Multidimensional Modeling of Learning Progression-based Vertical Scales [1]

Nina Deng

deng.nina@measuredprogress.org

Louis Roussos

roussos.louis@measuredprogress.org

Lee LaFond

leelafond74@gmail.com

[1] This paper is presented at the 2017 annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Abstract

The learning progressions (LPs) are vertically aligned content across grades and lay the groundwork for a meaningful vertical scale. One open question is whether unidimensional IRT models can sufficiently capture the potential multidimensionality nature of LP-based vertical scales. This study examined the LP-specific multidimensionality of the data collected from a large-scale mathematics assessment, where tests in Grades 3-8 were vertically linked via common items between adjacent grades. Exploratory and confirmatory analyses were performed to determine the degree of multidimensionality and the extent to which the LPs contributed to the multidimensionality. A series of multidimensional models (including traditional and bifactor IRT models) were applied, and the LP-based multidimensional scores were compared with the overall unidimensional scores. The results informed practitioners the performance and utility of MIRT models with empirical data utilizing LP-related information.

Key words: multidimensional item response theory, learning progression, vertical scale, bifactor model

Introduction

A key purpose of modern assessment is the measurement of growth. For this reason, vertical scaling has an intuitive appeal as growth in student ability over time, often in terms of successive grades, is the underlying interpretation for vertical scales. Briggs (2013) states that the purpose of vertical scales is to facilitate inferences about changes in magnitude with respect to a common unit of measurement. However, Dadley and Briggs (2012) reports that despite the wide spread usage of vertical scales, virtually no state used their scales to make criterion referenced interpretations of growth. The reason Briggs provides is that most states simply do not trust the inferences provided by their assessments' vertical scales to provide meaningful interpretations of growth. Briggs (2015) indicates that there is gap between intuition and practice, that perhaps the best way to resolve this issue is through improved design.

Kolen and Brennan (2004) provide two different conceptualizations of growth in the presence of a vertical scale: "Grade to Grade" and "Domain". "Grade to Grade" is growth in content that is administered to students in adjacent grade levels. In contrast, "Domain" growth takes into consideration the entire range of content across all grades. However, both definitions assume the underlying construct being measured does not change across grades. However, is "mathematics" in grade 3 the same as in grade 8 in terms of relevant content and skill domains? Martineau (Martineau, 2004, 2005, 2006) argues that the changing domains over results in construct shift that represents a threat to the validity of growth interpretations. In addition, the basic item response theory (IRT) assumption of unidimensionality, on which most vertical scales are designed, is particularly strained over time as content and skill domains shift over grades.

This study investigated two different strategies for combating these issues in the design of a vertical scale: Learning progressions and Multidimensional Item Response Theory (MIRT). Learning progressions address the construct shift concern and MIRT addresses the dimensionality issue.

The learning progressions (LPs) are vertically aligned content across grades and lay the groundwork for a meaningful vertical scale. Where "Domain" growth assumes the same underlying construct across time, learning progressions take a finer grained approach. Confrey (Confrey, 2012)

describes learning progressions as a likely path for learning from which instruction flows. Taking the Common Core State Standards Mathematics (CCSS-M) as an example, one could frame growth as occurring within individual content standards contained within a particular domain (e.g. Geometry or Functions) rather than a single "mathematics" construct. Rather than covering all grades, these individual learning progressions may or may not be included in each grade. The measurement of growth within each these separate domains requires separate ability estimates.

One advantage of using unidimensional IRT (UIRT) models is their simplicity, however, it is an open question whether UIRT models can sufficiently capture the inherent multidimensionality nature of LPs (Li, 2006). More recent studies have investigated the multidimensional modeling of vertical scales; nevertheless, many focused on the grade-specific multidimensionality (Li & Lissitz, 2012; Boughton, Lorié & Yao, 2005), not content-based LP-specific. Given the lack of literature in the field and the increasing popularity of LP-based vertical scales in K-12 assessment, it is of particular interest to investigate the multidimensionality nature of LPs, and whether multidimensional IRT (MIRT) models can be efficiently and usefully employed in LP-based vertical scales.

Item Response Theory Model

Multidimensional Item Response Theory Model

The standard item response theory IRT models assume unidimensionality, and that the item-ability relationship is adequately modeled by a single latent trait. Reckase (2009), however, suggests that examinees are likely to bring more than one ability to bear when responding to a particular test item, and that unidimensional item response theory (UIRT) overlooks this particular aspect. He proposes that a multidimensional item response theory (MIRT) model, where multiple abilities are taken into account, may provide a more realistic fit.

