Construct Validity and Measurement Invariance of Computerized Adaptive Testing: Application to Measures of Academic Progress (MAP) Using Confirmatory Factor Analysis

Shudong Wang
NWEA

Marty McCall
Smarter Balance Assessment Consortium

Hong Jiao
University of Maryland

Gregg Harris
NWEA

Paper presented at the annual meeting of the American Educational Research Association (AERA). April 12-16, 2012, Vancouver, British Columbia, Canada.

Send correspondence to:
Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97206
Shudong.Wang@NWEA.org

Abstract

The purposes of this study are twofold. First, to investigate the construct or factorial structure of a set of Reading and Mathematics computerized adaptive tests (CAT), *Measures of Academic Progress* (MAP), given in different states at different grades and academic terms. The second purpose is to investigate the invariance of test factorial structure across different grades, academic terms and states. Because of the uniqueness of CAT data (different student receive different items), traditional factor analysis based on fixed form data is no longer practically possible at the item level. This study illustrates how to overcome the difficulty of applying factor analysis in CAT data and study results provide evidences for valid interpretation MAP tests scores across grades at different academic terms for different states.

Construct Validity and Measurement Invariance of Computerized Adaptive Testing: Application to Measures of Academic Progress (MAP) Using Confirmatory Factor Analysis

Objectives

The purposes of this study are twofold: first, to investigate the construct or factorial structure of CAT MAP Reading and Mathematics tests at different grades, academic terms, and states; second, to investigate the invariance of test factorial structure across different grades, academic terms and states.

Perspectives

Recently, computerized adaptive testing (CAT) has been seen as a particularly effective method of measuring an individual student's status and growth over time in K-12 assessment (Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010). The major reason is that the CAT has advantages, such as a short test, immediate feedback on student scores, better reliability, and accuracy (Lord, 1977; Kingsbury & Weiss, 1983; Steinberg, & Thissen, 1990) over traditional paper-pencil tests. Its unique advantages in K-12 assessment include cost savings, multiple testing opportunities for formative and interim assessments, and better validity (Way, 2006).

Right now, Oregon, Delaware, and Idaho use CAT in their state assessments, and several other states (Georgia, Hawaii, Maryland, North Carolina, South Dakota, Utah, and Virginia) are in various stages of CAT development. As a matter of fact, one of the two consortia created as part of the Race to the Top initiative, the SMARTER Balanced Assessment Consortium (SBAC) consisting of over half of the states, is committed to a computerized adaptive model because it represents a unique opportunity to create a large-scale assessment system that provides maximally accurate achievement results for each student (Race to the Top Assessment Program, 2010).

Because high stakes decisions about students are based on state test results, these tests should be evaluated using professional testing principles, such as validity and reliability. Validity (and fairness), according to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999), is the most important consideration in test development and evaluation.

The MAP Reading and Mathematics tests, like most CATs, use a unidimensional item response model (IRT) model based on the premise that correlations among responses to test questions can be explained by a single underlying trait. Traits like reading and math are

obviously complex, representing many component skills and facts combined in specific ways. The claim of unidimensionality is that these components work together to manifest a coherent whole. Although tests are often structured around goal areas, this is done to provide adequate domain sampling rather than to measure different traits. While individuals may have strengths and weaknesses in goal areas on a unidimensional test, any systematic relationship among goals should be explained by the effect of the unitary latent trait on item responses.

Detecting dimensionality in adaptive assessments is tricky. Because of the uniqueness of CAT tests (different persons respond to different items), conducting factor analysis is more challenging for CAT data than for linear or fixed form data.

First of all, there are no common test forms, so data are very sparse. Observable (manifest) item variables differ across persons for both the overall test event and at the cluster (goal or subtest) level. So, although goal score variables are the same, the context differs; i.e. subtest scores are derived from different sets of items. One possible solution is to conduct confirmatory factor analysis (CFA) on the entire item bank, but the large amount of missing data (typically, the missing rate is above 90% if the ratio of test length to item bank size is 20) makes this unwieldy. For MAP Reading and Mathematics tests, typical missing rates are around 98% because the ratio of test length to item bank counts is around 50. The common imputation methods (Rubin, 1987) may statistically help the missing issue, but are difficult to execute.

The second issue is that the adaptive algorithm operates on the assumption of local independence (LI), thus restricting covariance among items. Items are selected to maximize information at the estimated latent trait level so that for dichotomous items the probability of a correct answer is about .5, responses are randomly distributed, and item covariance is low. Since the goal of factor analysis is to summarize patterns of correlation among observed variables, this restriction may lead to singularly uninformative factor analysis results for CATs. McCall and Hauser (2006) used Yen's (1984) Q statistic to get around the sparse data problem. The Q statistic operates on pairwise relationships between items and looks for covariance unexplained by the observed score. Because item selection is conditioned on the momentary achievement estimate, the range of ability is restricted, thus limiting variance and covariance. Values of the Q statistic were so small compared to those for fixed form tests that they were difficult to interpret.

