Validation of longitudinal achievement constructs of vertically scaled computerised adaptive tests: a multiple-indicator, latent-growth modelling approach

Shudong Wang*

Northwest Evaluation Association (NWEA),
121 NW Everett St.,
Portland, OR 97206, USA
E-mail: Shudong.Wang@NWEA.org
*Corresponding author

Hong Jiao

Department of Human Development and Quantitative Methodology,
University of Maryland,
College Park, MD 20742, USA
E-mail: hjiao@umd.edu

Liru Zhang

Delaware Department of Education,
401 Federal Street, Suite 2,
Dover, DE 19901, USA
E-mail: Liru.zhang@doe.k12.de.us

Abstract: It is a commonly accepted assumption by educational researchers and practitioners that an underlying longitudinal achievement construct exists across grades in K-12 achievement tests. This assumption provides the necessary assurance to measure and interpret student growth over time. However, evidence is needed to determine whether the achievement construct remains consistent or shifts over grades or time. The current investigative study uses a multiple-indicator, latent-growth modelling (MLGM) approach to examine the longitudinal achievement construct and its invariance for the measures of academic progress® (MAP®), a computerised adaptive test in reading and mathematics. The results of the analyses from ten states suggest that with repeated measures, the construct of both MAP reading and mathematics remained consistent at different time points. The findings support the achievement construct's invariance throughout different grades or time points and provide empirical evidence for measuring student growth.

Keywords: validity; longitudinal achievement construct; computerised adaptive test; CAT; multiple-indicator latent-growth modelling; MLGM.

Reference to this paper should be made as follows: Wang, S., Jiao, H. and Zhang, L. (2013) 'Validation of longitudinal achievement constructs of vertically scaled computerised adaptive tests: a multiple-indicator, latent-growth modelling approach', *Int. J. Quantitative Research in Education*, Vol. 1, No. 4, pp.383–407.

Biographical notes: Shudong Wang is Senior Research Scientist, Northwest Evaluation Association, Portland, OR. His research interests include linking and scaling in educational assessment, item response theory, test validity, computerised adaptive test, and generalised linear mixed model applications in educational research.

Hong Jiao is Associate Professor, Measurement, Statistics and Evaluation, Department of Human Development and Quantitative Methodology, University of Maryland. Her primary research interests include item response theory (IRT), extended IRT modelling (multilevel IRT modelling and mixture IRT modelling), and psychometrics in large-scale assessments.

Liru Zhang is Education Associate at the Delaware Department of Education. Her research interests include linking and equating, computerised adaptive test, and student growth in large-scale assessments.

1 Introduction

Most current achievement tests can be characterised by their test algorithm and administration platforms. Two major testing algorithms are commonly employed in achievement tests: one is the linear test, in which all students answer all test items. Another is the computerised adaptive test (CAT), in which different students may get different items. During adaptive testing, each item that administrates to an individual student adapts to his or her provisional achievement ability by selecting the most appropriate item difficulty. The most common platforms for achievement tests are paper and pencil and computer. When the linear test algorithm is used in computer platforms, the test is called a 'computer-based test'; when the CAT algorithm used in computer platforms, the test is a 'CAT'.

For decades, achievement tests delivered either in linear or CAT algorithms have been constructed to provide formative or summative measures about students' achievement status at grade level (for example, Amy's third-grade mathematics test score in Fall 2005), academic growth over time from longitudinal data (for example, Amy's third-grade mathematics test score in Fall 2005 and her seventh-grade mathematics score in Fall 2009) or achievement across grades from cross-sectional data (for example, Amy's third-grade, Johnny's fourth-grade and Tim's fifth-grade mathematics scores in Fall 2005) (Hamilton et al., 2008; Patz, 2007; Smith and Yen, 2006; Yen, 2007, 2009). In recent educational reforms, assessing individual students' growth has been required for high-stakes decisions by state and federal education policy, such as the race to the top initiatives (RTTT) (U.S. Department of Education, 2009). These requirements have put tremendous pressures on states and testing companies to develop high-quality and high-utility assessment systems. As RTTT states: "(Student) growth may be measured by a variety of approaches, but any approach used must be statistically rigorous and based on student achievement data, and may also include other measures of student learning in order to increase the construct validity and generalizability of the information" [U.S. Department of Education, (2009), p.37812]. CATs are considered more effective than linear tests for measuring individual students' growth over time (Way et al., 2010). The advantages of CATs over linear tests include shorter tests, immediate student score reporting, higher reliability and measurement accuracy (Kingsbury and Weiss, 1983; Lord, 1977; Thissen and Mislevy, 1990), cost savings and multiple testing opportunities for formative and interim assessments (Way, 2006).

Measuring individual student growth has two fundamental requirements (Bergman et al., 1991; Betebenner and Linn, 2010; Doran and Cohen, 2005; Linn, 1993; Mislevy, 1992; Williamson et al., 1991). First, there are multiple measures of achievement construct along the growth trajectory. Second, the achievement construct should be invariant from different grade levels or points in time.

This means that in order to measure student growth, achievement tests must satisfy two necessary conditions: First, there must be a continuous construct that the tests are designed to measure over grades or time; i.e., the tests are scored on a longitudinal construct. Second, the construct measured by multiple tests must be invariant, or constant, across grade levels and time; i.e., the construct does not shift over grade levels or time (Wang and Jiao, 2009).

Many researchers have investigated whether achievement tests satisfy these two assumptions from the perspectives of content and vertical scaling (Cizek, 2005; Linn, 2001; Lissitz, 2006; Martineau et al., 2007; Wang and Jiao, 2009; Wise, 2004). The study by Wang and Jiao (2009) provided empirical evidence to support construct invariance across grades for a vertically scaled norm-referenced test. However, few studies examine the longitudinal achievement construct and construct invariance across grade levels and points in time with CAT.

Achievement construct invariance is an important validity issue. According to the *Standards for Educational and Psychological Testing* [American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME), 1999], validity is the most important consideration in test development and evaluation. Validity refers to the degree to which empirical evidence and theoretical rationale support the inferences and actions based on test scores (Messick, 1989) or the degree to which evidence and theory support the interpretations of test scores for the proposed use of tests (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). Interpretations of test results that measure achievement ability or traits are subject to many validity threats, including two major threats: construct-under representation (CUR) and construct-irrelevant variance (CIV) (Messick, 1984). The construct of a test is a theoretical representation of the underlying traits, concepts, attributes, processes or structures the test is designed to measure, and it directly relates to test validity (Cronbach, 1971; Messick, 1989). Five sources of validity evidence specified in the *Standards* include

a test content

b response process

c internal structure

d relations to other variables

e consequences of testing.

CUR and CIV could be identified in the process of test development, administration and use of test results. This study focuses on validity evidence of internal structure and the invariance of the internal structure of achievement tests across grades and/or over time.

Factorial validity can be a valuable component of validity evidence (Guilford, 1946; Messick, 1995). Although evidence of a test's internal structure is routinely reported in many test technical reports or manuals from state assessment programmes and test companies, construct invariance is rarely addressed from a longitudinal perspective when test results are routinely used to determine student growth. Many test publishers have used the first-order latent growth curve model (FLGM) (Bollen and Curran, 2006; McArdle et al., 2009; Muthén, 1995) in longitudinal studies of student growth, based on the assumption that there is a continuous test construct. Unfortunately, little attention has been drawn to examine the assumption in practice. The second-order latent growth model, or the multiple-indicator latent growth model (MLGM), is rarely used to evaluate the longitudinal construct validity for adaptive test achievement based on CAT. There are two advantages of MLGM over FLGM (Bollen and Curran, 2006; Ferrer et al., 2008; Muthén and Muthén, 2007; Sayer and Cumsille, 2001). First, MLGM can test, instead of making assumptions, whether the same latent construct is measured at each point in time or grade level. Second, it accounts for important information of the psychometric properties on the indicators. MLGM can be used to evaluate factorial invariance across points in time and determine whether the same latent construct is measured across time or grade levels so as to assure that changes of test scores quantify achievement growth rather than the shift of the achievement test construct.

FLGM uses total scores as an observable variable or indictor in the longitudinal achievement growth analysis, while MLGM uses observed indicators either at the item level or the cluster of items level (testlet, subtest, goal score, etc.). The choice of the level of an indicator in MLGM or in general factor analysis has a significant effect on evaluating the construct, because item cluster involves averaging item scores and using summed scores as observed variables in analyses (Bandalos, 2002; Bandalos and Finney, 2001; Hall et al., 1999; Little et al., 2002; Nasser and Wisenbaker, 2003). Some reasons to use clusters of items are to reduce the problems of non-normality, to have fewer free parameters to be estimated compared to the number of observations and to improve data-model fit. Arguments state that using clusters of items increases the chance of combining items that truly measure multiple dimensions and therefore results in severe bias. Overall, clustering items is a commonly used technique based upon theoretical rationale.