In MIRT, instead of a single latent trait θ there is a $\boldsymbol{\theta}$-vector which is a vector of latent trait parameters describing the location of the examinee in a multidimensional space. Reckase (2009) defines the multidimensional extension of the 3PL model as:

$$P(\boldsymbol{\theta}) = c_i + (1 - c_i)\frac{\exp[1.7(\boldsymbol{a_i\theta'} + d_i)]}{1+\exp[1.7(\boldsymbol{a_i\theta'} + d_i)]}. \quad (1)$$

where parameter c_i is the pseudo-guessing parameter. Recall that in the 3PL model the exponential term was $a_i(\theta - b_i)$, and through distribution this becomes $a_i\theta - a_ib_i$. This can be changed to a slope-intercept form $a_i\theta - d_i$ where the intercept term a_ib_i is replaced by d_i. Also, the $\boldsymbol{a_i\theta'} + d_i$ term can be expanded to:

$$\boldsymbol{a_i\theta'} + d_i = a_{i1}\theta_1 + a_{i2}\theta_2 + \cdots + a_{im}\theta_m + d_i, \quad (2)$$

where $\boldsymbol{a_i}$ describes a discrimination/slope vector and the d_i is an intercept parameter.

Note that in MIRT, while the unidimensionality assumption is relaxed, versions of the local independence and monotonicity assumption still hold. MIRT models assume that the probability of correctly responding to an item increases as any element in the $\boldsymbol{\theta}$-vector increases. In addition, examinees are assumed to respond to each item as an independent event. In other words, the probability of correctly responding to a particular item depends solely on that examinee's $\boldsymbol{\theta}$-vector and the parameters of that item. In MIRT, this local independence assumption leads to:

$$P(U_1 = 1, U_2 = 1, \ldots, U_n = 1|\boldsymbol{\theta}) =$$

$$P(U_1 = u_1|\boldsymbol{\theta})P(U_2 = u_2|\boldsymbol{\theta}) \ldots P(U_n = u_n|\boldsymbol{\theta}). \quad (3)$$

The above states that the probability of an examinee of a given ability vector earning a particular summed-score is equal to the product of the probabilities of the individual responses on an *n* item test.

Also, it is important to note that while MIRT is a much more complicated model than UIRT, it is still an idealization of reality. Reckase (2009) notes that MIRT only gives an approximation to the relationship between a person's capabilities and the responses to test items. Despite MIRT's apparent complexities, current models are still relatively simply, and as more is learned concerning the item-ability relationship even more sophisticated models will be developed.

Bifactor Model

Full-information item bifactor analysis (Gibbons, et al., 2007; Gibbons & Hedeker, 1992) allows for multidimensionality between item types. A sample item bifactor measurement structure could have the following factor pattern:

$$\begin{pmatrix} a_{10} & a_{11} & 0 \\ a_{20} & a_{21} & 0 \\ a_{30} & 0 & a_{32} \\ a_{40} & 0 & a_{42} \\ a_{50} & 0 & a_{52} \end{pmatrix}$$

The first subscript denotes the item and the second corresponds to a dimension. For example, the above pattern could represent the passage-related factor structure for 5 items with 2 nested within one passage (y_1 and y_2) and 3 in another (y_3, y_4, and y_5). The first column represents a general dimension while the second and third are dimensions specific to each of the passages. Thus, this model controls for the format effect that potentially violates the unidimensionality assumptions in the UIRT model.

The bifactor IRT model (Cai et al., 2011) is an extension of the standard UIRT model. For dichotomous items the bifactor model with general factor θ_0 and specific factor θ_s is:

$$P(\theta_0, \theta_s) = c_i + (1 - c_i)\frac{\exp[1.7(a_{0i}\theta_0 + a_{si}\theta_s + d_i)]}{1+\exp[1.7(a_{0i}\theta_0 + a_{si}\theta_s + d_i)]}. \quad (4)$$

Above, c_i is the item pseudo-guessing parameter, d_i is the item intercept, a_0 is the item slope on the general factor, and a_{si} is the item slope on the specific factor *s*. Note that the item slopes are similar in interpretation to discrimination in the IRT model, but are specific to each factor. Comparing the bifactor equation in Equation 4 to the full multidimensional model in Equation 1 shows that the bifactor model is a specific case with two ability parameters.