One way to get around the sparse data problem is to conduct CFA at the item cluster level (goal or sub-content level). Since items in each CAT test event are balanced among goal areas based on content specifications, a reasonable option may be to assume that items within each goal area are content homogeneous across persons and that goal scores may be used as

observable variables. This is the method used here. If multiple traits based on goal areas explain item responses, this might show up as differential factor loading of the goal scores on the overall score. Furthermore, patterns of factor loading might differ with the different goal structures used in different states or among grades within the same state. For illustration purpose only, Figure 1 shows the factor model under the IRT assumption of LI. This is the model that is difficult to test with computerized adaptive test data. Figure 2 shows a factor model that uses testlet, goal score, or item clusters as observable variables. At the testlet level, this model still satisfies the LI assumption, but the LI assumption might or might not hold at the individual item level.

Methods

Data source or Participants

All data used in this study were collected from MAP Reading and Mathematics tests administered from Spring 2009 to Spring 2011 twice during the academic year. The MAP tests were used with grade 3 to grade 9 across 50 states. The data for this study focuses on 10 states (Colorado, Illinois, Indiana, Kansas, Kentucky, Michigan, Minnesota, South Carolina, Washington, and Wisconsin) that have the largest MAP sample sizes among the 50 states. Reading and Mathematics sample sizes for each state are presented in Table 1. Samples were collected for five academic terms: Spring 2009, Fall 2009, Spring 2010, Fall 2010, and Spring 2011. For each academic term, the samples contained results from five grades, with the grade range depending on the academic term. Table 2 and 3 (due to the limited space, can't list 10 state tables) list the frequency and percentages of samples across grades and terms for the Illinois Mathematics test and South Carolina Reading test. For each state, samples were randomly drawn from state records. Approximately 20% students for each state were selected under the constraints that student has to have scores for five academic terms and is in the grade range of 3 to 9 in the first term.

Instruments

The MAP tests of Reading and Mathematics for grades 3 to 9 were used in this study. MAP tests are computerized adaptive assessments that have been published by Northwest Evaluation Association (NWEA) since 2000. The purpose of MAP tests is to provide educators with information to inform teaching and learning in Reading, Mathematics, and Science (NWEA, 2011). For each state, the MAP tests are aligned to specific state content standards by assembling pools of items that address state content standards. Test algorithms survey the pools within goal

or strand areas to assure domain coverage. The marginal reliabilities of tests across 50 states and grades are consistently in the low to mid 0.90's (NWEA, 2011). Because items selected during the CAT test for each student are based on the student's provisional ability, these items have a limited range of difficulty for a given test taker. However, all items administered to each student have to satisfy the content requirements of the test to insure content validity and domain coverage. Table 4 lists test length (fixed length CAT) and numbers of goals (subtests) of both Reading and Mathematics tests. The examples of content specifications for Colorado Reading and Indiana Mathematics are shown in Table 5.

Data Analysis

Using Proc TCALIS in SAS® 9.2 (SAS Institute Inc., 2008), both confirmatory factor analysis (CFA) and multi-group confirmatory factor analysis (MGCFA) were conducted to determine the adequacy of fit of the factor structures of MAP tests and invariance of factor models across grades and academic terms (invariance across terms were not statistically tested). All estimation in this study use the maximum likelihood method.

All MAP tests assume there is only one latent factor (student achievement) that accounts for covariance among observed variables at item or item cluster levels. All MAP tests were scaled based on unidimensional Rasch model (Rasch, 1980) and *RIT* (Rasch unIT) scale that is linearly transformed from logit (*RIT* = *logit* x 10 + 200, NWEA, 2011). Figure 3 present CFA and MGCFA models of MAP tests and the detailed information of the represented models can be found in papers of McArdle (1988) and McDonald (1985).

The one-factor model with goal scores (or subtests) as observed variables and CFA was used to evaluate the adequacy of model to fully account for the relationships among subtests. Once adequacy of model fit was determined, MGCFA was used to test whether the same model holds across different groups. According to Steenkamp and Baumgarther (1998), the invariance of factor loadings is sufficient for construct comparability across groups. In this study, the additional condition of invariance of factor variance was also tested. Three levels of invariance across 5 grades at each of the academic calendars tested are no constraint (NC), equal factor loading (L), and equal factor loadings and factor variances (LV, see Appendix A).

Several well-known goodness-of-fit indexes (GOF) were used to evaluate model fit: (1) absolute indexes that include chi-square χ^2, unadjusted goodness-of-fit indexes (GFI), and standardized root mean square residual (SRMR); (2) incremental indexes that include the

comparative fit index (CFI) and Bentler-Bonett normal fit index (NFI); (3) parsimony index, the root mean square error of approximation (RMSEA). For group comparisons with increased constraints, the χ^2 value provides the basis of comparison with the previously fitted model, although χ^2 is not considered as the best practice because it is sample size dependent. A non-significant difference in χ^2 values between nested models reveals that all equality constraints hold across the groups. Therefore, the measurement model remains invariant across groups as the constraints are increased. A significant χ^2 does not necessarily indicate a departure from invariance when the sample size is large. Hu and Bentler (1999) recommended using combinations of GOF indices to obtain a robust evaluation of data-model fit in structural equation modeling. The cutoff criterion values of good model fit they recommended are CFI, GFI, NFI > 0.95, RMSEA < 0.06, and SRMR < 0.08. It is worth to note that many researchers (March 2007a, Marsh, Hau, & Grayson, 2005) showed that GOF criteria from Hu and Bentler (1999) are too restrictive.