Compared with linear tests, it is almost impossible to use indicators at the item level for adaptive tests because each student receives different items tailored to his or her ability. Using unique test forms with CATs makes evaluating longitudinal achievement constructs complicated. Compared with linear tests, CATs present two major challenges in using observable variables for MLGM. First, observable variables are different across persons at the item level. Second, even though the observable variables are the same at the cluster level, the context of the observable variables is different; i.e., the same subtest score may consist of different items for the same content. One possible solution to the first problem is to conduct a confirmatory factor analysis at the item level on the entire item bank. However, the drawback is the large amount of missing response data in the dataset. For example, the missing rate for the measure of academic progress (MAP, Northwest Evaluation Association, 2011) is around 98% for the reading and mathematics tests. Besides, the data are very sparse, because the ratio of the test length to the size of the item bank is about 1:50 (Wang and Harris, 2011). The commonly used imputation method (Rubin, 1987) may statistically help solve the problem, but it cannot deal with the missing data issue from the content perspective. Another possible solution for improvement is to treat items as nested inside person within the framework of generalisability theory, which provides information on parcel scores reliability.

The purpose of this study is to investigate the longitudinal achievement constructs of a standardised, large-scale CAT in reading and mathematics across ten states.

2 Method

2.1 Data source and participants

All data used in this study were collected from the measures of academic progress (MAP) assessment system from Spring 2009 through Spring 2011. The MAP reading and mathematics tests were administered to students in grades 3–10 across 50 US states. The data analysis only focused on the ten states (Colorado, Illinois, Indiana, Kansas, Kentucky, Michigan, Minnesota, South Carolina, Washington and Wisconsin) that supplied the largest samples among the 50 states. For each state, the data were collected as part of a five-wave panel design (Spring 2009/Grade 5, Fall 2009/Grade 6, Spring 2010/Grade 6, Fall 2010/Grade 7 and Spring 2011/Grade 7). Each state sampled the corresponding population under the constraints that students must have five academic calendar records (five-wave) from grade 5. Student demographic data is collected legally and ethically through the Master Services Agreement with each NWEA partner. And at no time is personally identifiable information (PII) released without the written consent and permission of the associated partner. All data collection, storage and use in studies meets the requirements of the Family Educational Rights and Privacy Act (FERPA, U.S. Department of Education, 2012). Prior to each test administration, NWEA partners roster students who will take the MAP, and it is through this rostering process that student PII is provided (student first, last and middle name; date of birth, grade, gender and ethnicity). This data assists in properly identifying and tracking students from one test administration to another in order to assign the correct test scores, goal data and other relevant test characteristics.

Although this paper does not focus on the effect of heterogeneity of samples on parameter estimates of the MLGM, a previous study (Zhang and Wang, 2012) found statistically significant different rates of change by gender, using across-grade cohorts with a non-linear hierarchical linear model (HLM).

Table 1 and Table 2 list the frequency distributions and percentages of reading and mathematics for sample states (due to limited space) across five-wave by state, gender and *ethnicity*.

Table 1 Samples of MAP reading tests for ten states

State	*Gender*		*Ethnicity**								*Total*
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Colorado	F	N	78	51	119	729	4	1,351	57	93	2,482
		%	1.53	1.00	2.34	14.34	0.08	26.58	1.12	1.83	48.84
	M	N	59	50	113	708	6	1536	52	76	2,600
		%	1.16	0.98	2.22	13.93	0.12	30.22	1.02	1.50	51.16
Illinois	F	N	58	754	698	2,299	13	5,486	220	342	9,870
		%	0.29	3.77	3.49	11.50	0.07	27.43	1.10	1.71	49.35
	M	N	53	692	725	2,320	15	5,733	259	333	10,130
		%	0.27	3.46	3.63	11.60	0.08	28.67	1.30	1.67	50.65
Indiana	F	N	20	77	365	254		4,247	162	124	5,249
		%	0.19	0.74	3.50	2.43		40.71	1.55	1.19	50.31
	M	N	9	60	295	249	1	4,285	156	129	5,184
		%	0.09	0.58	2.83	2.39	0.01	41.07	1.50	1.24	49.69
Kansas	F	N	189	130	226	355	8	3,236	103	105	4,352
		%	2.13	1.46	2.54	3.99	0.09	36.38	1.16	1.18	48.93
	M	N	204	138	256	404	3	3,329	105	103	4,542
		%	2.29	1.55	2.88	4.54	0.03	37.43	1.18	1.16	51.07
Kentucky	F	N	4	41	362	95	1	2,491	26	264	3,284
		%	0.06	0.63	5.55	1.46	0.02	38.21	0.40	4.05	50.37
	M	N	5	21	337	98	6	2,446	37	286	3,236
		%	0.08	0.32	5.17	1.50	0.09	37.52	0.57	4.39	49.64
Michigan	F	N	25	100	642	139	6	2,295	9	159	3,375
		%	0.37	1.46	9.40	2.03	0.09	33.59	0.13	2.33	49.39
	M	N	30	80	582	109	7	2,454	3	193	3,458
		%	0.44	1.17	8.52	1.60	0.10	35.91	0.04	2.82	50.61
Minnesota	F	N	203	439	607	551	1	7,883	2	262	9,948
		%	1.02	2.20	3.04	2.76	0.01	39.43	0.01	1.31	49.75
	M	N	206	448	616	529	2	8,015	6	224	10,046
		%	1.03	2.24	3.08	2.65	0.01	40.09	0.03	1.12	50.25
South Carolina	F	N	24	118	3,542	518	12	5,514	226	2	9,956
		%	0.12	0.59	17.71	2.59	0.06	27.57	1.13	0.01	49.78
	M	N	22	101	3,561	545	9	5,585	216	5	10,044
		%	0.11	0.51	17.81	2.73	0.05	27.93	1.08	0.03	50.22

Notes: *1. Native American/Alaskan Native; 2. Asian; 3. African American; 4. Hispanic; 5. Native Hawaiian or Other Pacific Islander; 6. White; 7. Multi-ethnic; 8. Not specified or other.

Table 1 Samples of MAP reading tests for ten states (continued)

State	*Gender*		*Ethnicity**								*Total*
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Washington	F	N	156	228	169	1,765	32	3,558	169	214	6,291
		%	1.20	1.76	1.30	13.62	0.25	27.46	1.30	1.65	48.55
	M	N	153	270	177	1905	51	3,739	181	190	6,666
		%	1.18	2.08	1.37	14.70	0.39	28.86	1.40	1.47	51.45
Wisconsin	F	N	126	240	462	525	3	5,755	8	226	7,345
		%	0.84	1.61	3.10	3.52	0.02	38.56	0.05	1.51	49.21
	M	N	99	237	428	527	1	6,044	6	239	7,581
		%	0.66	1.59	2.87	3.53	0.01	40.49	0.04	1.60	50.79

Notes: *1. Native American/Alaskan Native; 2. Asian; 3. African American; 4. Hispanic; 5. Native Hawaiian or Other Pacific Islander; 6. White; 7. Multi-ethnic; 8. Not specified or other.

Table 2 Samples of MAP mathematics tests across five-wave from ten states

State	*Gender*		*Ethnicity**								*Total*
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Colorado	F	N	78	52	123	805	4	1,396	60	92	2,610
		%	1.44	0.96	2.27	14.86	0.07	25.77	1.11	1.70	48.18
	M	N	64	53	131	830	6	1,594	51	78	2,807
		%	1.18	0.98	2.42	15.32	0.11	29.43	0.94	1.44	51.82
Illinois	F	N	52	739	744	2,186	14	5,543	230	360	9,868
		%	0.26	3.70	3.72	10.93	0.07	27.72	1.15	1.80	49.34
	M	N	63	680	745	2,152	11	5,866	260	355	10,132
		%	0.32	3.40	3.73	10.76	0.06	29.33	1.30	1.78	50.66
Indiana	F	N	22	76	370	222	0.00	4,185	163	128	5,166
		%	0.21	0.74	3.61	2.17	0.00	40.84	1.59	1.25	50.41
	M	N	9	61	298	216	1	4,214	152	130	5,081
		%	0.09	0.60	2.91	2.11	0.01	41.12	1.48	1.27	49.59
Kansas	F	N	180	129	238	356	8	3,285	102	108	4,406
		%	2.01	1.44	2.65	3.97	0.09	36.64	1.14	1.20	49.14
	M	N	203	137	271	395	4	3,337	107	106	4,560
		%	2.26	1.53	3.02	4.41	0.04	37.22	1.19	1.18	50.86
Kentucky	F	N	4	42	381	101	1	2,486	27	278	3,320
		%	0.06	0.64	5.77	1.53	0.02	37.66	0.41	4.21	50.287
	M	N	5	22	340	102	6	2,478	36	293	3,282
		%	0.08	0.33	5.15	1.54	0.09	37.53	0.55	4.44	49.712

Notes: *1. Native American/Alaskan Native; 2. Asian; 3. African American; 4. Hispanic; 5. Native Hawaiian or Other Pacific Islander; 6. White; 7. Multi-ethnic; 8. Not specified or other.