Methods

Data

Data were collected from a large-scale Mathematics achievement assessment, where tests were vertically linked via common items between adjacent grades. There were 252 items in total across the six grades in Grades 3-8, each was specified being associated with one of ten learning progressions identified by the content specialists: Number and Number Systems (Num), Fractions (Fra), Operations (Ope), Algebraic Thinking (Alg), Functions (Fun), Measurement (Mea), Geometric Properties (Geo), Spatial geometry (Spa), Statistics and Data (Sta), Probability (Pro). Each grade had a single form consisting of from 35 to 44 core items (items provided to every student and counted into students' total scores), and 18 to 39 matrix items (items provided to a subset of students and not counted into students' total scores). The number of common items shared between adjacent grades ranged 37-39. The numbers of items broken down by the LP in each grade can be found in Table 1 and Table 2.

77,988 students in total across the six grades took the tests. Due to the very large sample size which caused unnecessary complexities in computation and IRT modeling, a random sample of 2,000 students were selected per grade (12,000 in total for six grades) for the purpose of analyses in this paper.

Dimensionality analysis

Dimensionality analysis was conducted to explore the degree of multidimensionality and the extent to which the LPs contributed to the multidimensionality. A number of various analyses were conducted. The eigenvalues were computed using polychoric correlations as appropriate for ordinal data, and results were evaluated in terms of scree plot and factor loadings. Exploratory linear and non-linear factor analyses were used to examine the factorial structure, using both investigative and targeted approaches. The nonparametric methods, DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999), were used to detect whether the unidimensionality was violated, based on the covariance between each pair of items conditioned on the expected total score. The analyses were conducted within a single grade and then expanded to multiple vertically-linked grades.

IRT models

A number of LP-specific multidimensional IRT models were applied to assess the suitability of these models. The confirmatory multidimensional IRT models were fitted so that, each item loaded on one LP-specific dimension to which they were expected to be related. The correlations between the different LP-specific dimensions were freely estimated. We started with the 10-dimensional model which included the ten LPs originally identified by the content specialists. Then we evaluated the mode utility, convergence, and parameter estimates, and experimented on other MIRT models by combining some LPs if necessary. In the end, we calibrated all the items concurrently using the standard unidimensional IRT (UIRT) models for the comparison purpose.

Different from the traditional MIRT models, the bifactor model specifies for each item nonzero loadings on two and only two factors, one general and one group factor, achieving the so-called *bifactor* model. The general and group factors are constrained to be uncorrelated so that they explain the total variance independently. The items are considered essentially unidimensional to the extent that the majority of item variance is explained by the general factor, despite the presence of group factors. The bifactor model was of interest given that it has been used to examine the unidimensionality of complex concepts. The bifactor model was applied to the dataset in the way that, the LP-specific factors in the multidimensional models described above were retained as the *group* factors, and an overall *general* factor ("math" dimension) was imposed on all the items. A number of bifactor models were experimented, each corresponding to one of the above multidimensional models.

Evaluation criteria

The results of different MIRT, bifactor, and UIRT models were compared and evaluated in terms of model utility, convergence, model fit statistics, ability estimates, and between-grade difference and growth. The differences were evaluated both on overall summative and LP-specific dimension scores.

Results

Dimensionality Analyses

DIMTEST and DETECT were first applied to the data within each grade. The data were split into a training sample and a cross-validation sample to avoid the error due to capitalization on chance. The DIMTEST null hypothesis was rejected at a significance level of 0.01 for every dataset, which was not surprising because strict unidimensionality is an idealization that rarely holds exactly for a given dataset. DETECT was then used to estimate the effect size of the violations of local independence found by DIMTEST. Table 3 displays the multidimensional effect size estimates from DETECT for each grade. The DETECT values (0.2 - 0.4) indicate from weak to moderate multidimensionality across the grades. Next, the analyses were applied to the across-grade dataset – a supper sparse matrix dataset combining the response data across the six grades. However, the program failed to work due to a lot of missing data – the design that the vertical linking items were put in the matrix slots thus only a small subset of students were given to those items.