Results

1. *Results of CFA*

Tables 6 and 7 present the summaries of GOF indexes for independent models of Washington MAP Reading and South Carolina MAP Mathematics tests for by grade and term (because of limited space, only partial results are listed for two of 10 states). Although not shown in the table, all factor loadings of models across content, grades, and states are statistically significant. There are mixed results on the statistically significant χ^2 tests (Washington Reading tests are not significant and South Carolina Mathematics are significant) and very similar patterns of χ^2 tests results hold for the rest of the states tests. However, given the large sample sizes across states, it is not surprising to have statistically significant χ^2 tests results for some states. All values of fit indexes (except RMSEAs for Michigan MAP Mathematics tests) satisfy the Hu and Bentler (1999) criteria and show that each model fits data extremely well for different content areas, grades, terms, and states. Overall results suggest that the one-factor (unidimensional) model is the most reasonable model for MAP tests in these 10 states.

2. *Results of MGCFA*

Tables 8 and 9 display the summaries of GOF indexes of the nested models that tested for measurement invariance across grades for Kansas MAP Reading and Michigan Mathematics Tests. In the nested model comparison, the effect of constraints (NC, L, and LV) imposed on

less restricted modes can be evaluated by using the difference between χ^2 (called $\Delta\chi^2$) because it is distributed as χ^2 with the degree of freedom equal to the difference in degrees of freedom between the two models. The null hypothesis of no significant difference in fit is tested by evaluating whether the chi-square difference is significant. If the difference is significant, then the null hypothesis is rejected (Loehlin, 2004). However, the χ^2 test may be misleading because (1) the more complex the model, the more likely a good fit, (2) the larger the sample size, the more likely the rejection of the model and the more likely a Type II error, and (3) the chi-square fit index is also very sensitive to violations of the assumption of multivariate normality. To address these limitations, the difference of other GOF (CFI, GFI, NFI, RMSEA, and SRMR) as adjuncts to the χ^2 statistic can also be used to assess model fit. For the Kansas MAP Reading Tests (see Table 8), χ^2 increases ($\Delta\chi^2$) are significant for testing L invariance at different terms, but not significant for testing LV invariance. The rest of states results show a similar pattern. For Michigan Mathematics Tests, all χ^2 increases are significant for both L and LV invariance. All fit indexes for both Reading and Mathematics tests for different grades and academic years from 10 states satisfied Hu and Bentler's criteria, except RMSEAs and SRMRs for Michigan Mathematics Tests. In summary, the results provide clear support for the metric invariance for all tests except for Michigan Mathematics Tests, and at least, there are configure invariances for all tests.

These results suggest that constructs of MAP tests are well defined, proved to be unidimensional equivalent across grades, and have the same patterns across academic years.

Scientific Significance of the Study

The factor structure of test for a particular grade is directly related to the construct validity interpretation of the test, and validity is one of the most important considerations when evaluating a test. The factor invariance across grades is a fundamental requirement for use in vertical scaling and interpretation of student growth based on the test scores. There are many challenges to providing validity evidence for CAT tests because of its uniqueness compared to fixed form tests. This study using real data provides empirical evidence of construct and invariance construct of MAP scales across grades at different academic calendars for 10 different states. Results show the consistency and reasonableness of interpretation of the MAP RIT scale across grades and academic calendar years for the different states.

Table 1. Sample Sizes for MAP Reading and Mathematics Tests across States

State Name	Reading	Mathematics
Colorado	256310	259600
Illinois	444485	433595
Indiana	262740	247905
Kansas	217730	211070
Kentucky	149785	148725
Michigan	150945	151645
Minnesota	457630	448470
South Carolina	473135	465525
Washington	316980	316925
Wisconsin	351740	351690

Table 2. Frequency and Percentages* of Samples across Grades and Academic Calendars for Illinois MAP Mathematics Test

Grade	Spring 2009	Fall 2009	Spring 2010	Fall 2010	Spring 2011	Total
3	20000					20000
	(4.61)					(4.61)
4	20000	20000	20000			60000
	(4.61)	(4.61)	(4.61)			(13.84)
5	20000	20000	20000	20000	20000	100000
	(4.61)	(4.61)	(4.61)	(4.61)	(4.61)	(23.06)
6	20000	20000	20000	20000	20000	100000
	(4.61)	(4.61)	(4.61)	(4.61)	(4.61)	(23.06)
7	6719	20000	20000	20000	20000	86719
	(1.55)	(4.61)	(4.61)	(4.61)	(4.61)	(20.00)
8		6719	6719	20000	20000	53438
		(1.55)	(1.55)	(4.61)	(4.61)	(12.32
9				6719	6719	13438
				(1.55)	(1.55)	(3.10)
Total	86719	86719	86719	86719	86719	433595
	(20.00	(20.00)	(20.00)	(20.00)	(20.00)	(100.00)

*: Percentage in parentheses

Table 3. Frequency and Percentages* of Samples across Grades and Academic Calendars for South Carolina MAP Reading Test