Table 2 Samples of MAP mathematics tests across five-wave from ten states (continued)

State	*Gender*		*Ethnicity**								*Total*
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	
Michigan	F	N	22	86	627	172	5	2,301	9	136	3,358
		%	0.32	1.26	9.18	2.52	0.07	33.70	0.13	1.99	49.19
	M	N	24	74	563	137	6	2,479	3	183	3,469
		%	0.35	1.08	8.25	2.01	0.09	36.31	0.04	2.68	50.81
Minnesota	F	N	217	432	628	518	1	7,897	2	254	9,949
		%	1.09	2.16	3.14	2.59	0.01	39.49	0.01	1.27	49.75
	M	N	225	457	600	523	2	8,026	3	215	10,051
		%	1.13	2.29	3.00	2.62	0.01	40.13	0.02	1.08	50.26
South Carolina	F	N	22	110	3622	521	11	5,477	227	2	9,992
		%	0.11	0.55	18.11	2.605	0.055	27.385	1.135	0.01	49.96
	M	N	24	102	3532	540	6	5,591	205	8	10,008
		%	0.12	0.51	17.66	2.7	0.03	27.95	1.025	0.04	50.04
Washington	F	N	167	264	189	1,881	31	3,985	172	265	6,954
		%	1.17	1.85	1.32	13.17	0.22	27.91	1.20	1.86	48.70
	M	N	178	276	188	2,004	54	4,205	186	233	7,324
		%	1.25	1.93	1.32	14.04	0.38	29.45	1.30	1.63	51.30
Wisconsin	F	N	131	241	481	549	4	5,933	19	214	7,572
		%	0.85	1.57	3.13	3.58	0.03	38.64	0.12	1.39	49.31
	M	N	102	236	449	539	4	6,212	14	228	7,784
		%	0.66	1.54	2.92	3.51	0.03	40.45	0.09	1.48	50.69

Notes: *1. Native American/Alaskan Native; 2. Asian; 3. African American; 4. Hispanic; 5. Native Hawaiian or Other Pacific Islander; 6. White; 7. Multi-ethnic; 8. Not specified or other.

2.2 Instruments

MAP has been published by Northwest Evaluation Association™ (NWEA™) since 1976, and all MAP tests are computerised and presented adaptively. The purpose of MAP is to provide educators with information to inform teaching and learning in reading, language usage, mathematics and science (NWEA, 2011). The MAP tests align with the content standards of each state by assembling a customised item pool to measure the specific standards. The variation in the items pools selected across states is reflected in different state tests in terms of the number of goals shown in the Table 3. For example, the MLGM for the South Carolina reading test will consist of three goals, and the MLGM for the Michigan mathematics test will consist of six goals.

Unlike state assessment programmes used to report proficiency under NCLB, MAP tests allow certain off-grade items to be used for on-grade assessment for the purpose of measuring growth. Test algorithms survey the pool within each content standard goal or strand level to assure the content domain coverage. The marginal reliabilities of test scores are consistently in the low- to mid-0.90s range across grades, tests and states

(Northwest Evaluation Association, 2011). MAP items are calibrated with the Rasch model (Rasch, 1961), and all MAP tests are vertically scaled (Northwest Evaluation Association, 2011).

In the process of item selection, all items administered to each student must satisfy the content requirements of each test to ensure the content validity of the test. Table 3 lists test length (fixed-length CAT) and numbers of goals (subtests) for reading and mathematics for the 10 states. A sample of content specifications for reading in Colorado and mathematics in Indiana is shown in Table 4. Due to the uniqueness of CAT test forms, the observed variables used in this study are goal scores (item clusters) under the assumption that the item cluster for each goal area contains homogeneous content across students. All goal scores are scale scores on the same metric across goals and time periods.

Table 3 Test length and numbers of goals (subtests) of reading and mathematics tests for grades 5 to 7 across states

State	*Reading*		*Mathematics*	
	Test length	*Number of goals*	*Test length*	*Number of goals*
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 4 Content specifications of Colorado reading and Indiana mathematics for grades 5 to 7 across states

Colorado reading		*Indiana mathematics*	
Goal	*% items per goal*	*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%

Figure 1 The MLGM at five time points with linear growth structure and invariance of factor loadings

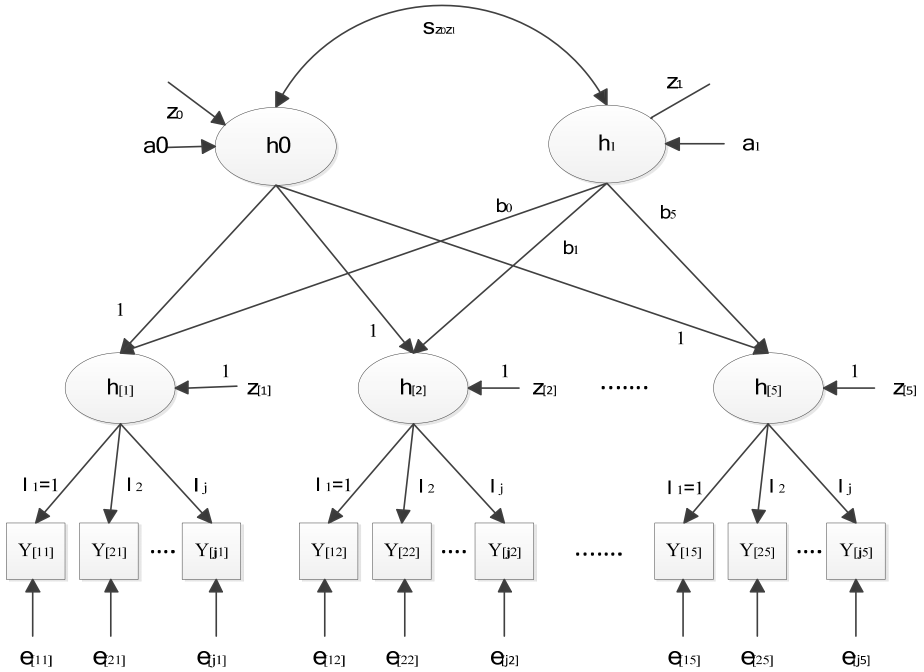
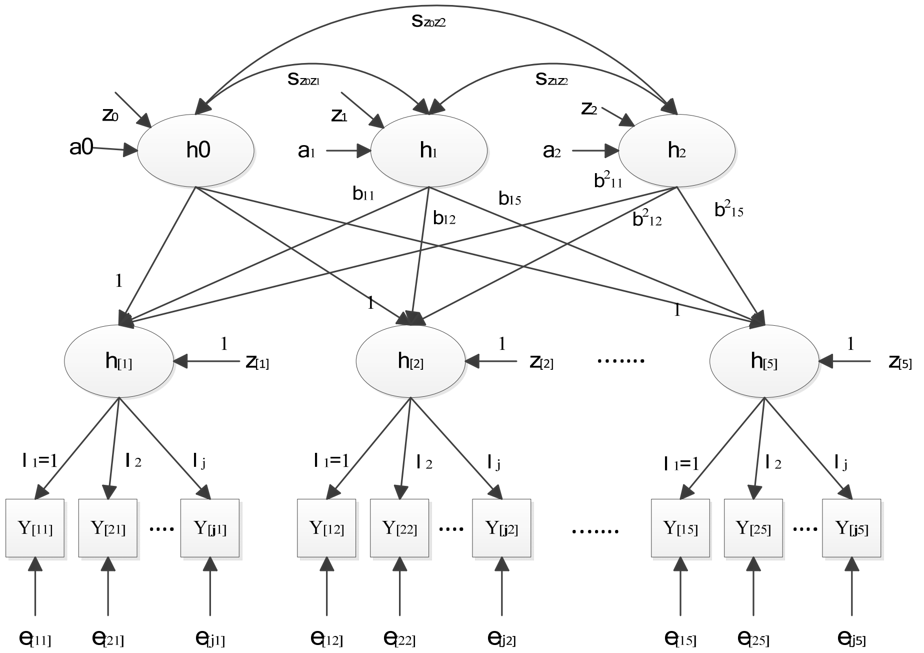


