

The Vertical Scaling of Science Achievement Tests

Mark D. Reckase
Joseph Martineau
Michigan State University

October 2004

Paper commissioned by the Committee on Test Design for K-12 Science
Achievement
Center for Education
National Research Council

Copyright © 2004 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the draft papers are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Committee on Test Design for K-12 Science Achievement or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

The No Child Left Behind Act of 2001 requires that states assess the performance of school children in science beginning in the 2007-2008 academic year. The goal of that assessment is to determine how much students know and can do in science from grades 3 through 8. When tests are given every year, a natural question is how much the students gain in knowledge from year to year. It is generally fairly easy to assess gain, such as noting the increase in height from year to year, when the thing being measured can be characterized as varying along a single dimension. However, the skills and knowledge included in the science curriculum are fairly complex, including many different subject matter areas, such as life science, earth science, and physical science, and a variety of skills, such as the design of experiments, laboratory procedures, and the critical evaluation of the results of research. As a result of the complexity of the science curriculum, school related science assessments such as the National Assessment of Educational Progress Science Assessment, have very complex frameworks that includes a variety of content and cognitive levels (National Assessment Governing Board, 1996).

The easiest way to assess gain from year to year is to give the same test each year and compute differences in scores. This approach could potentially work for two adjacent grade levels, but when the span of grades from 3 to 8 is considered, the shift in content and challenge of the materials is too great to be appropriate for all levels. It is unlikely that 3rd grade students could reasonably attempt material designed for 8th grade students, and 8th grade students would likely find 3rd grade test items trivially easy. Instead of a common test approach, tests are typically designed for each grade level with some common items in the tests for adjacent levels. The common items allow the scales of the tests to be linked together so that gain over grades can be estimated. The process of linking the scales of the tests together when they are designed to be of increasing difficulty is called vertical scaling.

Vertical scaling has been considered as a very challenging psychometric procedure for many years (see Feuer, Holland, Green, Bertenthal & Hemphill, 1999 for a summary of issues). The reason is that procedures for linking score scales make assumptions that the tests are measuring the same constructs and that they are reasonably parallel in their construction. Neither of these assumptions is met when the tests are designed for different grade levels. For some content such as reading, an argument can be made that adjacent grade levels have test constructs that are fairly similar. If that is the case, the linking of the test score scale may work sufficiently well to give some meaningful results. However, for science assessments, the content is likely to shift in many different ways from grade to grade. At one grade level the emphasis might be on life science. At the next grade level the emphasis might be on earth science. It is not likely that the scales of science tests on such different content can be

linked together using procedures that assume common constructs, even if there are common items in the test forms for the adjacent grade level.

One approach that might make such linkages possible is to acknowledge the multiple dimensions assessed by the tests and use a multidimensional model for the linkage. The purpose of this report is to investigate the possible use of such a multidimensional model for linking science assessments. Actual test data from a series of grade level science tests will be analyzed to determine whether they can be successfully linked to allow meaningful reporting of gains over grades.

This report has three sections. The first describes the multidimensional item response theory model and its use for analyzing science achievement data. The second section summarizes the analysis of the actual test data. The final section discusses the overall results and the usefulness of the approach for vertically scaling grade level science tests.

A Multidimensional Model for Responses to Achievement Test Items

The model that is used for the research reported here is the multidimensional extension of the three-parameter logistic model. This model is given in Equation 1

$$P(u_{ij} = 1 | \vec{q}_j, \vec{a}_i, d_i, c_i) = c_i + (1 - c_i) \frac{e^{\vec{a}_i \vec{q}_j + d_i}}{1 + e^{\vec{a}_i \vec{q}_j + d_i}} \quad (1)$$

where u_{ij} is the item score, 0 or 1, for person j on item i ,

\vec{q}_j is a vector of parameters that describe the location of person j in an n -dimensional space,

\vec{a}_i is a vector of parameters that specifies the discrimination power of the item i on each of the n dimensions in the space,

d_i is a parameter related to the difficulty of item i ,

c_i is a parameter that specifies the probability of correct response for persons who are low on all of the dimensions,

and e is the mathematical constant 2.7182818 . . .