Grade	Spring 2009	Fall 2009	Spring 2010	Fall 2010	Spring 2011	Total
3	20000					20000
	(4.23)					(4.23)
4	20000	20000	20000			60000
	(4.23)	(4.23)	(4.23)			(12.68)
5	20000	20000	20000	20000	20000	100000
	(4.23)	(4.23)	(4.23)	(4.23)	(4.23)	(21.14)
6	20000	20000	20000	20000	20000	100000
	(4.23)	(4.23)	(4.23)	(4.23)	(4.23)	(21.14)
7	14627	20000	20000	20000	20000	94627
	(3.09)	(4.23)	(4.23)	(4.23)	(4.23)	(20.00)
8		14627	14627	20000	20000	69254
		(3.09)	(3.09)	(4.23)	(4.23)	(14.64)
9				14627	14627	29254
				(3.09)	(3.09)	(6.18)
Total	94627	94627	94627	94627	94627	473135
	(20.00	(20.00)	(20.00)	(20.00)	(20.00)	(100.00)

*: Percentage in parentheses

Table 4. Test Length and Numbers of Goals (subtests) of Reading and Mathematics Tests for Grades 3 to 9 across States

State Name	Reading		Mathematics	
	Test Length	Number of Goal	Test Length	Number of Goal
Colorado	40	4	50	6
Illinois	40	4	50	5
Indiana	40	5	50	7
Kansas	40	5	50	4
Kentucky	40	5	50	5
Michigan	40	4	50	6
Minnesota	40	4	50	4
South Carolina	40	3	50	5
Washington	40	5	50	4
Wisconsin	40	4	50	5

Table 5. Content Specifications of Colorado Reading and Indiana Mathematics for Grades 3 to 9

Colorado Reading Goal	% items per goal	Indiana Mathematics Goal	% items per goal
Reading Strategies, Comprehending Literary Texts	25%	Number Sense	14%
Comprehending Informative and Persuasive Texts	25%	Computation	14%
Word Relationships and Meanings	25%	Algebra and Functions	14%
Total operational items	25%	Geometry	14%
		Measurement	14%
		Statistics, Data Analysis, and Probability	14%
		Problem Solving	14%

Table 6. Summary of Goodness-of- Fit Indexes of Models of Washington MAP Reading Tests for Each Grade at Each Academic Calendar

Academic Calendar	Grade	N	χ^2	df	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	3	12795	24.94	5	1.00	1.00	1.00	0.02	0.00
	4	13296	7.19	5	1.00	1.00	1.00	0.01	0.00
	5	12957	9.11	5	1.00	1.00	1.00	0.01	0.00
	6	14285	7.98	5	1.00	1.00	1.00	0.01	0.00
	7	10065	3.33	5	1.00	1.00	1.00	0.00	0.00
Fall 2009	4	12795	10.07	5	1.00	1.00	1.00	0.01	0.00
	5	13296	15.13	5	1.00	1.00	1.00	0.01	0.00
	6	12957	12.55	5	1.00	1.00	1.00	0.01	0.00
	7	14285	15.06	5	1.00	1.00	1.00	0.01	0.00
	8	10065	6.52	5	1.00	1.00	1.00	0.00	0.00
Spring 2010	4	12795	17.56	5	1.00	1.00	1.00	0.01	0.00
	5	13296	19.52	5	1.00	1.00	1.00	0.01	0.00
	6	12957	10.78	5	1.00	1.00	1.00	0.00	0.00
	7	14285	5.72	5	1.00	1.00	1.00	0.00	0.00
	8	10065	7.29	5	1.00	1.00	1.00	0.00	0.00
Fall 2010	5	12795	15.39	5	1.00	1.00	1.00	0.01	0.00
	6	13296	5.28	5	1.00	1.00	1.00	0.00	0.00
	7	12957	11.66	5	1.00	1.00	1.00	0.01	0.00
	8	14285	18.77	5	1.00	1.00	1.00	0.01	0.00
	9	10065	6.06	5	1.00	1.00	1.00	0.00	0.00
Spring 2011	5	12795	4.4	5	1.00	1.00	1.00	0.00	0.00
	6	13296	10.63	5	1.00	1.00	1.00	0.01	0.00
	7	12957	15.08	5	1.00	1.00	1.00	0.01	0.00
	8	14285	13.90	5	1.00	1.00	1.00	0.01	0.00
	9	10065	4.52	5	1.00	1.00	1.00	0.00	0.00

Table 7. Summary of Goodness-of- Fit Indexes of Models of South Carolina MAP Mathematics Tests for Each Grade at Each Academic Calendar