The above issue of missing data by design caused similar computational problems in the eigenvalue and EFA analyses. Specifically, the covariance matrix of pairwise items could not be estimated due to the empty cells. Therefore, the average correlation of non-empty cells was computed (about 0.2) and imputed in the covariance matrix to overcome the computational issue. After imputation, the first eigenvalue was 51.5, explaining 20% of total variance and 5 times as large as the second (eigenvalue of 10.7 explaining 4% of total variance). The scree plot had a clear elbow at the second factor, suggesting a dominant first factor. There were some negative eigenvalues with insignificant magnitudes, indicating too many variables were highly correlated.

Finally, the non-linear factor analyses were conducted using the software flexMIRT (Cai, 2012) to explore the factorial structure of the dataset. Both “blind” and target rotations were used. Although there was an indication of some degree of multidimensionality, the factor loadings did not present any discernible relationship between the multidimensionality and specified LPs. Specifically, the items with a salient factor loading on each factor (e.g., factor loading > 0.4) were associated with multiple LPs. This suggests that the learning progression factor did not sufficiently explain the variance of the data, and that a better understanding of the underlying factors of the multidimensionality warrants further investigation.

Some possible factors include the item type, item complexity level, instruction opportunity, and grade-specific factors, which however are beyond the learning progression factor and thus are not the focus of this study.

IRT models

Despite that the dimensionality analyses did not present a discernible relationship between LP-specific factors and multidimensionality, the learning progressions were conceived as different dimensions from a content perspective. Therefore it was still of interest to find out how the LP-specific MIRT model fits the data. We started with the 10-dimensional model in which each item loaded on one of the ten LPs originally identified by the content specialists. Not surprisingly, the program failed to produce an output due to the highly demanding computation requirements. In the next, we evaluated the standards of the ten LPs against the common core math standards, and collapsed them into 7 LPs. However, the computation problem persisted. After that, we re-evaluated and further combined them into 4 LPs based on content consideration. The 4-dimensional model got converged, however, we found that some item a-parameters were negative and that two dimensions were almost perfectly correlated. Finally, after evaluating the results, we further combined the two highly-correlated dimensions into one dimension, and fitted a 3-dimensional model. The final 3-dimensional model was based on three combined super learning progressions, which are one combined LP of Algebraic Thinking, Fractions, Functions, Number and Operations (called D1), another combined LP of Geometric Properties, Measurement and Spatial Geometry (called D2), and the third combined LP of Probability, Statistics and Data (called D3), respectively. Table 4 provides a summary of the specification of the different MIRT models we investigated, and the detailed information of learning progressions in each model.

After that, four bifactor models were fitted, each applying one general factor to one of the four multidimensional models described above, that is, the 10-D, 7-D, 4-D, and 3-D models, respectively. The LP specific factors in the multidimensional models were retained as group factors in the bifactor models. None of the four bifactor models converged in our experimental runs, suggesting that the LP-specific

bifactor model specification did not help in revealing the factorial structure of the dataset. This possibly was due to the reason that the underlying factors of multidimensionality were not LP-specific, therefore it was problematic for the general and LP-based group factors to be constrained uncorrelated in the bifactor models. In other words, after the general "math" dimension was extracted, there probably was not much information left over for LP specific dimensions given the lack of LP specific multidimensionality, as suggested in the dimensionality analyses. On the contrary, the traditional MIRT model allowed the LP-specific factors to be correlated which seemed to be a better fit than the bifactor simple structure.

Between-grade growth

The between-grade difference were evaluated when the tests were vertically scaled using the unidimensional versus LP-specific multidimensional models. The expected a posteriori (EAP) abilities were estimated for the 12,000 students across grades 3-8. The mean and standard deviations of the theta scores were summarized (Table 5 and Figure 1). As can be seen, there was a gradual increase of ability moving across Grades 3-6, however, the differences became more flat or even slightly negative between Grades 6 and 8. The growth deceleration (between-grade growth that decreases over time) has been observed in math tests by Dadey & Briggs (2012). Besides, we found a similar trend of the between-grade differences in the unidimensional and LP-specific multidimensional models.

Since the unidimensional and multidimensional theta scores were not linked to a common scale, therefore their scores and absolute differences were not directly comparable. The standardized mean differences (Cohen, 1988) and Yen's separation index (Yen, 1986) of the theta scores between grades were computed. The two types of statistics yielded highly similar values. Therefore only the effect sizes of standardized mean differences were reported and summarized in Table 6 and Figure 2.

