A Multidimensional Multi-Group IRT Model for Vertical Scales with Complex Test Structure: An Empirical Evaluation of Student Growth using Real Data

Keith A. Boughton
William Lorié
Lihua Yao

CTB/McGraw-Hill

Paper presented at the 2005 annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.
A Multidimensional Multi-Group IRT Model for Vertical Scales with Complex Test Structure: An Empirical Evaluation of Student Growth using Real Data

Perspectives/Theoretical Framework

Vertical or developmental test scales are metrics that span more than one point in time along a growth trajectory. The most common types of vertical scales in psychometric research and practice are those that measure year-to-year growth for individual examinees or student cohorts in a particular content area. Currently, several nationally-normed tests (e.g., CTB/McGraw-Hill’s TerraNova, Harcourt Educational Measurement’s Stanford 10, and the Iowa Tests of Basic Skills) and custom state built instruments are used to track student growth on a developmental continuum.

Research on vertical scales has advanced in at least three directions in recent years. Although the issue has been extensively studied, it remains an open question whether or not a unidimensional IRT model can be successfully applied to a vertical scale that may not be strictly unidimensional in nature (Camilli, 1999; Camilli, Yamamoto, and Ming-mei, 1993; Patz, Yao, Chia, Lewis, Hoskins, 2003; Yao and Mao, 2004; Yen, 1985; Yen and Burket, 1997). A second but related issue is whether or not it is better to estimate IRT parameters for a vertical scale using a concurrent calibration across all grades or separately by grade, with the scale built up through a grade-by-grade linking procedure (e.g., Stocking and Lord, 1983), when, by design, the scales are inherently multidimensional (Beguin and Hanson, 2001; Beguin, Hanson, and Glas, 2001; Hanson and Beguin, 1999; Patz, et al., 2003; Yao and Mao, 2004). A third direction of research on vertical scaling is whether a multidimensional model can be realistically employed in place of a unidimensional model within an operational vertical scaling context (Patz, et al., 2003; Yao and Mao, 2004). Harris, Hendrickson, Tong, Shin, & Shyu (2004) question whether or not test design/construction should be based on a preexisting growth model, as done with the hierarchical multi-group IRT model for growth proposed by Patz, et al. (2003), or whether it should be developed based on an empirical evaluation from an assessment built to a particular curriculum.

The most recent research on developmental scales addresses some of these issues combined by exploring the performance of separate versus concurrent calibration in the construction of vertical scales in both unidimensional and multidimensional contexts (Yao and
Mao, 2004). Yao and Mao found that when a multidimensional model truly underlies test performance, separate calibration was better than concurrent calibration when applying a unidimensional model, but the reverse was true when applying a multidimensional model to that same data. However, as with most research of this type, the issue of concurrent versus separate calibration in the presence of multidimensionality will not be resolved through simulation studies alone.

The extent to which multidimensional vertical scales can be usefully employed in operational testing programs depends on whether students truly develop over time as suggested by the results of these models. A vertical scale implies a learning model; however, is it possible to choose one vertical scaling method over another, given that the very nature of true growth is unknown (Harris et al., 2004)? Addressing this question, and others like it, requires more research with real data applications utilizing both unidimensional and multidimensional modeling, together with substantive considerations of growth.

**Objectives /Purpose**

The central question of this study is whether actual student growth trajectories follow those predicted by a unidimensional or multidimensional approach. To this end, this study proposes the use of a Bayesian multi-group multidimensional IRT model in order to determine the dimensional structure of a vertically scaled achievement test. A comparison of the calibration methods will be explored to determine the best representation of the data. Data collected from a longitudinal study across multiple years using the same vertical scale, thus tracking individual student growth, will be used to aid in assessing how well the proposed models fit actual student growth.

**Methods**

The vertical scales that were used in this study were developed by using a common items design, for both vertically linking Grades 5, 6, 7, 8, 9, and 10 in Mathematics and horizontally across three years (i.e., 2002, 2003, and 2004), with 2002 as the base year. In addition, three cohorts (sample of 1500 students per cohort) were tracked across three years.
Each grade had between 50 to 60 multiple-choice (MC) items and 12 to 15 constructed-response (CR) items per form.