Academic Calendar	Grade	N	χ^2	df	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	3	20000	58.25	5	1.00	0.99	1.00	0.02	0.01
	4	20000	109.70	5	0.99	1.00	1.00	0.03	0.00
	5	20000	152.66	5	1.00	1.00	0.99	0.04	0.00
	6	20000	88.94	5	1.00	0.99	1.00	0.03	0.01
	7	13205	65.05	5	0.99	1.00	1.00	0.03	0.00
Fall 2009	4	20000	70.17	5	0.99	1.00	1.00	0.03	0.00
	5	20000	113.07	5	1.00	0.99	1.00	0.03	0.01
	6	20000	58.18	5	1.00	1.00	0.99	0.02	0.00
	7	20000	88.11	5	1.00	1.00	1.00	0.03	0.00
	8	13205	74.75	5	1.00	1.00	0.99	0.02	0.00
Spring 2010	4	20000	149.51	5	1.00	0.99	1.00	0.04	0.01
	5	20000	180.02	5	1.00	0.99	1.00	0.03	0.01
	6	20000	145.66	5	1.00	1.00	0.99	0.04	0.00
	7	20000	128.38	5	1.00	1.00	1.00	0.04	0.00
	8	13205	78.99	5	1.00	1.00	0.99	0.03	0.00
Fall 2010	5	20000	151.35	5	1.00	0.99	1.00	0.04	0.01
	6	20000	37.83	5	1.00	1.00	1.00	0.02	0.00
	7	20000	70.91	5	1.00	1.00	0.99	0.03	0.00
	8	20000	102.66	5	1.00	0.99	1.00	0.03	0.00
	9	13205	66.81	5	1.00	1.00	1.00	0.03	0.00
Spring 2011	5	20000	140.04	5	1.00	0.99	1.00	0.04	0.01
	6	20000	89.24	5	1.00	1.00	1.00	0.03	0.00
	7	20000	158.49	5	1.00	1.00	1.00	0.04	0.00
	8	20000	201.38	5	1.00	1.00	1.00	0.04	0.01
	9	13205	75.93	5	0.99	1.00	1.00	0.03	0.00

Table 8. Results of Comparisons of Model Invariance of Kansas MAP Reading Tests across Five Grades*

Academic Calendar	Grade/ Group	Model	χ^2	df	$\Delta\chi^2$	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	G3 to G7	1. NC	67.00	25		1.00	1.00	1.00	0.01	0.00
		2. L	528.76	41	461.76	1.00	1.00	0.99	0.02	0.02
		3. LV	619.04	45	90.28	1.00	0.99	1.00	0.02	0.04
Fall 2009	G4 to G8	1. NC	29.88	25		1.00	1.00	1.00	0.00	0.00
		2. L	225.73	41	195.84	1.00	1.00	1.00	0.02	0.03
		3. LV	258.16	45	32.44	1.00	0.99	1.00	0.03	0.04
Spring 2010	G4 to G8	1. NC	61.99	25		1.00	1.00	1.00	0.01	0.00
		2. L	509.85	41	447.86	1.00	0.99	1.00	0.04	0.03
		3. LV	566.77	45	56.92	0.99	0.99	0.99	0.04	0.05
Fall 2010	G5 to G6	1. NC	27.54	25		1.00	1.00	1.00	0.00	0.00
		2. L	333.64	41	306.10	1.00	0.99	1.00	0.03	0.03
		3. LV	379.73	45	46.09	1.00	1.00	1.00	0.03	0.04
Spring 2011	G5 to G6	1. NC	65.99	25		1.00	1.00	1.00	0.01	0.00
		2. L	353.53	41	287.54	1.00	0.99	1.00	0.03	0.03
		3. LV	487.08	45	133.54	1.00	1.00	1.00	0.03	0.06

* The levels of model constraints restricted to be equal across grades are:
1. NC: No Constraint (Model structure).
2. L: Factor loading .
3. LV: Factor loading + Factor Variance.

Table 9. Results of Comparisons of Model Invariance[1] of Michigan MAP Mathematics Tests across Five Grades[*]

Academic Calendar	Grade/ Group	Model	χ^2	df	$\Delta\chi^2$	CFI	GFI	NFI	RMSEA	SRMR
Spring 2009	G3 to G7	1. NC	1844.03	45		0.98	0.98	0.98	0.08	0.02
		2. L	3410.33	65	1566.30	0.97	0.97	0.96	0.09	0.07
		3. LV	3726.05	69	315.72	0.96	0.96	0.96	0.09	0.12
Fall 2009	G4 to G8	1. NC	1995.99	65		0.98	0.98	0.98	0.08	0.02
		2. L	4008.02	69	2012.02	0.97	0.96	0.97	0.10	0.08
		3. LV	4328.11	45	320.10	0.97	0.96	0.97	0.10	0.13
Spring 2010	G4 to G8	1. NC	2628.19	45		0.98	0.97	0.98	0.09	0.02
		2. L	4699.64	65	2071.45	0.97	0.95	0.97	0.11	0.08
		3. LV	152033.96	69	147334.32	0.97	0.95	0.97	0.11	0.14
Fall 2010	G5 to G6	1. NC	3212.61	45		0.98	0.97	0.98	0.10	0.02
		2. L	3843.60	65	630.99	0.97	0.96	0.97	0.10	0.04
		3. LV	4298.94	69	455.34	0.97	0.95	0.97	0.10	0.14
Spring 2011	G5 to G6	1. NC	2705.84	45		0.98	0.97	0.98	0.10	0.02
		2. L	4254.03	65	1548.19	0.97	0.96	0.97	0.10	0.06
		3. LV	4326.35	69	72.32	0.97	0.96	0.97	0.10	0.08

* The levels of model constraints restricted to be equal across grades are:
1. NC: No Constraint (Model structure only).
2. L: Factor loading .
3. LV: Factor loading + Factor Variance.

References

Bryne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issues of partial measurement invariance. *Psychological Bulletin, 105(3)*, 456-466.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Loehin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Mahwah, NJ: Erlbaum.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.*), New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-238). New York: Academic Press.

Marsh, H.W., Muthén, B., Asparouhov, A., Lüdtke, O., Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling, 16*, 439-476.