This model is based on the assumption that persons who take a test vary on a large number of cognitive dimensions. The model gives the hypothesized relationship between a person's status on the cognitive dimensions and the probability of a correct response to a test item. This model is sometimes labeled as a compensatory model because a high level on one dimension can compensate for a low level on another dimension. Although more complex models exist that do not have this compensatory property, such as the partially

compensatory model proposed by Sympson (1978), the model given in Equation 1 has been shown to be useful in a large number of cases (see Reckase, 1997 for one example) and no studies have shown that the more complex model are better at describing the processes behind item responses. Therefore, the model in Equation 1 will be used here to model the performance of students on grade level science tests.

To give additional insight into the characteristics of this model, three different ways of representing the model will be shown for the simple case of two cognitive abilities, θ_1 and θ_2 . Figure 1 shows the graph of the relationship between the probability of correct response to an item that has an \vec{a}_i -vector of $[1.5 \ .75]$, $d_i = 1.2$, and $c_i = .2$.

Note that the probability increases more quickly as θ_1 increases than it does as θ_2 increases. This is because of the different sizes of the elements in the \vec{a}_i -vector. Also, the lowest probability is .2 when the level on both of the dimensions is very low. A straight line has been drawn on the surface to show the places on the surface that have the steepest slope in a direction perpendicular to the surface. This line is given by $a_1\mathbf{q}_1 + a_2\mathbf{q}_2 + d = k$ where k is a constant value related to the level of the probability. For the line along the points of steepest slope, $k = 0$ and the probability is half way between the c -parameters, .2, and 1.0, which is .6.

Another way of representing the relationship between the location of a person in the ability space and the probability of a correct response to the item is through a plot that shows the contours of the surface that have equal probabilities. The contour plot for the item shown in Figure 1 is given in Figure 2.