Figure 2 The MLGM at five time points with quadratic growth structure and invariance of factor loadings



2.3 Multiple-indicator, latent-growth model (MLGM)

MLGM is a multivariate extension of FLGM (Bollen and Curran, 2006; Ferrer et al., 2008; McDonald, 1985; McArdle, 1988; Muthén, 1991; Tisak and Meredith, 1990). If y_{jti} denote the observed variables (goals) for individual *i*, indicator *j* and time point *t*, and let η_{ti} denote a latent variable construct, the level-1 model for measurement part is:

$$y_{jti} = \tau_{jt} + \lambda_{jt}\eta_{ti} + \varepsilon_{jti} \tag{1}$$

where τ_{jt} is intercept for the j^{th} indicator in the t^{th} time period, λ_{jt} is the factor loading for the j^{th} indicator at the t^{th} time point, and ε_{jti} is the random error for the i^{th} individual in the t^{th} time point and the j^{th} indicator. Level-1 models for a latent variable with both linear and quadratic growth are:

Linear growth

$$\eta_{ti} = \eta_{0i} + \eta_{1i}\beta_{1t} + \zeta_{ti} \tag{2}$$

Quadratic growth

$$\eta_{ti} = \eta_{0i} + \eta_{1i}\beta_{1t} + \eta_{2i}\beta_{1t}^2 + \zeta_{ti} \tag{3}$$

where ζ_{ti} is the random normal error for the i^{th} individual in the t^{th} time point; η_{0i}, η_{1i} and η_{2i} are the intercept, slope and quadratic of latent factors, respectively, for individual *i*; and β_{1t} and β_{1t}^2 represent t^{th} time point coefficients that determine the shape of the growth curve. Level 2 models are:

Linear growth

$$\eta_{0i} = \alpha_0 + \zeta_{0i} \tag{4}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i} \tag{5}$$

Quadratic growth

$$\eta_{0i} = \alpha_0 + \zeta_{0i} \tag{6}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i} \tag{7}$$

$$\eta_{2i} = \alpha_2 + \zeta_{2i} \tag{8}$$

where ζ_{0i}, ζ_{1i} and ζ_{2i} are normal random errors; α_0, α_1 and α_2 are latent means of intercept, slope and quadratic terms for individual *i* and w_i is weight. All MAP tests have been tested as one latent factor (Wang et al., 2011). Figure 1 and Figure 2 present the MLGM with *linear* and *quadratic* latent growth.

2.4 Measurement invariance of using MLGM

The measurement invariance (Drasgow, 1987; Ferrer et al., 2008; Meredith, 1993) evaluated in this study is the invariance across time points of testing. Although values of manifest variables are different across time in longitudinal studies, they should be on the same measurement scale to derive an equal definition of a latent construct across time. Widaman and Reise (1997) classified two types of factorial invariance as non-metric (configural) and metric. Configural invariance (CI) refers to the same indicators of the latent construct. The metric factorial invariance has three hierarchical levels, which are categorised as weak invariance (WI), where the factor loading of each indicator is invariant over time; strong invariance (SI), where the factor loading and intercept of each indicator are invariant over time; and strict factorial invariance (SFI), where the factor loading, intercept and unique variance of each indicator are invariant over time. Sayer and Cumsille (2001) showed that the SFI is unlikely to hold because heterogeneous variance across time is often observed. In this study, only CI, WI and SI are analysed. The invariance tested is summarised in three conditions: the CI, the WI that can be expressed as equation (9) and the SI that can be expressed in equation (10) and equation (11):

$$\mathrm{H}_0 : \lambda_{j1} = \lambda_{j2} = \ldots \lambda_{jT} = \lambda_j \tag{9}$$

$$\mathrm{H}_0 : \lambda_{j1} = \lambda_{j2} = \ldots \lambda_{jT} = \lambda_j \tag{10}$$

$$\mathrm{H}_0 : \alpha_{j1} = \alpha_{j2} = \ldots \alpha_{jT} = \alpha_j \tag{11}$$

where $\lambda_1 = 1$, $\alpha_0 = 0$ and variances of ε_{jti} and ζ_{ti} may vary over time. And for structural differences, the mean of η_{ti} and variance of η_{ti} vary over time.

Several well-known goodness-of-fit (GOF) indices were used to evaluate the model fit. They are

1 absolute indices that include chi-square χ^2 and standardised root mean square residual (SRMR)

2 incremental indices that include the comparative fit index (CFI) and Tucker-Lewis Index (TLI)

3 parsimony index, the root mean square error of approximation (RMSEA).

For nested models that include a different shape of growth (e.g., linear versus quadratic), both Akaike's Information Criterion (AIC; Akaike, 1987) and Bayesian Information Criterion (BIC) (Schwartz, 1978) are obtained for each model tested. According to Raftery (1995), the values of BIC difference (BIC of quadratic model – BIC of linear model) ranging from 0 to 2 are interpreted as weak evidence for quadratic model, values of 2 to 6 are interpreted as moderate, values of 6 to 10 are interpreted as strong and values > 10 are interpreted as very strong. For model comparisons with increased constraints, the χ^2 value also provides the basis of comparison with the previously fitted model in addition to AIC and BIC. χ^2 is not considered best in 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 time. Therefore, the measurement model remains invariant across groups as the constraints are increased. A significant χ^2value 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 for model-data fit in structural equation modelling. The cut-off criterion values of a good model-fit that they recommended are CFI > 0.95, TLI > 0.95, RMSEA < 0.06 and SRMR < 0.08. It is worth noting that many researchers (Marshet al., 2005; Marsh, 2007; Sayer and Cumsille, 2001) consider the GOF criteria from Hu and Bentler too restrictive. All analyses are conducted in *Mplus* 5.1 (Muthén and Muthén, 2007).

2.5 Analytic approach

The different MLGM models have been used for different state MAP test results because different states have different test specification or test blueprints. For each state, both linear and quadratic MLGM models are fitted to MAP test data. All missing data are handled by using Mplus default and according to Muthén et al. (1978), it is preferable to pairwise deletion (LISTWISE = ON missing data is dealt using pairwise deletion) because pairwise deletion requires missing completely at random (MCAR), not missing at random (MAR).

Figure 3 Observed individuals growth of $Y_{[11]}$ of MAP mathematics test across five-wave for Colorado

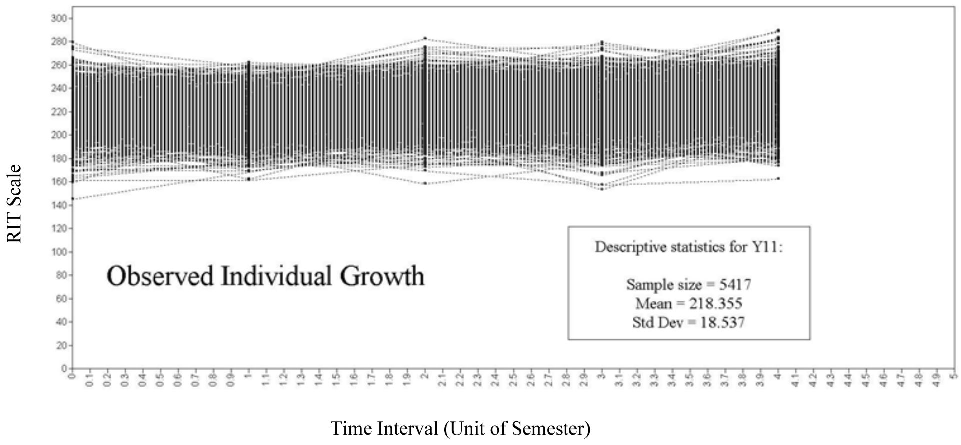


Figure 4 Estimated individuals quadratic growth of $Y_{[11]}$ of MAP mathematics test across five-wave for Colorado