McCall, M. & Hauser, C. (2006). Item Response Theory and Longitudinal Modeling: The Real World is Less Complicated than We Fear. In R. Lissitz, (Ed.), *Assessing and Modeling Cognitive Development in School.* (pp 143-174). Maple Grove, MN: JAM Press.

McArdle, J. J. (1988). Dynamic but Structural Equation Modeling of Repeated Measures Data. In J. R. Nesselroade & R. B. Cattell (Eds.), *The Handbook of Multivariate Experimental Psychology*. New York: Plenum Press.

McDonald, R. P. (1985). *Factor Analysis and Related Methods*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Northwest Evaluation Association. (2011, January). *Technical manual for Measure of Academic Progress & Measure of Academic Progress for Primary Grades*. Portland, Oregon.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Steenkamp, J.B., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research, 25*: 78-90.

Washington State, on behalf of the SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program Application for New Grants*. Retrieved from http://www.k12.wa.us/SMARTER/pubdocs/SBAC_Narrative.pdf.

Way, W. (2006). *Online Testing Research: Information and Guiding Transitions to Computerized Assessments*. A white paper from Pearson Educational Measurement.

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved June 11, 2010, from www.ets.org/s/commonassessments/pdf/AdaptiveTesting.pdf.

Yen, W. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. Applied Psychological Measurement, 8(2), pp 125-145.