The current study used a Markov chain Monte Carlo (MCMC) algorithm to estimate the multidimensional multi-group IRT model parameters for both item types, as implemented in the Bayesian Multidimensional IRT software program (BMIRT; Yao, 2003). A multidimensional compensatory three-parameter logistic model (M3PL; Reckase, 1985; Yao, 2003) will be applied to the dichotomous items. The probability of a correct response for an examinee with abilities, \( \vec{\theta} \), using the M3PL is:

\[
P_{i,j} = P(X_{i,j} = 1 \mid \vec{\theta}, \vec{\beta}_j) = \frac{1 - \beta_{3,j}}{1 + e^{-\beta_{2,j} \vec{\theta}^T \vec{\beta}_j}},
\]

where \( X_{i,j} \) is a random variable with a 0 or 1 score. \( \vec{\theta} \), is a 1 by \( D \) vector, with \( D \) being the number of ability dimensions. \( \beta_{1,j} \) is the item difficulty parameter, with \( \beta_{2,j} \) as the discrimination parameter in a 1 by \( D \) vector for the \( j \)th item, and \( \beta_{3,j} \) as the guessing parameter. \( \vec{\beta}_{2,j} \Theta \vec{\theta}_i \), is a dot product, which is a scalar. \( \vec{\beta}_j \equiv \left( \vec{\beta}_{2,j}, \beta_{1,j}, \beta_{3,j} \right) \) are the set of parameters for item \( j \).

A multidimensional compensatory generalized partial credit model (MGPCM; Muraki, Carlson, & Bolt, 1996; Yao, 2003) will be used for the polytomous items, \( j \). The probability of a correct response \( k \) for an examinee with abilities, \( \vec{\theta} \), for the MGPCM is:

\[
P_{i,j,k} = P(X_{i,j} = k - 1 \mid \vec{\theta}, \vec{\beta}_j) = \frac{e^{(k-1) \vec{\beta}_{2,j} \vec{\theta}_i} \sum_m \beta_{5,j}}{ \sum_{m=1}^{K_j} e^{m \vec{\beta}_{2,j} \vec{\theta}_i} \sum_m \beta_{5,j}},
\]

where \( \vec{\beta}_j = (\vec{\beta}_{2,j}, \beta_{5,j}, ..., \beta_{8,j}) \) are the parameters for the \( j \)th item, \( \beta_{5,j} = 0 \), and \( k=1,2,...,K_j \), where \( K_j \) is the number of response categories for the \( j \)th item.
This research exploits the use of a multidimensional model in order to allow for dimensional shifts while moving sequentially from grade to grade. However, a single score is could also be used to compare multidimensional results with the unidimensional outcomes and can be found by using a recursion formula suggested by Lord & Wingersky (1984). The distribution of score $x$ with a sample of $N$ examinees and $J$ items is:

$$ f(x) = \frac{1}{N} \sum_{i=1}^{N} f_j(x|\theta_i), $$

where $f_j(x|\theta_i)$ is the distribution of $x$ for examinee with ability $\theta_i$ and $x=0,1,2,\ldots,k$. $k$ is the total score of the test, thus producing a single score distribution for each test (Yao & Mao, 2004).

Since several years of data on a single vertical scale were available (i.e., the same students were tracked over consecutive years), item parameters derived from the multidimensional model established in the baseline year was used to compute scores for students in subsequent years. The yearly growth of the baseline year cohorts was compared to the growth trajectory suggested by the vertical scale in order to assess whether student growth follows the model.

**Within-Year Study**

For the within-year vertical scaling study, a random sample of 1500 students was drawn from each of the grades 5 through 10. This was done once for each year in which within-year profiles were to be studied – 2002, 2003, and 2004.

For each of these years, an item map showing shared items between grades was used to construct calibration input files. Each grade had a single mathematics form consisting of from 50 to 60 multiple choice items and from 12 to 15 constructed response items. The number of linking items shared between grades ranged from 15 to 20. Two types of calibrations were conducted using the models described above, the first was unidimensional, and the second was multidimensional.

For the multidimensional item parameters, the common items loaded on two dimensions (i.e., adjacent grades) and non-common items (on-level only) were constrained to load on the on-level dimension only. For the multidimensional ability parameters, Grade 5 students would have
a grade 5 theta and a grade 6 theta, while the Grade 6 students had a grade 5 theta, grade 6 theta, and a grade seven theta. The dimensions have been named according to the central grade in which they appear. For example, DIM5 is the dimension on which 5th and 6th graders can be tracked, and DIM6 is the dimension on which 5th, 6th, and 7th graders can be tracked.

The IRT models applied yielded six profiles of student ability across 6 different cohorts. For simplicity, only the first three cohorts were studied in depth for the within-year profiles of student performance. In the unidimensional case, the 8th and 9th grades were added for the 2003 and 2004 within-year analyses, to track the (samples of) students from the three cohorts who were in grades 5, 6, and 7, respectively, in 2002.

Although clearly desirable, it is not a guarantee that scale score means increase from one year to the next, when a vertical scale is being set. This is due to cohort differences, which might have their basis in differential instruction or other factors. For example, in a particular year of scaling, the 7th grade cohort might not have performed, overall, as well as the 6th grade cohort.

For the three grades we chose to study, it was the case that across cohorts increasing in grade, means on both unidimensional and multidimensional scores increased, as will be shown in the results.

