	2938.650	232.476	2643.872
	V2	225.611	3545.664	231.159	2883.384	234.279	2898.043	233.426	2221.654
	V3	226.372	3081.243	233.018	3285.909	235.374	3612.094	233.062	3192.025
	V4	226.514	3247.698	231.929	3440.590	234.642	3649.205	234.466	3774.032
	V5	226.682	3452.936	231.101	3434.921	233.830	3523.088	234.183	3318.833
Factor Variance									
	F1,F1	1.0	N/A	1.0	N/A	1.0	N/A	1.0	N/A

Discussion

A confirmatory factor analysis of the mathematics construct on the Oregon statewide assessment was compared for different grade level groups and for two forms of the assessment. The primary purpose of the analyses was to determine the equivalence and comparability of the standard English form and the dual language form. Models were evaluated for goodness of fit and also relative fit. The CFA models fit well for all groups. The nested invariance tests showed that the forms were comparable for almost all comparisons. Even the model with the most constraints (i.e. Strict Invariance) fit very well for all grades. Similarly, each grade demonstrated at least Weak Invariance between the English and English Side by Side forms. In terms of relative fit, we demonstrated Strict Invariance for grades 6- 8 and 10 but there was evidence that form differences in strand score means may be present for grades 3-5.

Despite the relatively small change in the fit indices, there is a noticeable increase in the chi-square when the intercepts (i.e. strand score means) are constrained to be equal across the two groups. This is not surprising and is evident in the raw data. Students who use the dual language form tend to have lower scale scores than students who take the English form. While this may be attributable to demographics and SES, it also may be a function of opportunity to learn or a number of other concomitant variables. For several of these possible factors, the assessment should reflect differences in the means. The lack of differences observed in the fit indices for the models in the majority of the comparisons in this study may be due to the increased heterogeneity of the population's opportunity to learn and the subject matter such that the form itself is not as strong of a predictor of performance as compared to the lower grades. Alternatively, in these later grades, constraining the means to be equal may mitigate differences in the residuals that might otherwise be reducing the fit of the model.

However, without random assignment to form, it should not be expected that means will be equivalent. The present study provides a strong demonstration that the characteristics of the assessment instruments

are comparable in terms of structural configuration, measurement relations between strand scores and the construct of interest. Certainly, one can argue that with at least the equivalence of the factor structure, the Spanish form of the assessment will be at least an equal (if not more) valid representation of these students' knowledge and skills than an assessment delivered in a language in which they are not proficient (i.e. English).

While results of these analyses suggest strong comparability of scores across forms, further research regarding form comparability is still warranted and planned. . In addition, we will investigate options for conducting an experimental design in future research wherein students can be randomly assigned to specific forms in a manner that does not negatively impact a school's accountability results but also doesn't create different motivation conditions. Finally, we will refine this analysis to partition the groups according to students presented with the same items. This analysis may better identify any sources of discrepancies.

REFERENCES

Arbuckle, J. L. (2006). Amos (Version 7.0) [Computer Program]. Chicago: SPSS.

Bentler, P.M. (1990) Comparative fit indices in structural models, *Psychological Bulletin*, *107*. 238-246

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Structural Equation Modeling, 6*(1), 1-54.

Little, T.D. (1997). Mean and covariance, structure (MACS) analyses of cross-cultural data: practical and theoretical issues. *Multivariate Behavioral Research, 32* (1), 53-76

Steiger, J. H. (1989). EzPATH: Causal modeling. Evanston, IL: SYSTAT

Wu, A. D., Zhen L. & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12* (3), 1-25.