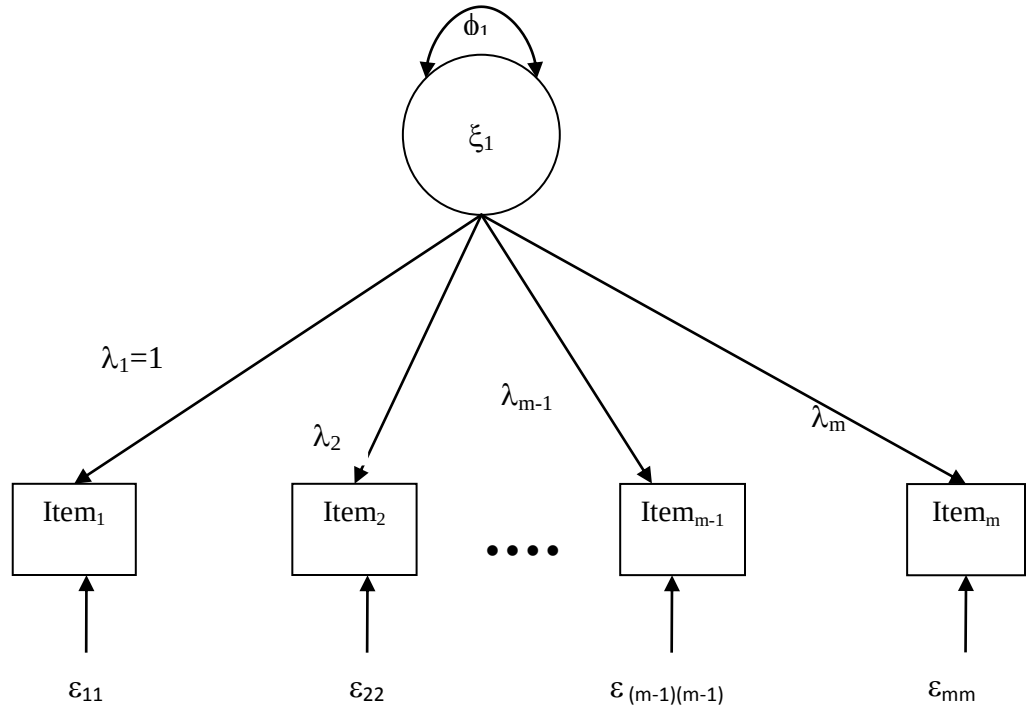


Figure 1. Individual item as observables

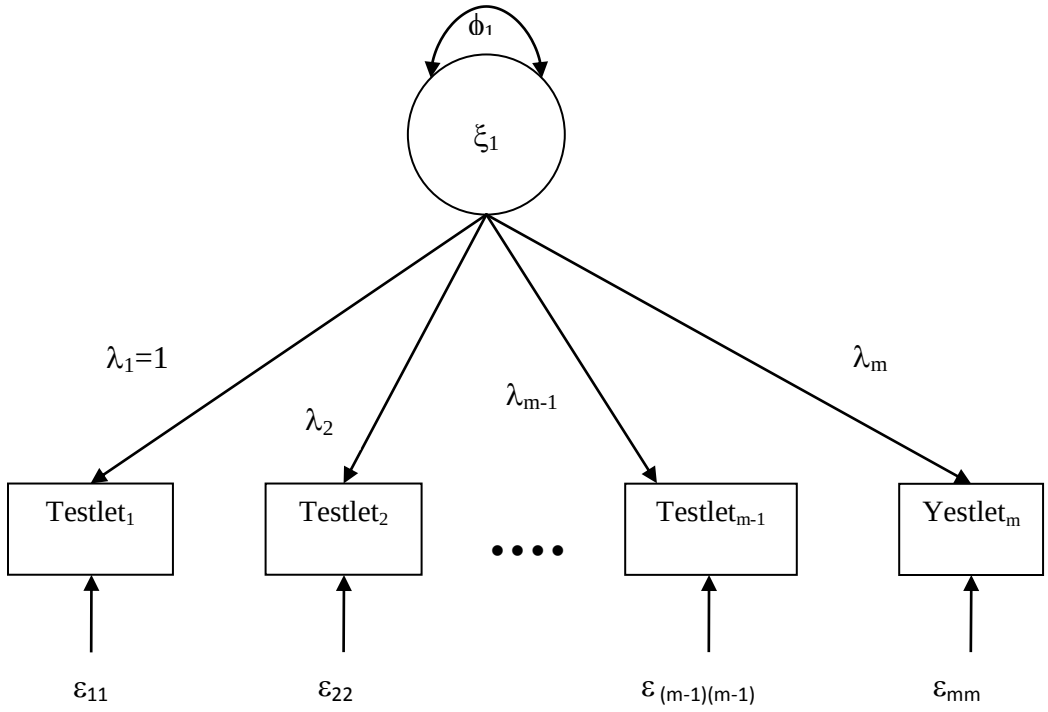
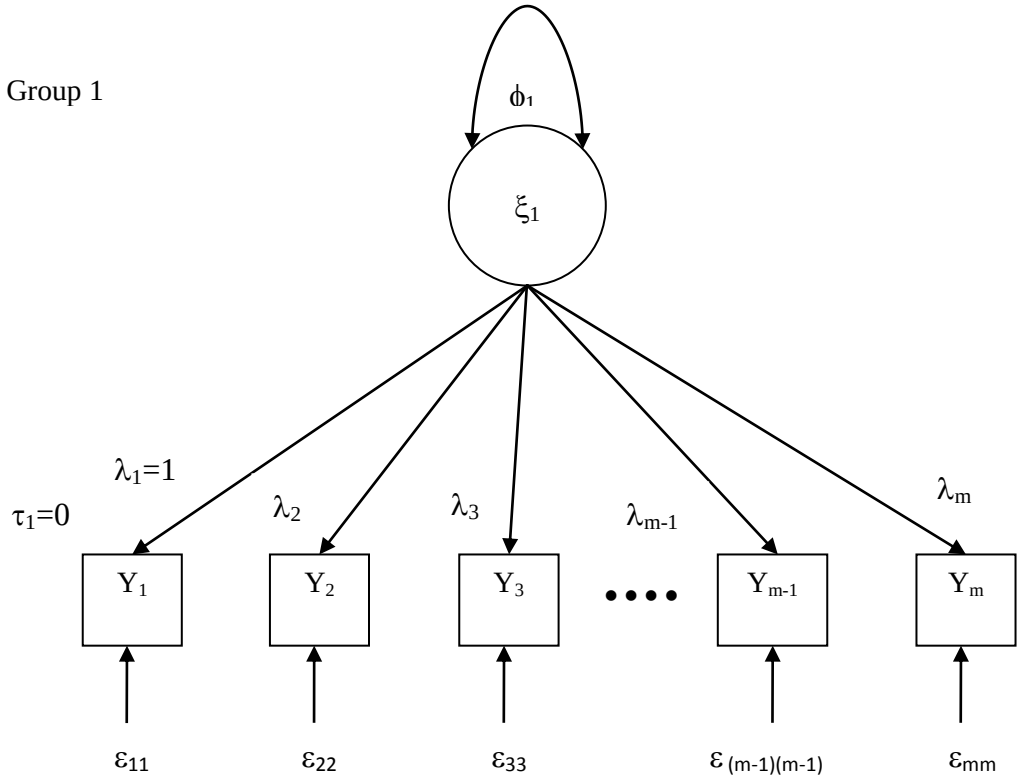


Figure 2. Testlets as observable variables


Group 1
ϕ1
ξ1
λ1=1
τ1=0
λ2
λ3
λm-1
λm
Y1
Y2
Y3
Ym-1
Ym
ε11
ε22
ε33
ε (m-1)(m-1)
εmm