Across-Year Study

Records for 1500 students who increased in grade from 2002 to 2004 and whose records could be matched, were selected for the cross-year study. There were three cohorts studied: The cohort who was in the 5th in 2002 (2002 G5 cohort), the cohort who was in the 6th in the same year (2002 G6 cohort), and the cohort in the 7th grade that year (2002 G7 cohort).

Scores were established for the unidimensional scale defined by the 2002 data, and for the multidimensional scale define by data for that same year. This required all parameters to be equated to that baseline year. As will be shown in the section on findings, these data produce trend lines for each of the cohorts.
Results

Within-Year Study: Unidimensional

Data were calibrated using the BMIRT software program (Yao, 2003) to obtain parameters on a scale encompassing grades 5 through 10, however, grade 10 will not be presented, as the three cohorts that were followed in this study, were not yet in grade 10. The mean scale scores for each grade in the 2002 vertical scaling are shown in Figure 1a. Student scale score means showed general progression from grade to grade, but growth was neither uniform nor guaranteed. A number of factors can account for this – most obviously, cohort differences. The growth profiles for the 2003 (Figure 1b) and 2004 (Figure 1c) scaling effort provide evidence that cohort differences explain the declines in grades 8 and 9 for the 2002 scaling. In 2003, the decline is observed only in grade 9, and all declines disappear by 2004, when the cohort which is in grade 5 in 2002 has advanced two years. The ordinal relationship of the means for the 2002 grade 5, 6, and 7 cohorts is maintained as they advance grades (grade 6, 7, and 8 in 2003 and grades 7, 8, and 9 and 2004). The 2002 grade 7 is a higher performing group, as seen across the years in Figures 1a, b, and c (i.e., grade 8 in 2003 and grade 9 in 2004).

![Figure 1a. Student means for 2002 vertical scale across grades 5 to 9.](image-url)
Within-Year Study: Multidimensional

It is instructive to compare the within-year unidimensional profiles with those obtained from multidimensional calibration. In the figures below, the same data were calibrated using a six-dimensional model in BMIRT. The dimensions were restricted to a maximum of three grades each and named after the central grade.
In 2002, all dimensions display the expected ordinal growth relationships, except that the on-grade groups for DIM8, DIM9, and DIM10 show a decline, as shown in Figure 2a. This corresponds to the observed 2002 declines in grades 8 and 9. However, the multidimensional profiles for 2003 (Figure 2b) and 2004 (Figure 2c) are not always aligned with their corresponding unidimensional growth patterns. In 2003, for example, the decline in grade 8 DIM8 remains, even though it is not as pronounced. And in 2004, the decline has disappeared in grade 8 DIM8, and for the most part in grade 9 DIM9, but two unexpected declines materialize – one in grade 7 DIM6 and another smaller one in grade 8 DIM7. There is no evidence in the unidimensional profiles to suggest that there should be declines here.

One conclusion that might be reached from these findings is that while a unidimensional vertical model might predict year-to-year growth, some of the individual dimensions of a multidimensional model might not. The higher scale score means for the 7th graders in DIM7 (compared to the 6th graders) might outweigh the lower scale score means for the 7th graders on DIM8, resulting in an overall observed score growth between these two grades, for a unidimensional model.

![Student Means on 2002 Multidimensional Vertical Scale](image)

Figure 2a. Student means for 2002 multidimensional vertical scale across grades 5 to 10.
Across-Year Study: Unidimensional Growth

Three cohorts were tracked from year to year – the cohorts were in grade 5, 6, and 7 in 2002. Figure 3 shows that these cohorts advance in scale score means on a unidimensional
metric. The amount of year-to-year growth never exceeds about approximately 4/10th of a standard deviation for the grades studied.

![Means by Cohort across Grades and Years](image)

**Figure 3.** Unidimensional means for three cohorts across grades and years.

**Across-Year Study: Multidimensional Growth**

The unidimensional and multidimensional growth patterns for the 2002 Grade 5, 6, and 7 Cohorts are presented in Figures 4a, 4b, and 4c. Student growth from a multidimensional perspective, as depicted in Figures 4a to 4c, is not always increasing for certain types of skills or dimensions, as in the unidimensional case. The unidimensional growth is close to the average of the multidimensional growth patterns with some exceptions. It is important to note that the 2002 scaling was the base scale, with all years being scaled onto the year’s metric. Figure 1a shows the means across grades 5, 6, 7, 8, and 9 for 2002, with Grade 7 being a more able group, then the Grade 8 sample. There are several reasons why the grade 8 sample may not look as able as the grade 7 sample: 1) The Grade 8 sample is in fact not as able as the Grade 7 sample or 2) A linking or scaling error has occurred, or 3) Multidimensionality is present, or 4) Some combination. Given that each sample was chosen based on matching students across years, the
Figure 4a. Multidimensional and Unidimensional means for the Grade 5 Cohort across 2002, 2003, and 2004.