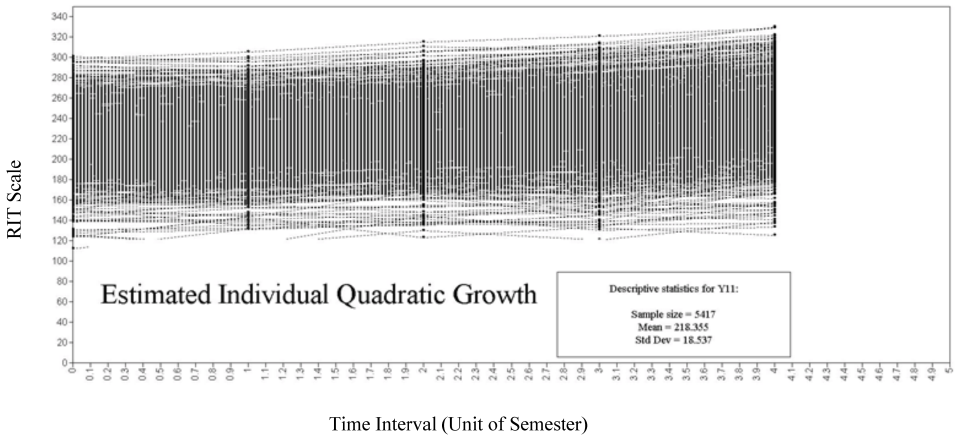


Table 5 Goodness-of fit indexes of MLGM models of MAP reading tests for different states

State	*Model*	*N*	χ^2	*df*	$\Delta\chi^2$	*CFI*	*TLI*	*RMSEA**	*SRMR*	*AIC*	*BIC*
Colorado	Linear	5,082	619.234	170		0.996	0.996	0.023(.021, .025)	0.044	741,066.826	741,458.834
	quadratic		537.906	166	81.328	0.997	0.996	0.021 (.019, .023)	0.043	740,993.498	741,411.640
Illinois	Linear	20,000	3,039.390	170		0.993	0.992	0.029 (.028, .030)	0.050	2,890,494.433	2,890,968.642
	quadratic		2,783.299	166	256.091	0.994	0.993	0.028 (.027, .029)	0.049	2,890,246.342	2,890,752.165
Indiana	Linear	10,433	2,056.374	275		0.992	0.991	0.025 (.024, .026)	0.040	1,928,420.139	1,928,964.094
	quadratic		1,979.157	271	77.217	0.992	0.992	0.025 (.024, .026)	0.041	1,928,350.922	1,928,923.888
Kansas	Linear	8,894	2,110.578	275		0.991	0.990	0.027 (.026, .028)	0.043	1,634,280.583	1,634,812.568
	quadratic		2,090.995	271	19.583	0.991	0.990	0.027 (.026, .028)	0.043	1,634,269.000	1,634,829.357
Kentucky	Linear	6,520	1,179.054	275		0.994	0.994	0.022 (.021, .024)	0.029	1,206,322.986	1,206,831.683
	quadratic		1,082.308	271	96.746	0.995	0.994	0.021 (.020, .023)	0.022	1,206,234.240	1,206,770.068
Michigan	Linear	6,833	1,299.092	170		0.992	0.991	0.031 (.030, .033)	0.052	989,251.073	989,660.844
	quadratic		1,106.289	166	192.803	0.993	0.992	0.029 (.027, .030)	0.050	989,066.269	989,503.359
Minnesota	Linear	19,994	3,713.725	170		0.991	0.990	0.032 (.031, .032)	0.043	2,878,450.276	2,878,924.468
	quadratic		3,466.721	166	247.004	0.992	0.991	0.032 (.031, .032)	0.046	2,878,211.273	2,878,717.077
South Carolina	Linear	20,000	4,609.603	90		0.988	0.986	0.050 (.049, .051)	0.076	2,123,841.525	2,124,197.182
	quadratic		3,456.152	86	1153.451	0.991	0.989	0.044 (.043, .046)	0.038	2,122,696.074	2,123,083.345
Washington	Linear	12,957	2,033.200	275		0.995	0.994	0.022 (.021, .023)	0.026	2,385,763.775	2,386,323.979
	quadratic		1,923.339	271	109.861	0.995	0.995	0.022 (.021, .023)	0.025	2,385,661.914	2,386,251.996
Wisconsin	Linear	14,926	2,718.167	170		0.992	0.991	0.032 (.031, .033)	0.063	2,148,535.797	2,148,992.449
	quadratic		2,490.324	166	227.843	0.992	0.991	0.031 (.030, .032)	0.062	2,148,315.954	2,148,803.050

Note: *Values in parentheses are lower and upper 90% confidence limits of RMSEA.

Table 6 GOF indexes of MLGM models of MAP mathematics tests for different states

State	*Model*	*N*	χ^2	*df*	$\Delta\chi^2$	*CFI*	*TLI*	*RMSEA**	*SRMR*	*AIC*	*BIC*
Colorado	Linear	5,417	3,200.482	405		0.986	0.985	0.036 (.035, .037)	0.033	1,202,562.013	1,203,155.770
	quadratic		2,983.488	401	216.994	0.987	0.986	0.034 (.033, .036)	0.033	1,202,353.011	1,202,973.156
Illinois	Linear	20,000	11,337.522	275		0.982	0.981	0.045 (.044, .046)	0.029	3,618,796.729	3,619,389.491
	quadratic		10,322.301	271	1,015.221	0.984	0.982	0.043 (.042, .044)	0.027	3,617,789.507	3,618,413.883
Indiana	Linear	10,247	148,965.438	560		0.580	0.553	0.161 (.160, . 162)	3.559	2,790,203.181	2,790,962.829
	quadratic		71,963.157	556	77,002.281	0.798	0.784	0.112 (.111, . 113)	0.995	2,713,208.900	2,713,997.487
Kansas	Linear	8,966	4,764.568	170		0.979	0.977	0.055 (.054, .056)	0.048	1,264,358.923	1,264,784.995
	quadratic		3,921.295	166	843.273	0.983	0.980	0.050 (.049, .052)	0.045	1,263,523.650	1,263,978.127
Kentucky	Linear	6,602	3,465.561	275		0.983	0.982	0.042 (.041, .043)	0.020	1,211,001.089	1,211,510.724
	quadratic		2,988.207	271	477.354	0.986	0.984	0.039 (.038, .040)	0.018	1,210,531.735	1,211,068.550
Michigan	Linear	6,827	4,184.200	275		0.981	0.979	0.046 (.044, .047)	0.024	1,245,720.842	1,246,232.990
	quadratic		3,822.681	271	361.519	0.983	0.981	0.044 (.043, .045)	0.024	1,245,367.323	1,245,906.785
Minnesota	Linear	20,000	15,870.194	170		0.970	0.966	0.068 (.067, .069)	0.093	2,857,014.335	2,857,488.544
	quadratic		13,337.092	166	2,533.102	0.975	0.971	0.063 (.062, .064)	0.082	2,854,489.233	2,854,995.056
South Carolina	Linear	20,000	13,233.352	275		0.980	0.978	0.049 (.048, .049)	0.027	3,649,802.550	3,650,395.312
	quadratic		11,124.677	271	2,108.675	0.983	0.981	0.045 (.044, .045)	0.021	3,647,701.876	3,648,326.251
Washington	Linear	14,278	8,230.010	170		0.978	0.976	0.058 (.057, .059)	0.046	2,049,743.767	2,050,197.755
	quadratic		7,615.228	166	614.782	0.980	0.977	0.056 (.055, .057)	0.042	2,049,136.985	2,049,621.239
Wisconsin	Linear	15,356	9,389.812	275		0.980	0.978	0.046 (.046, .047)	0.024	2,780,059.981	2,780,632.926
	quadratic		8,310.063	271	1,079.749	0.982	0.980	0.044 (.043, .045)	0.022	2,778,988.233	2,779,591.735

Note: *Values in parentheses are lower and upper 90% confidence limits of RMSEA.

3 Results

3.1 Results of MLGM

Figure 3 and Figure 4 illustrate a sample of observed and estimated individual quadratic growth based on MGLM across five-wave academic terms. Table 5 and Table 6 display the summaries of GOF indexes of MLGM data fit for linear and quadratic growth in reading and mathematics across states. All values of the fit indexes satisfy the Hu and Bentler (1999) criteria in both content areas and show that each model fits the data extremely well across states, with one exception in Indiana for mathematics.

The overall results suggest that both linear and quadratic MLGMs are reasonably good models for MAP tests in ten states. For AIC, the lower value (positive or negative) indicates a better fit than the higher value. The results show that the quadratic model fits data better than the linear model in the nested modelling comparison. For BIC, all differences between quadratic and linear models are greater than 10 in both reading and mathematics, which indicate that the quadratic model fits the data better than the linear model. The statistically significant χ^2 difference between the quadratic and linear models provides additional evidence to support the conclusion that the quadratic model is a better fit for the data than the linear model. It is also important to note that both the linear and quadratic models show that the longitudinal achievement construct underlying achievement measures equally well in growth.