Figure 1
Item Response Surface

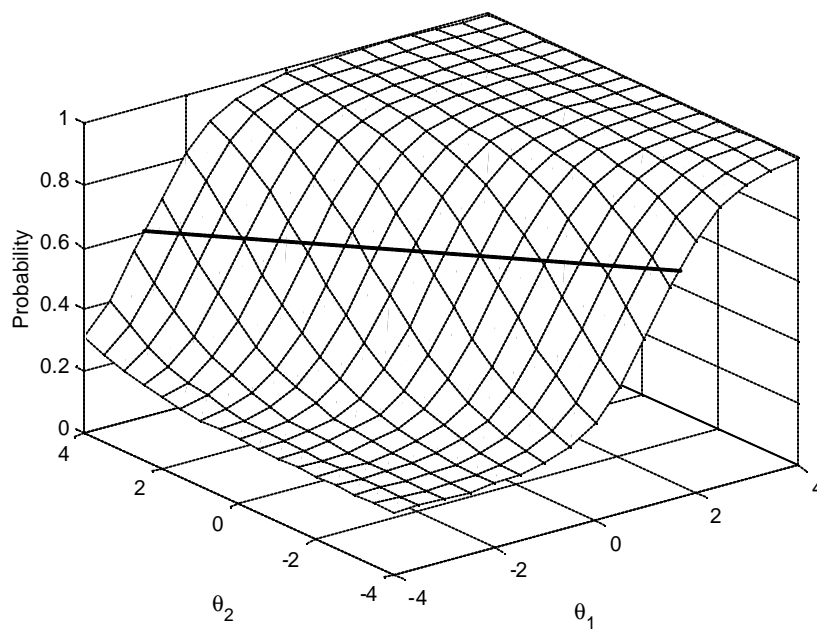
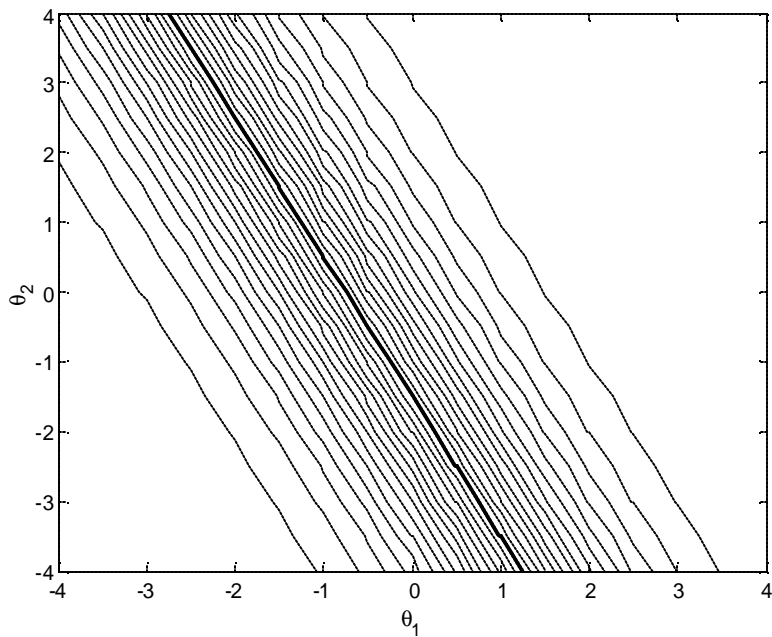
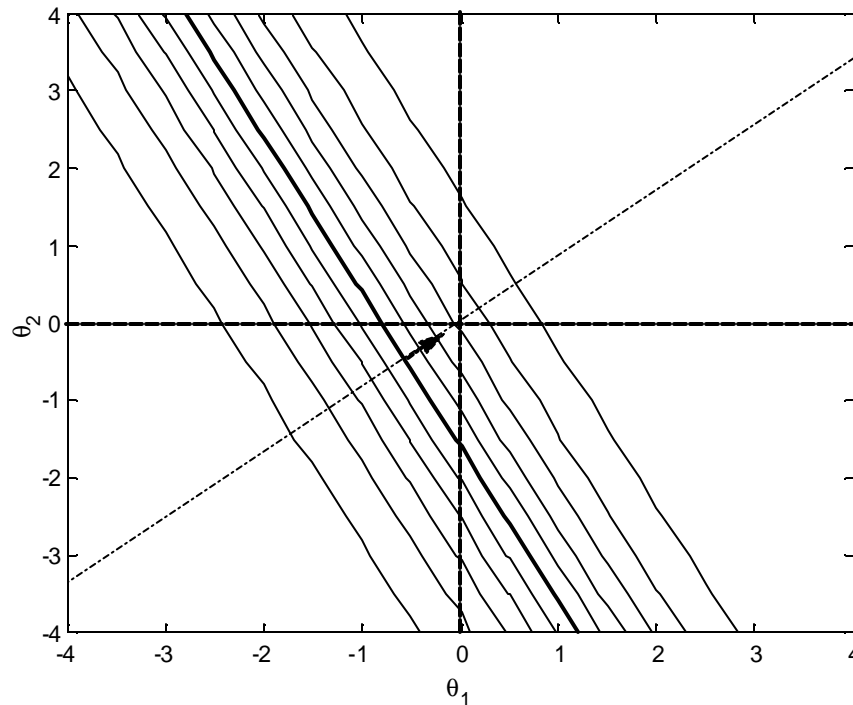


Figure 2
Contour Plot for the Item



The surface plot and contour plot are useful for showing the relationship between a point in the ability space and the probability of correct response for one item, but when it is desirable to show more than one item, these types of graphs quickly become confusing. Therefore, a method has been developed to represent a single item by a vector. The general characteristics of this vector are shown in Figure 3.

Figure 3
Item Vector on a Contour Plot



This Figure shows a line through the origin of the space that is perpendicular to the equiprobable contours for the item. The line of steepest slope is indicated on the plot as the contour that is darker than the others. The vector has its base on the contour that corresponds to the line of steepest slope and its direction is at right angles to that contour on the line that goes through the origin of the space.

The location of the base of the vector and the direction are a function of the item parameters for the item (see Reckase & McKinley, 1991 for the derivations). The distance of the base of the vector from the origin of the space is given by the multidimensional difficulty of the item, $MDIF_i$

$$MDIF_i = \frac{-d_i}{\sqrt{a_i a_i}} \quad (2)$$

The angle of the vector with each axis is given by

$$\mathbf{a}_{ij} = \arccos \frac{a_{ij}}{\sqrt{\bar{a}'_i \bar{a}_i}} \quad (3)$$

where a_{ij} is the j -th element of the vector of discrimination parameters for Item i . To add some additional information to the vector, the length is usually made a function of the value of $\sqrt{\bar{a}'_i \bar{a}_i}$. This value is called the multidimensional discrimination of the item. Using item vectors, the difficulty of the items, the direction of maximum discrimination, and the discriminating power of the items can be shown for a number of items in the same graph.