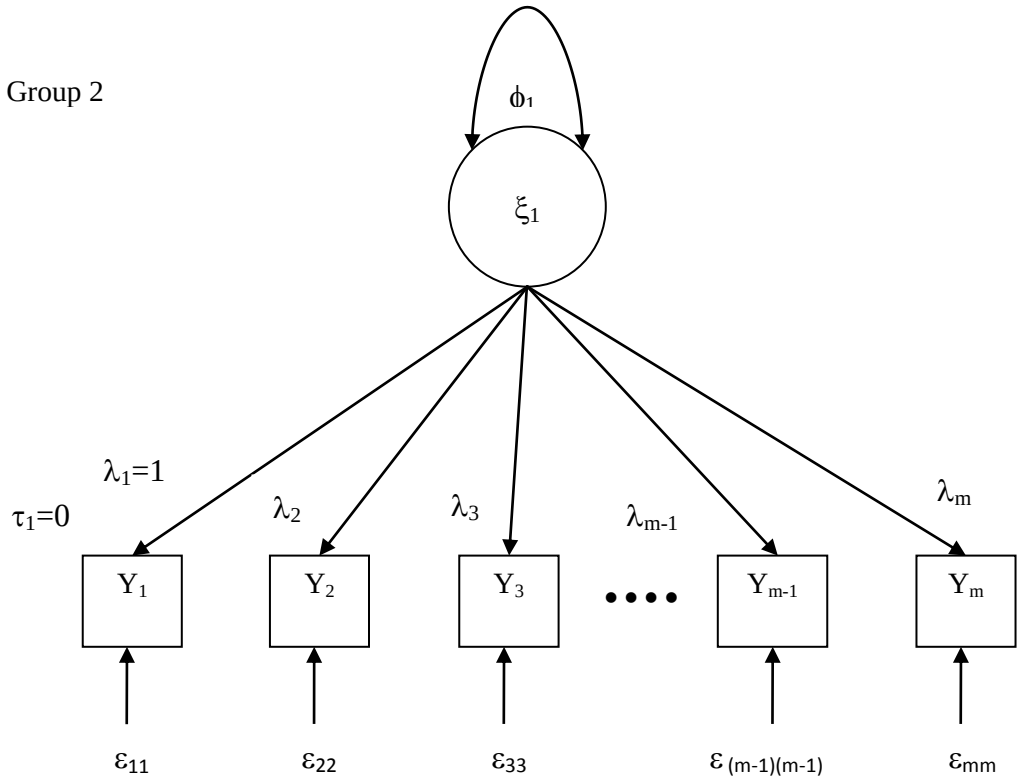

Group 2
ϕ1
ξ1
λ1=1
τ1=0
λ2
λ3
λm-1
λm
Y1
Y2
Y3
Ym-1
Ym
ε11
ε22
ε33
ε (m-1)(m-1)
εmm

Figure 3. Models of MAP Tests across Groups (group indicators omitted for simplicity)

Appendix A.

1. CFA Measurement Model

$$\mathbf{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \quad \text{(A.1)}$$

where **Y** is the vector of manifest indicator (goal scores in this study), $\boldsymbol{\tau}$ is a vector of measurement intercepts, $\boldsymbol{\Lambda}$ is the matrix of factor loading, and $\boldsymbol{\varepsilon}$ is a vector of residuals. The model-implied covariance matrix is

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \quad \text{(A.2)}$$

where $\boldsymbol{\Phi}$ is the latent variable (achievement in this study) covariance matrix and $\boldsymbol{\Theta}$ is the residual covariance matrix. Because we expect the mean achievement will be different across the grades, mean structure is not our concern in this study. In this study, all measurement intercepts were set to zero.

2. CFA Measurement Model in the Multiple Group

$$\mathbf{Y}^g = \boldsymbol{\tau}^{\mathbf{g}} + \boldsymbol{\Lambda}^{\mathbf{g}}\boldsymbol{\xi}^{\mathbf{g}} + \boldsymbol{\varepsilon}^{\mathbf{g}} \quad \text{(A.3)}$$

Where *g* is group indicator and *g* = 1,2,...5 in this study. $\mathbf{Y}^g$ is the vector of manifest indicator (goal scores in this study), $\boldsymbol{\tau}^{\mathbf{g}}$ is a vector of measurement intercepts, $\boldsymbol{\Lambda}^{\mathbf{g}}$ is the matrix of factor loading, and $\boldsymbol{\varepsilon}^{\mathbf{g}}$ is a vector of residuals. The model-implied covariance matrix is

$$\boldsymbol{\Sigma}^g = \boldsymbol{\Lambda}^{\mathbf{g}}\boldsymbol{\Phi}^{\mathbf{g}}\boldsymbol{\Lambda}'^{\mathbf{g}} + \boldsymbol{\Theta}^{\mathbf{g}} \quad \text{(A.4)}$$

where $\boldsymbol{\Phi}^{\mathbf{g}}$ is the latent variable (achievement in this study) covariance matrix for group g and $\boldsymbol{\Theta}^{\mathbf{g}}$ is the residual covariance matrix for g group. Because we expect the mean achievement will be different across the grades, mean structure is not our concern in this study and all measurement intercepts were set at zero.

According to many researchers (Bryne, Shavelson & Muthén, 1989; Jöreskog ,1971; Marsh, Muthén, Asparouhov, Lüdtke, Robitzsch, Morin &Trautwein, 2009), the invariance of

the parameter matrices implied by equation (A.4) means, the covariance matrices for G groups will only be identical if all of the factor loadings, factor variance and covariances, and residual variance are identical across groups. Although there are total 13 partially nested models (named differently for different researchers) can be tested (Marsh et al., 2009) for model invariance. In this study, three invariance tests conducted are: (1) configure invariance (congeneric invariance) without constraint imposed on parameters; (2) weak factor invariance (tau-equivalent or metric invariance) with constraint of equal factor loading; and (3) invariance of factor loading and factor variance. The invariance tested in this study is summarized as following:

1. No constraint, baseline model (NC)
2. Equal factor loadings (L)

$$H_0: \ \boldsymbol{\Lambda}^1 = \boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}^3 = \boldsymbol{\Lambda}^4 = \boldsymbol{\Lambda}^5$$

3. Equal factor loadings and factor variance (LV)

$$H_0: \ \boldsymbol{\Lambda}^1 = \boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}^3 = \boldsymbol{\Lambda}^4 = \boldsymbol{\Lambda}^5$$
$$H_0: \ \boldsymbol{\Phi}^1 = \boldsymbol{\Phi}^2 = \boldsymbol{\Phi}^3 = \boldsymbol{\Phi}^4 = \boldsymbol{\Phi}^5$$