3.2 Results of invariance of MLGM

Table 7 and Table 8 present the summaries of GOF indexes with nested linear MLGM that was used for measurement invariance across five waves in reading and mathematics. Nearly all fit indexes satisfied Hu and Bentler's criteria in both reading and mathematics tests across states except Indiana. Some SRMRs seemed to be slightly above Hu and Bentler's criteria. In evaluating measurement invariance, the simple model is a restricted model and the complex model is an unrestricted model. The effect of constraints imposed on the less restricted model can be evaluated by using the difference of χ^2 ($\Delta\chi^2$) for nested model comparisons, because the degree of freedom is equal to the difference in the degrees of freedom of two models. Results indicate that all χ^2 increases ($\Delta\chi^2$) are statistically significant for evaluating the differences of invariance between unrestricted and restricted models. As χ^2 becomes statistically significant, a more complex model should be chosen. However, the limitations of the χ^2 test are the sample size dependency (Cheung and Rensvold, 2002) and the difference of other GOF indexes (such as CFI) as adjuncts to the χ^2 statistic, which can also be used to assess model fit. According to Cheung and Rensvold (2002), if the difference of CFI (ΔCFI) is less than 0.01 between the two models, the simple model is not worse than the complex model. The value of all ΔCFIs less than 0.01 in both tests indicates that constrained parameters are invariant across time.

In summary, the results of analyses in this study provide clear support to the CI, WI and SI for all tests except the Indiana mathematics test. These results suggest that longitudinal constructs of MAP tests are well defined for measuring student achievement growth.

Table 7 GOF indexes of invariances of linear MLGM models of MAP reading tests

State	*Model**	*N*	χ^2	*df*	$\Delta\chi^2$	*CFI*	*TLI*	*RMSEA*	*SRMR*	*AIC*	*BIC*
Colorado	CI	5,082	619.234	170		0.996	0.996	0.023	0.044	741,066.826	741,458.834
	WI		764.490	182	145.256	0.995	0.995	0.025	0.092	741,188.082	741,501.688
	SI		1,264.094	194	499.604	0.991	0.991	0.033	0.096	741,663.686	741,898.891
Illinois	CI	20,000	3,039.390	170		0.993	0.992	0.029	0.050	2,890,494.433	2,890,968.642
	WI		3,220.062	182	180.672	0.993	0.992	0.029	0.073	2,890,651.105	2,891,030.472
	SI		6,975.913	194	3,755.851	0.984	0.984	0.042	0.079	2,894,382.956	2,894,667.482
Indiana	CI	10,433	2,056.374	275		0.992	0.991	0.025	0.040	1,928,420.139	1,928,964.094
	WI		2,296.707	291	240.333	0.991	0.991	0.026	0.080	1,928,628.472	1,929,056.383
	SI		4,603.455	307	2,306.748	0.981	0.981	0.037	0.088	1,930,903.220	1,931,215.088
Kansas	CI	8,894	2,110.578	275		0.991	0.990	0.027	0.043	1,634,280.583	1,634,812.568
	WI		2,524.264	291	413.686	0.988	0.988	0.029	0.099	1,634,662.268	1,635,080.763
	SI		4,233.895	307	1,709.631	0.980	0.980	0.038	0.099	1,636,339.899	1,636,644.904
Kentucky	CI	6,520	1,179.054	275		0.994	0.994	0.022	0.029	1,206,322.986	1,206,831.683
	WI		1,234.003	291	54.949	0.994	0.994	0.022	0.049	1,206,345.935	1,206,746.110
	SI		2,077.320	307	843.317	0.989	0.989	0.030	0.058	1,207,157.252	1,207,448.905

Notes: *The levels of model constraints restricted to be equal across grades are CI: configural invariance, WI: weak invariance, SI: strong invariance.

Table 7 GOF indexes of invariances of linear MLGM models of MAP reading tests (continued)

State	*Model**	*N*	χ^2	*df*	$\Delta\chi^2$	*CFI*	*TLI*	*RMSEA*	*SRMR*	*AIC*	*BIC*
Michigan	CI	6,833	1,299.092	170		0.992	0.991	0.031	0.052	989,251.073	989,660.844
	WI		1,394.311	182	95.219	0.991	0.991	0.031	0.086	989,322.292	989,650.109
	SI		2,839.156	194	1,444.845	0.981	0.981	0.045	0.092	990,743.137	990,989.000
Minnesota	CI	19,994	3,713.725	170		0.991	0.990	0.032	0.043	2,878,450.276	2,878,924.468
	WI		3,910.535	182	196.81	0.991	0.991	0.032	0.070	2,878,623.087	2,879,002.440
	SI		8,089.829	194	4,179.294	0.981	0.981	0.045	0.078	2,882,778.381	2,883,062.896
South Carolina	CI	20,000	4,609.603	90		0.988	0.986	0.050	0.076	2,123,841.525	2,124,197.182
	WI		4,891.837	98	282.234	0.987	0.986	0.049	0.112	2,124,107.759	2,124,400.188
	SI		6,308.980	106	1,417.143	0.983	0.983	0.054	0.118	2,125,508.902	2,125,738.104
Washington	CI	12,957	2,033.200	275		0.995	0.994	0.022	0.026	2,385,763.775	2,386,323.979
	WI		2,181.778	291	148.578	0.994	0.994	0.022	0.052	2,385,880.353	2,386,321.047
	SI		3,608.474	307	1,426.696	0.990	0.990	0.029	0.060	2,387,275.050	2,387,596.233
Wisconsin	CI	14,926	2,718.167	170		0.992	0.991	0.032	0.063	2,1485,35.797	2,148,992.449
	WI		2,901.072	182	182.905	0.991	0.991	0.032	0.093	2,148,694.703	2,149,060.024
	SI		5,040.002	194	2,138.93	0.984	0.984	0.041	0.099	2,150,809.632	2,151,083.623

Notes: *The levels of model constraints restricted to be equal across grades are CI: configural invariance, WI: weak invariance, SI: strong invariance.

Table 8 GOF indexes of invariances of linear MLGM models of MAP mathematics tests

State	*Model**	*N*	χ^2	*df*	$\Delta\chi^2$	*CFI*	*TLI*	*RMSEA*	*SRMR*	*AIC*	*BIC*
Colorado	CI	5,417	3,200.482	405		0.986	0.985	0.036	0.033	1,202,562.013	1,203,155.770
	WI		3,327.537	425	127.055	0.985	0.985	0.036	0.057	1,202,649.064	1,203,110.875
	SI		5,630.506	445	2,302.969	0.974	0.974	0.046	0.065	1,284,353.034	1,284,814.845
Illinois	CI	20,000	11,337.522	275		0.982	0.981	0.045	0.029	3,618,796.729	3,619,389.491
	WI		12,480.270	291	1,142.748	0.980	0.980	0.046	0.082	3,619,907.480	3,620,373.786
	SI		20,906.282	307	8,426.012	0.967	0.968	0.058	0.088	3,628,301.492	3,628,641.342
Indiana	CI	10,247	148,965.438	560		0.580	0.553	0.161	3.559	2,790,203.181	2,790,962.829
	WI					**	**	**	**	**	**
	SI		135,770.462	608		0.617	0.625	0.147	**	2,776,912.205	2,777,324.585
Kansas	CI	8,966	4,764.568	170		0.979	0.977	0.055	0.048	1,264,358.923	1,264,784.995
	WI		5,255.879	182	491.311	0.977	0.976	0.056	0.121	1,264,826.234	1,265,167.091
	SI		9,666.501	194	4,410.622	0.957	0.958	0.074	0.126	1,269,212.855	1,269,468.498
Kentucky	CI	6,602	3,465.561	275		0.983	0.982	0.042	0.020	1,211,001.089	1,211,510.724
	WI		3,652.142	291	186.581	0.982	0.982	0.042	0.059	1,211,155.670	1,211,556.582
	SI		6,026.435	307	2,374.293	0.970	0.970	0.053	0.070	1,213,497.964	1,213,790.154

Notes: *The levels of model constraints restricted to be equal across grades are CI: configural invariance, WI: weak invariance, SI: strong invariance.
**Not available.