Figure 4
Item Vectors for a 10-Item Test

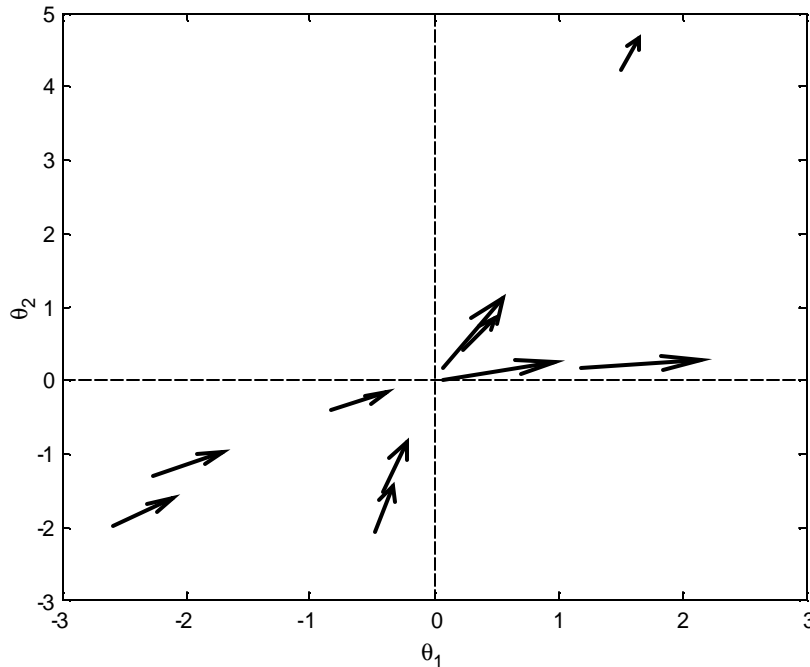


Figure 4 shows an example of an item vector representation of a 10-item test. The example shows items that vary substantially in difficulty. Those in the lower left quadrant are fairly easy items and those in the upper right quadrant are more difficult. The vector to the far upper right represents a very difficult item. The items in this example tend to point in two different directions. One set tends to fall more along the θ_1 -axis while the second set tends to fall more along the θ_2 -axis. These two sets of items are measuring somewhat different constructs, although the constructs are related because the directions the vectors are pointing are not orthogonal to each other. The lengths of the vectors are related to the discriminating power of the items. Those that are longer

represent more discriminating items. The hard item at the upper right is much less discriminating than the item immediately above the θ_1 -axis. As indicated in Figure 3, the base of each vector is on the line of steepest slope for the item response surface. This vector representation of the items can show a large amount of information about the items in a test in fairly concise form.