Figure 4b. Multidimensional and Unidimensional means for the Grade 6 Cohort across 2002, 2003, and 2004.
sample of 1500 examinees is not necessarily representative of the actual population. For convenience, this research only used the sample of 1500 matched students who were tracked across grades and after a thorough review, it was found that the Grade 7 group sampled was slightly more able than the Grade 7 population in 2002, while the Grade 8 sample was less able than the average of the Grade 8 population in 2002. In and of itself, this doesn’t pose a problem, except for the fact that we also scaled 2003 and 2004 to the 2002 scale and a bias may have been introduced. For example, in Figure 4b, the 2003 means are all higher than 2002, while the 2004 means are significantly lower than the 2002 and 2003 means. This may be caused by the scaling, as the 2004 (Figure 4b) Grade 7 dimension is scaled back to the 2002 cohort and since the 2004 Grade 7 cohort are being compared to a higher performing base cohort, a bias may have been introduced, pulling the mean down lower than it would have been with a more representative sample for both years. The grade 8 dimension in 2004 (Figure 4b) may have been biased in a similar direction, but for a different reason. The Grade 8 dimension in 2004 is scaled to a lower performing 2002 cohort, Figure 4c shows that in 2002, overall, the Grade 7 cohort did
better than the grade 6 cohort (Figure 4b), while the Grade 7 cohort in 2003, shows lower means then in 2002.

Vector Plots for Multidimensional MC and CR Items

Figure 4a shows a vector representation of items loading on Grade 5 (theta 1) and Grade 6 (theta 2) dimensions for both MC and CR items. Figure 4b displays the items that measure the Grade 6 (theta 1) and Grade 7 (theta 2) dimensions. Figure 4c shows the items that measure Grade 7 (theta 1) and Grade 8 (theta 2) dimensions. Each item starts at the origin and vectors in the first quadrant are more difficult items, while the easier items lie in the third quadrant. Vectors that align closer to the theta 1 axis, measure more of that grade level dimension, while vectors that align with the theta 2 axis, measure more of that dimension. If a vector lies in between the two axes, then it measures both dimensions equally.

Some Grade 5 Items

Figure 4a. Common item across Grade 5 and Grade 6.
Figure 4b. Common item across Grade 6 and Grade 7.

Figure 4c. Common item across Grade 7 and Grade 8.
The items vectors in Figure 4a (Grade 5/Grade 6 dimensions) show that most of the items measure more of the Grade 5 dimension, then the Grade 6. Most of the items fall in the third quadrant and thus, are not very difficult. For the Grade 6/Grade 7 dimensions in Figure 4b, we see that there is a mixture of items, some measuring more of Grade 6 and some more of the Grade 7 dimensions. Figure 4c shows that the items have medium difficulty, with some items measuring more of the grade 8 dimension and some the grade 9 dimension. Note that item 2 is a more difficult item measuring both dimensions equally and item 10 is somewhat easy and measures more of Grade 8, then the Grade 9 dimension.

Discussion and Conclusion

The No Child Left Behind Act requires students to be tested in key content areas across grades 3 through 8, thus giving rise to the vertical scaling efforts in many states. The often-heard criticism of vertical scales is that of fitting a unidimensional IRT model to content areas that cannot possibly be unidimensional when spanning several grades. The promise of multidimensional vertical scaling is that they provide more flexible models of student growth. Several successful simulations and real data applications have been completed in the area of multidimensional vertical scales, however, more real-data applications are needed in order to ascertain its accuracy and usefulness (Patz, et al., 2003; Karkee, Lewis, Hoskens, Yao, & Haug, 2003; Yao & Mao, 2004; Yao, Patz, & Lewis, 2003). In this study: 1) Real vertically-scaled data were used for comparison purposes across a unidimensional and multidimensional model of student growth, and 2) Individual students’ growth across years was tracked, allowing for an alternative view of student growth trajectories.

Specifically, this research explored how a multidimensional approach to student growth data can be implemented and how different grade levels measure different constructs. It was found that growth in mathematics does not occur uniformly across all of the skills and knowledge areas. Growth across grades is extremely complex; however, MIRT models can help to sort through such complexities, but test items would need to be written to cover the various dimensions in order to be able to model actual student growth.
This research has not yet been validated from a substantive content/construct view by test developers and the increase in model fit/complexity tradeoff for the MIRT model does not necessarily warrant a multidimensional parameterization. However, the ability to compare a unidimensional and multidimensional parameterization of items and growth is an extremely valuable tool for ensuring appropriate model selection.
References