Table 8 GOF indexes of invariances of linear MLGM models of MAP mathematics tests (continued)

State	*Model**	*N*	χ^2	*df*	$\Delta\chi^2$	*CFI*	*TLI*	*RMSEA*	*SRMR*	*AIC*	*BIC*
Michigan	CI	6,827	4,184.200	275		0.981	0.979	0.046	0.024	1,245,720.842	1,246,232.990
	WI		4,373.172	291	188.972	0.980	0.980	0.045	0.059	1,245,877.814	1,246,280.703
	SI		7,279.799	307	2,906.627	0.966	0.967	0.058	0.071	1,248,752.440	1,249,046.072
Minnesota	CI	20,000	15,870.194	170		0.970	0.966	0.068	0.093	2,857,014.335	2,857,488.544
	WI		17,282.766	182	1,412.572	0.967	0.966	0.069	0.139	2,858,402.908	2,858,782.275
	SI		27,832.970	194	10,550.204	0.947	0.948	0.084	0.142	2,868,929.112	2,869,213.637
South Carolina	CI	20,000	13,233.352	275		0.980	0.978	0.049	0.027	3,649,802.550	3,650,395.312
	WI		14,894.473	291	1,661.121	0.977	0.977	0.050	0.095	3,651,431.671	3,651,897.977
	SI		22,617.917	307	7,723.444	0.965	0.966	0.060	0.094	3,659,123.116	3,659,462.966
Washington	CI	14,278	8,230.010	170		0.978	0.976	0.058	0.046	2,049,743.767	2,050,197.755
	WI		9,544.553	182	1,314.543	0.975	0.974	0.060	0.125	2,051,034.309	2,051,397.500
	SI		1,4495.506	194	4,950.953	0.962	0.962	0.072	0.135	2,055,961.262	2,056,233.656
Wisconsin	CI	15,356	9,389.812	275	9,389.812	0.980	0.978	0.046	0.024	2,780,059.981	2,780,632.926
	WI		9,653.384	291	263.572	0.979	0.978	0.046	0.047	2,780,291.554	2,780,742.270
	SI		13,743.973	307	4,090.589	0.970	0.971	0.053	0.055	2,784,350.143	2,784,678.631

Notes: *The levels of model constraints restricted to be equal across grades are CI: configural invariance, WI: weak invariance, SI: strong invariance.
**Not available.

4 Discussion

Since the factor structure of a test is directly related to the construct validity interpretation of the test at a particular point in time, the longitudinal factor structure at different points in time is crucial for the longitudinal construct validity interpretation to measure student growth. The achievement constructs of a test at a particular time, grade level or semester calendar is well studied and reported in practice for given purposes and related interpretation of test scores. Although many standardised achievement tests in large-scale assessments report test scores on a vertical scale for student growth and group achievement trends, few studies reported the longitudinal achievement construct. Many researchers are interested in whether the longitudinal achievement construct remains the same over time or shifts from time to time from content standard and vertical scaling perspectives. A few studies have focused on validation of longitudinal achievement construct using the MLGM approach, especially for studies based on CAT longitudinal data.

First, this study examined the hypothesis of shapes of latent construct across time. As shown in Table 5, the shape of the growth fits both linear and quadratic MLGM well; however, quadratic growth has slight advantages over linear growth in terms of fit statistics. This means that between two competing interpretations of longitudinal constructs across grades, quadratic growth makes more real sense because, in general, the rate of changes of student achievement across grades are not constant. From longitudinal perspective, for example, lower grades growth is always faster than higher grades growth across time for both reading and mathematics achievement tests.

Second, the present study tested the hypothesis of factorial invariance of MAP reading and mathematics tests over time. The evidence collected in the study shows that with repeated measures, the construct of both reading and mathematics remained consistent at different points in time, which supports the internal structure of MAP design for intended purposes. The evidence also suggest that there are not only configure and WI, but also SI of the longitudinal construct in MAP reading and mathematics tests across different states (except the Indiana mathematics test), which supports valid interpretations of student growth.

The current study utilised the advantages of MLGM over FLGM to investigate the longitudinal construct of achievement test. No longitudinal achievement constructs studies using MLGM have been done on CAT data across states. The major difference between FLGM and MLGM is that the FLGM uses total scores as observable variable in analysis, while the MLGM uses item cluster (or item) as observable variable. However, since the study was based on longitudinal data, students who missed one test were excluded from the sample, which might introduce sample bias.

In summary, this study underscores the importance of empirical evidence in validating longitudinal achievement constructs to support the interpretation of student growth. In particular, the study explored the feasibility of assessing the internal structure of MAP tests using CAT data. The results support consistent and reasonable interpretations of the MAP reading and mathematics tests across academic calendar years used by different states. This study carries the validation process beyond a traditional construct validation process in which validation evidence is usually collected at one point in time only, but used to support the longitudinal achievement construct for student growth. It is indeed important to investigate the longitudinal achievement construct to ensure that the same construct is measured over time for a valid interpretation of student achievement growth. We strongly recommend that achievement test publishers and users continue investigating the longitudinal achievement construct and construct invariance over time in the near future to support valid interpretations of student academic growth.

5 Limitations of the study

First, although the current study takes the advantages of MLGM over FLGM, the study could not avoid using item parcels or clustering items as indicators of the MLGM for CAT data. The effect of using aggregated indicators vs. individual items is unknown and item parcels' interpretability has not been sufficiently examined with respect to content. Second, it is well-known fact that heterogeneity in sample may bias the model estimates (Muthén, 1989). All samples used in this study are not homogeneous with respect to at least gender and ethnicity, and the impact of heterogeneity on model parameter estimates is unknown.

Acknowledgements

The authors would like to thank the two anonymous reviewers for their comments and suggestions that have greatly improved this article.

References

Akaike, H. (1987) 'Factor analysis and AIC', *Psychometrika*, Vol. 52, No. 3, pp.317–332.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, DC.

Bandalos, D.L. (2002) 'The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling', *Structural Equation Modeling*, Vol. 9, No. 1, pp.78–102.

Bandalos, D.L. and Finney, S.J. (2001) 'Item parceling issues in structural equation modeling', in Marcoulides, G.A. and Schumacker, R.E. (Eds.): *Advanced Structural Equation Modeling: New Developments and Techniques*, Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Bergman, L.R., Eklund, G. and Magnusson, D. (1991) 'Studying individual development: problems and methods', in Magnusson, D., Bergman, L.R., Rudinger, G. and Täorestad, B. (Eds.): *Matching Problems and Methods in Longitudinal Research*, pp.1–28, Cambridge University Press, Cambridge.

Betebenner, D.A. and Linn, R.L. (2010) *Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis and Accountability* [online] http://www.isbe.state.il.us/peac/pdf/growth_in_student_achieve.pdf (accessed 2 February 2012).

Bollen, K.A. and Curran, P.J. (2006) *Latent Curve Models: A Structural Equation Perspective*, Wiley, Hoboken, NJ.

Cheung, G.W. and Rensvold, R.B. (2002) 'Evaluating goodness-of-fit indexes for testing measurement invariance', *Structural Equation Modeling*, Vol. 9, No. 2, pp.233–255.

Cizek, G.J. (2005) 'Adapting testing technology to serve accountability aims: the case of vertically-moderated standard setting', *Applied Measurement in Education*, Vol. 18, No. 1, pp.1–10.

Cronbach, L.J. (1971) 'Test validation', in Thorndike, R.L. (Ed.): *Educational Measurement*, 2nd ed., American Council on Education, Washington, DC.

Doran, H.C. and Cohen, J. (2005) 'The confounding effect of linking bias on gains estimated from value-added models', in Lissitz, R.W. (Ed.): *Value-Added Models in Educations: Theory and Applications*, pp.80–104, JAM Press, Maple Grove, MN.

Drasgow, F. (1987) 'Study of the measurement bias of two standardized psychological tests', *Journal of Applied Psychology*, Vol. 72, No. 1, pp.19–29.

Ferrer, E., Balluerka, N. and Widaman, K.F. (2008) 'Factorial Invariance and the specification of second-order latent growth models', *Methodology*, Vol. 4, No. 1, pp.22–36.

Guilford, J.P. (1946) 'New standards for test evaluation', *Educational and Psychological Measurement*, Vol. 6, pp.427–439.

Hall, R.J., Snell, A.F. and Foust, M. (1999) 'Item parceling strategies in SEM: investigating the subtle effects of unmodeled secondary constructs', *Organizational Research Methods*, Vol. 2, No. 3, pp.233–256.

Hamilton, L.S., Stecher, B.M. and Yuan, K. (2008) *Standards-Based Reform in the United States: History, Research, and Future Directions*, RAND Corporation [online] http://www.rand.org/content/dam/rand/pubs/reprints/2009/RAND_RP1384.pdf (accessed 2 February 2012).

Hu, L. and Bentler, P.M. (1999) 'Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives', *Structural Equation Modeling*, Vol. 6, No. 1, pp.1–55.

Kingsbury, G.G. and Weiss, D.J. (1983) 'A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure', in Weiss, D.J. (Ed.): *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, pp.257–238, Academic Press, New York.