As can been seen in the results, in general the between-grade differences were quite small (effect size < 0.2), with the largest difference occurring between Grades 3 and 4, and the smallest between Grades 6 and 7. The LP-specific multidimensional scores had a similar trajectory across the grades compared to their unidimensional scores. More specifically, the third dimensional scores (combined LP of

Probability, Statistics and Data) were the most similar to the unidimensional scores. The first dimensional scores (combined LP of Algebraic Thinking, Fractions, Functions, Number and Operations), and the second dimensional scores (combined LP of Geometric Properties, Measurement and Spatial Geometry), were more slightly different from the unidimensional scores.

The LP-specific multidimensional theta scores were also compared to the theta score distribution derived from the unidimensional separate linking, in which each grade was calibrated separately and linked through common items between adjacent grades. The common items were evaluated based on their content and statistical properties. The results from separate linking had the between-grade effect sizes ranging from 0.35 to 0.17, with the largest effect size between Grades 3 and 4 (0.35), intermediate between Grades 4 and 5 (0.29), and the smallest between Grades 5-8 (0.17). Comparatively speaking, the separate linking produced larger between-grade differences and effect sizes than that of the concurrent linking, both unidimensionally and multidimensionally. Besides, no negative mean difference between Grades 6 and 7 was detected in the separate linking as was in the concurrent linking. Although there could be various reasons, at least two might contribute to explaining the differences: 1) different linking approaches. The separate linking started with the independent IRT calibration within each grade, and then linking the two adjacent grades via their common items, finally linking all the grades together onto a common scale in a chain-linking fashion; whereas the concurrent linking calibrated all the items across the grades together. The separate linking approach can be helpful in securing a gradually increased and ordered list of means of ability scores across the grades; 2) different vertical-linking items. A non-trivial proportion of vertical linking items were flagged and excluded from the linking set in the separate linking. The off-grade matrix items and about 32% of on-grade vertical linking items were removed from linking across the grades. The different linking item set is another potential important factor in resulting in the different results.

Conclusion

In recent years there has been increased interest in examining achievement growth on the basis of large-scale achievement tests. In order to compare the scores from different tests and grades, it is necessary for them to be reported on a common meaningful scale. The learning progressions are vertically aligned content across grades and lay the groundwork for a meaningful vertical scale. However, all of the state-level vertical scales currently used in practice assume that a single construct is measured within and across grades for each content area which consists of multiple learning progressions. There is suspect that the unidimensional model may not sufficiently capture the multidimensionality nature of the vertical scales. Besides, there have been few studies in literature evaluating the learning progression (LP)-specific multidimensionality of vertical scales.

The purposes of this paper were: 1) to examine the LP-specific multidimensionality on the data collected from a large-scale mathematics assessment where tests were vertically linked, 2) to explore the utility of multidimensional IRT models, and 3) to compare the LP-specific multidimensional vertical scores with the unidimensional vertical scores.

The dimensionality analyses suggested that the data presented a dominant factor with 2-3 secondary factors. Further investigation of the multidimensionality did not suggest a discernable relationship between the multidimensionality and the LP-specific factors. Nevertheless, several LP-based multidimensional and bifactor models were fitted ahead to the data. The fitting results showed that a 3-LP based MIRT model had the model convergence as well as provided reasonable parameter estimates. None of the bifactor models had the model convergence, presumably due to a model misspecification. The LP-specific multidimensional scores had very similar means, standard deviations, and growth across the grades compared to the concurrent unidimensional scores. There was no substantial difference or distortion detected in the concurrent unidimensional scores compared to the LP-specific multidimensional scores.

There are several limitations of the study which are noteworthy. One primary limitation is the missing data by design. Given that many vertical linking items were placed in the matrix slots, there were many pairwise items in the dataset which had non-overlapping student responses, causing major issues in

the exploratory factor and dimensionality analyses. The current study used the average of nonempty cell correlations to impute the missing cells in the correlation matrix. However, more sophisticated methods should be explored to better understand the dimensionality of the dataset. Another primary limitation is the inclusion of vertical linking items. In the concurrent linking, all the common items were used as the vertical linking items in the calibration, including both the on-grade core items and off-grade matrix items. A closer look at the item content suggested that some of the off-grade items may not be suitable for the lower grade students, e.g., the content had not yet been instructed to the students. Besides, the stability of the common items across grades is unknown, which is potentially problematic given that a large number of linking items was flagged in the separate linking. Finally, there are some methodological issues which are worthy of investigations in the future studies. First, there are some different ways of concurrent linking which should be evaluated, in addition to calibrating data Grades 3 -8. For example, dividing the six grades into two grade spans, Grades 3-5 and Grades 6-8, might be of interest given that more content and LPs were shared within each span. Another example is to concurrently calibrate adjacent grades given that the link was stronger between those grades. Secondly, it is desired that the multiple group IRT analyses be investigated in the concurrent calibration given that students across the grades are more likely nonequivalent groups.

Developing a meaningful and interpretable vertical scale is a challenging endeavor. Given the recent increasing popularity of constructing the learning progression-based vertical scales in the K-12 assessment, it is important to understand the potential multidimensionality of such scales and to have accurate interpretations of growth. It was the hope that this study will inform practitioners about the performance of LP-specific multidimensional models with some empirical data utilizing the LP-related information, and will encourage more future studies under these contexts.

References

Boughton, K. A., Lorié, W., & Yao, L. (2005). *A multidimensional multi-group IRT model for vertical scales with complex test structure: An empirical evaluation of student growth using real data*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204–226.

Briggs, D. C., & Dadey, N. (2015). Making sense of common test items that do not get easier over time: Implications for vertical scale designs. *Educational Assessment*, *20*(1), 1–22.

Cai, L. (2012). *flexMIRT* [Computer program]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221-248.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Confrey, J. (2012). Better measurement of higher cognitive processes through learning trajectories and diagnostic assessments in mathematics: The challenge in adolescence. In V. F. Reyna, S. B. Chapman, M. R. Dougherty,& J. Confrey (Eds.), *The adolescent brain: Learning, reasoning, and decision making* (pp. 155–182). Washington, D.C.: American Psychological Association.

Dadey, N.& Briggs, D. C. (2012). A Meta-Analysis of Growth Trends from Vertically Scaled Assessments. *Practical Assessment, Research & Evaluation, 17*(14).

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D., Frank, E., Grochocinski, V., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer Verlag.

Li, T. (2006). *The effect of dimensionality on vertical scaling*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Li, Y. & Lissitz, R. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36*(1), 3-20.

Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models* (Unpublished Dissertation). Michigan State University, East Lansing, MI.

Martineau, J. A. (2005). *Un-distorting measures of growth: Alternatives to traditional vertical scales*. Paper presented at the 35th Annual Conference of the Council of Chief State School Officers.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35–62.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W., Froelich, A., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-376). New York, NY: Springer-Verlag.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23* (4), 299-325.

Zhang, J., & Stout, W. (1999).The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.

Table 1. Numbers of on-grade core items and off-grade matrix items by learning progression in Grades 3-8

Learning Progression		On-grade Core Item						Off-grade Matrix Item					
		G3	G4	G5	G6	G7	G8	G3	G4	G5	G6	G7	G8
Algebraic Thinking	ALG	9	6	6	8	10	5	4	6	7	12	8	5
Fractions	FRA	7	9	9	2	2	0	1	7	7	4	1	1
Functions	FUN	0	2	4	10	7	12	2	1	5	6	13	3
Geometric Properties	GEO	8	6	9	7	4	2	7	11	6	5	4	1
Measurement	MEA	4	4	4	0	0	0	6	4	3	1	0	0
Number and Number Systems	NUM	3	7	3	5	0	4	9	2	7	1	2	0
Operations	OPE	2	3	3	1	4	3	1	3	2	1	0	1
Probability	PRO	0	0	0	0	5	0	0	0	0	2	0	3
Spatial Geometry	SPA	0	0	2	2	0	8	0	1	2	0	3	0
Statistics and Data	STA	2	2	2	8	6	10	0	2	0	5	7	4
	(total)	35	39	42	43	38	44	30	37	39	37	38	18

Table 2. Numbers of vertical linking items between adjacent grades by learning progression

Learning Progression		G3/G4	G4/G5	G5/G6	G6/G7	G7/G8
Algebraic Thinking	ALG	4	5	10	11	10
Fractions	FRA	6	8	4	2	1
Functions	FUN	2	3	6	8	11
Geometric Properties	GEO	9	9	8	4	1
Measurement	MEA	6	4	1	0	0
Number and Number Systems	NUM	7	4	5	2	0
Operations	OPE	3	3	0	1	1
Probability	PRO	0	0	0	2	3
Spatial Geometry	SPA	0	1	2	0	3
Statistics and Data	STA	0	2	2	7	7
	(total)	37	39	38	37	37