Linn, R.L. (1993) 'Linking results of distinct assessments', *Applied Measurement in Education*, Vol. 6, No. 1, pp.83–102.

Linn, R.L. (2001) *The Design and Evaluation of Educational Assessment and Accountability Systems*, No. CSE Technical Report 539, Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA.

Lissitz, R.L. (Ed.) (2006) *Longitudinal and Value Added Models of Student Performance*, JAM Press, Maple Grove, MN.

Little, T.D., Cunningham, W.A., Shahar, G. and Widaman, K.F. (2002) 'To parcel or not to parcel: exploring the question and weighing the merits', *Structural Equation Modeling*, Vol. 9, No. 2, pp.151–173.

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

Marsh, H.W. (2007) 'Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology', in Tenenbaum, G. and Eklund, R.C. (Eds.): *Handbook of Sport Psychology*, Third ed., pp.774–798, Wiley, New York.

Marsh, H.W., Hau, K.T. and Grayson, D. (2005) 'Goodness of fit in structural equation models', in Maydeu-Olivares, A. and McArdle, J.J. (Eds.): *Contemporary Psychometrics: A Festschrift for Roderick P. McDonald*, pp.225–340, Lawrence Erlbaum Associates, Mahwah, NJ.

Martineau, J.A., Paek, P., Keene, J. and Hirsch, T. (2007) 'Integrated, comprehensive alignment as a foundation for measuring student progress', *Educational Measurement: Issues & Practice*, Vol. 26, No. 1, pp.28–35.

McArdle, J.J. (1988) 'Dynamic but structural equation modeling of repeated measures data', in Nesselroade, J.R. and Cattell, R.B. (Eds.): *The Handbook of Multivariate Experimental Psychology*, Plenum Press, New York.

McArdle, J.J., Grimm, K.J., Hamagami, F., Bowles, R.P. and Meredith, W. (2009) 'Modeling lifespan growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement', *Psychological Methods*, Vol. 14, pp.126–149.

McDonald, R.P. (1985) *Factor Analysis and Related Methods*, Erlbaum, Hillsdale, N.J.

Meredith, W.M. (1993) 'Measurement invariance, factor analysis, and factorial invariance', *Psychometrika*, Vol. 58, No. 4, pp.525–543.

Messick, S. (1984) 'The psychology of educational measurement', *Journal of Educational Measurement*, Vol. 21, No. 3, pp.215–237.

Messick, S. (1989) 'Validity', in Linn, R.L. (Ed.): *Educational Measurement*, 3rd ed., pp.13–103, American Council on Education and Macmillan, New York.

Messick, S. (1995) *Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences*, *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES)*, Vol. II, pp.291–305, National Assessment Governing Board and National Center for Education Statistics, Washington, DC.

Mislevy, R.J. (1992) *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*, ETS Policy Information Center, Princeton, NJ.

Muthén, B. (1989) 'Latent variable modeling in heterogeneous populations. Presidential address to the Psychometric Society', *Psychometrika*, July, Vol. 54, No. 4, pp.557–585.

Muthén, B., Kaplan, D. and Hollis, M. (1987) 'On structural equation modeling with data that are not missing completely at random', *Psychometrika*, Vol. 52, No. 3, pp.431–462.

Muthén, B.O. (1991) 'Analysis of longitudinal data using latent variable models with varying parameters', in Collins, L. and Horn, J. (Eds.): *Best Methods for the Analysis of Change. Recent Advances, Unanswered Questions, Future Directions*, pp.1–17, American Psychological Association, Washington, DC.

Muthén, B.O. (1995) *Longitudinal Studies of Achievement Growth Using Latent Variable Modeling*, Technical Report, University of California, Los Angeles.

Muthén, L. and Muthén, B. (2007) *Mplus User's Guide*, 5th ed., Muthén & Muthén, Los Angeles, CA.

Nasser, F. and Wisenbaker (2003) 'A Monte Carlo study investigating the impact of item parcelling on measures of fit in confirmatory factor analysis', *Educational and Psychological Measurement*, Vol. 63, No. 5, pp.729–757.

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

Patz, J.R. (2007) *Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems*, Council of Chief State School Officers, Washington, DC.

Raftery, A.E. (1995) 'Bayesian model selection in social research (with discussion)', *Sociological Methodology 1995*, P.V. Marsden, ed., pp.111–163, American Sociological Association, Washington, DC.

Rasch, G. (1961) 'On general laws and the meaning of measurement in psychology', pp.321–334, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, IV, University of California Press, Available free from Project Euclid, Berkeley, California.

Rubin, D.B. (1987) *Multiple Imputation for Non-Response in Surveys*, John Wiley & Sons, New York.

Sayer, A.G. and Cumsille, P.P.E. (2001) 'Second-order latent growth models', in Collins, L.L.M. and Sayer, A.G. (Eds.): *New Methods for the Analysis of Change*, American Psychological Association, Washington, DC.

Schwartz, G. (1978) 'Estimating the dimension of a model', *The Annals of Statistics*, Vol. 6, No. 2, pp.461–464.

Smith, R.L. and Yen, W.M. (2006) 'Models for evaluating grade-to-grade growth', in Lissitz, R.W. (Ed.): *Longitudinal and Value Added Modeling of Student Performance*, pp.82–94, JAM Press, Maple Grove, MN.

Thissen, D. and Mislevy, R.J. (1990) 'Testing algorithms', in Wainer, H., Dorans, N., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. and Thissen, D. (Eds.): *Computerized Adaptive Testing: A Primer*, pp.103–135, Lawrence Erlbaum Associates, Hillsdale, NJ.

Tisak, J. and Meredith, W. (1990) 'Descriptive and associative developmental models', in Von Eye (Eds.): *Statistical Methods in Longitudinal Research*, Vol. 2, Academic, Boston.

U.S. Department of Education (2009) *Executive Summary, Race to the Top Assessment Program: Notice of Public Meetings and Request for Input* [online] http://ed.gov/programs/racetothetop-assessment/executive-summary.pdf (accessed 2 February 2012).

U.S. Department of Education (2012) *Family Educational Rights and Privacy Act (FERPA)* [online] http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html (accessed 2 February 2012).

Wang, S. and Harris, G. (2011) *Psychometric Evaluation of NWEA Item Calibration Procedure*, Technical Report, Northwest Evaluation Association, Portland, OR.

Wang, S. and Jiao, H. (2009) 'Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment', *Educational and Psychological Measurement*, Vol. 69, No. 5, pp.760–777.

Wang, S., McCall, M., Jiao, H. and Harris, G. (2011) 'Construct validity and measurement invariance of computerized adaptive testing: application to measures of academic progress (MAP) using confirmatory factor analysis', *Proposal for 2012 Annual Meeting of American Educational Research Association*.

Way, W. (2006) *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments*, Pearson Educational Measurement Research Report [online] http://www.education.pearsonassessments.com/NR/rdonlyres/EC965AB8-EE70-46E5-B1A5-036BE41AB899/0/RR_05_03.pdf (accessed 2 February 2012).

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S. and Mazzeo, J. (2010) *Some Considerations Related to the Use of Adaptive Testing for the Common Core Assessments* [online] http://www.ets.org/s/commonassessments/pdf/AdaptiveTesting.pdf (accessed 11 June 2010).

Widaman, K.F. and Reise, S.P. (1997) 'Exploring the measurement invariance of psychological instruments: applications in the substance use domain', in Bryant, K.J., Windle, M. and West, S.G. (Eds.): *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, pp.281–324, American Psychological Association, Washington, DC.

Williamson, G.L., Appelbaum, M. and Epanchin, A. (1991) 'Longitudinal analyses of academic achievement', *Journal of Educational Measurement*, Vol. 28, No. 1, pp.61–76.

Wise, L. (2004) *Vertically-Articulated Content Standards* [online] http://www.nciea.org/publications/RILS_LW04.pdf (accessed 5 June 2006).

Yen, W.M. (2007) 'Vertical scaling and no child left behind', in Dorans, N.J., Pommerich, M. and Holland, P.W. (Eds.): *Linking and Aligning Scores and Scales*, pp.273–283, Springer, New York, NY.

Yen, W.M. (2009) *Growth Models for the NCLB Growth Model Pilot*, ETS, Princeton, NJ.

Zhang, L.R. and Wang, S.D. (2012) 'An investigation of instructional effects on student growth in mathematics with repeated measures using computerized-adaptive test', in the Session of Technical Issues in Development and Implementation of Computerized Adaptive Test in K-12 Assessments, paper presented at the *NCME Annual Conference*, Vancouver, BC.