Table 3. DETECT multidimensionality effect size by grade

Grade	Effect Size
3	0.19
4	0.18
5	0.29
6	0.18
7	0.24
8	0.19
Average	0.21

Table 4. Summary of the MIRT model specification

Learning Progression		# of items	Dimension indicator			
			10-D	7-D	4-D	3-D
Algebraic Thinking	ALG	46	1	6	2	1
Fractions	FRA	29	2	1	1	1
Functions	FUN	35	3	2	2	1
Geometric Properties	GEO	39	4	3	3	2
Measurement	MEA	15	5	4	3	2
Number and Number Systems	NUM	25	6	5	1	1
Operations	OPE	16	7	6	2	1
Probability	PRO	5	8	7	4	3
Spatial Geometry	SPA	12	9	3	3	2
Statistics and Data	STA	30	10	7	4	3

Table 5. Mean and standard deviation (in parenthesis) of theta scores across grades in different IRT models

	G3	G4	G5	G6	G7	G8
UIRT	-0.21	-0.06	0.00	0.06	0.06	0.14
	(0.91)	(0.92)	(0.94)	(0.93)	(0.95)	(0.97)
MIRT-D1	-0.18	-0.01	0.02	0.10	0.05	0.10
	(0.99)	(0.99)	(0.97)	(0.95)	(0.94)	(1.01)
MIRT-D2	-0.19	-0.03	0.05	0.08	0.03	0.08
	(0.96)	(0.94)	(0.97)	(0.89)	(0.92)	(0.87)
MIRT-D3	-0.14	-0.01	0.02	0.06	0.05	0.10
	(0.84)	(0.77)	(0.79)	(0.87)	(0.97)	(0.97)

Table 6. Effect size of between-grade mean score difference in different IRT models

	G3/G4	G4/G5	G5/G6	G6/G7	G7/G8
UIRT	0.17	0.06	0.07	0.00	0.09
MIRT-D1	0.18	0.03	0.07	-0.05	0.05
MIRT-D2	0.17	0.08	0.04	-0.06	0.06
MIRT-D3	0.15	0.05	0.04	-0.01	0.05

Figure 1. Mean of theta scores across Grades 3 -8 in different IRT models

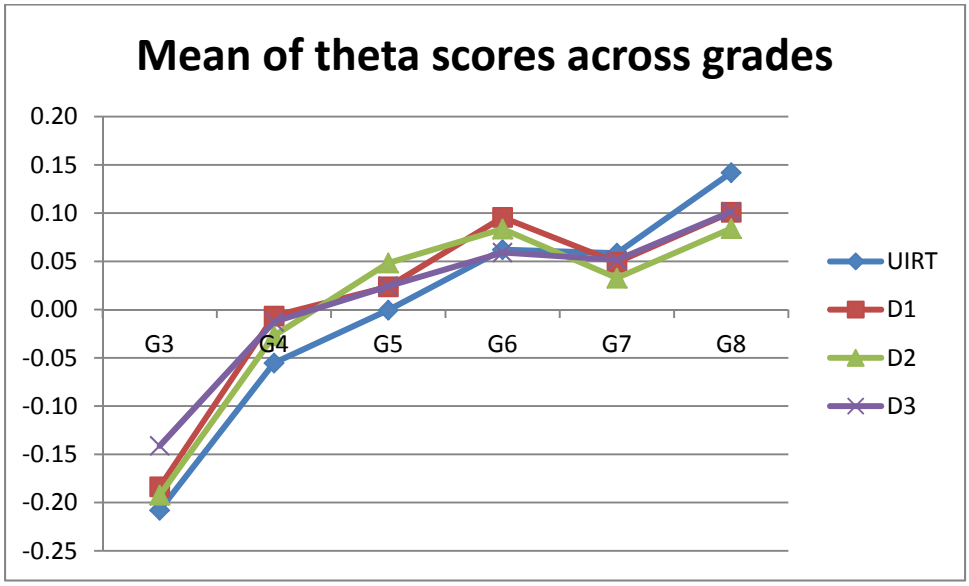

Figure 2. Effect size of between-grade mean score difference in different IRT models