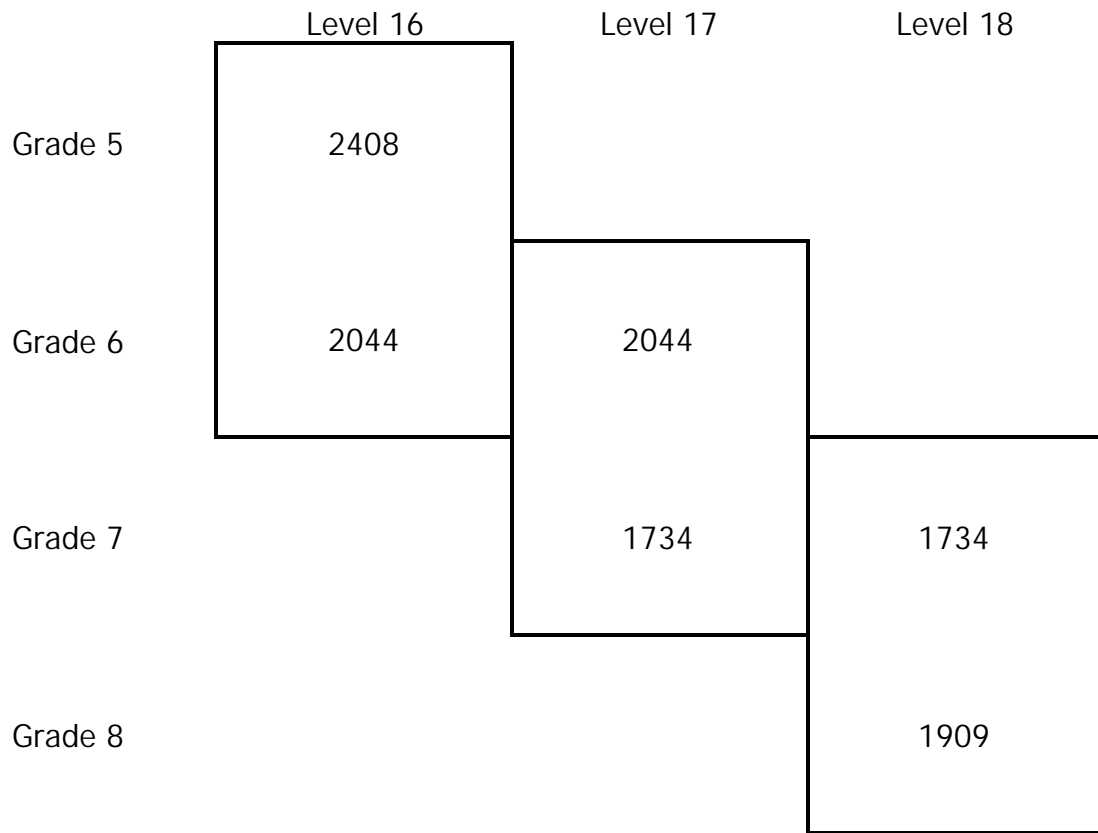
The vector representation of items can be used to identify sets of items that are measuring similar composites of skills. Those items that have item vectors that are pointing in the same direction measure the same combination of knowledge and skills. The angle between the directions of the items vectors is a measure of the difference in the skills measured by items. When the angle is 0 degrees, the items are measuring the same combinations of things. When the angle is 90 degrees, the items are measuring totally different things. Thus, the angle between the directions of item vectors, or some transformation like the cosine of the angle, can be used in a cluster analysis to identify sets of items that are measuring essentially the same composites of abilities.

In the context of the assessment of knowledge and skills in science, it is expected that the items in science tests will be fairly complex, requiring multiple skills and knowledge areas. Tests for different grade levels will probably also require different combinations of skills and knowledge for students to respond correctly to the items. By using multidimensional item response theory models to analyze the test data, it should be possible to identify some of the skill and knowledge components needed in each grade level test and determine how they change from grade to grade. Further, the multidimensional space of skills and knowledge at one grade level can be linked to that at another grade level to determine how the skills and knowledge change with instruction.

Analysis of a Set of Grade Level Science Tests

To determine how this multidimensional IRT model can be applied to the development of a vertical scale for grade level science assessments, test data were obtained from the vertical scaling study of a grade level science test. The data from the grade level tests were in the form shown in Figure 5. Students at each grade level took test forms at two levels. For example, 6th grade students took both levels 16 and 17, and 7th grade student took both levels 17 and 18. For the calibration of the items using the multidimensional IRT model, the level 17 items were considered as a common item set for the tests made up of level 16 and 17, and levels 17 and 18. Each of the levels has 25 items, so the calibration model had 50 item tests with 25 common items.

Figure 5
Data Structure for the Grade Level Science Tests



Each of the tests was calibrated using the TESTFACT program. The parameters for pairs of tests were rotated and translated to a common configuration using an oblique procrustes procedure that is a generalization of the methodology developed by Min (2002). After the transformation of the item parameters to a common coordinate framework in the multidimensional space, item vectors can be graphed to allow comparisons of the skills and knowledge measured by the various levels of the tests. The meaning of the score scale can also be described by looking at the dimensions of change represented by differences in scores at different points on the score scales.

A three dimensional analysis of the test levels will be shown first because that number of dimensions can be easily represented graphically. Later analyses use a higher dimensional solution that captures all of the dimensions of knowledge and skills and the different grade levels of the tests.

Item Vector Plots

To show how the tests change in their sensitivity to differences in performance on a various dimensions, the item vector plots are presented in three dimensions. Figure 6 shows the vector plot for the Level 16 test. From this plot it can easily be seen that the three easiest items on the test (those at the bottom front of the plot) are best measuring the first dimension, while the rest of the items tend to be measuring a composite of all three dimensions.

Figure 6
Item Vector Plot for Level 16

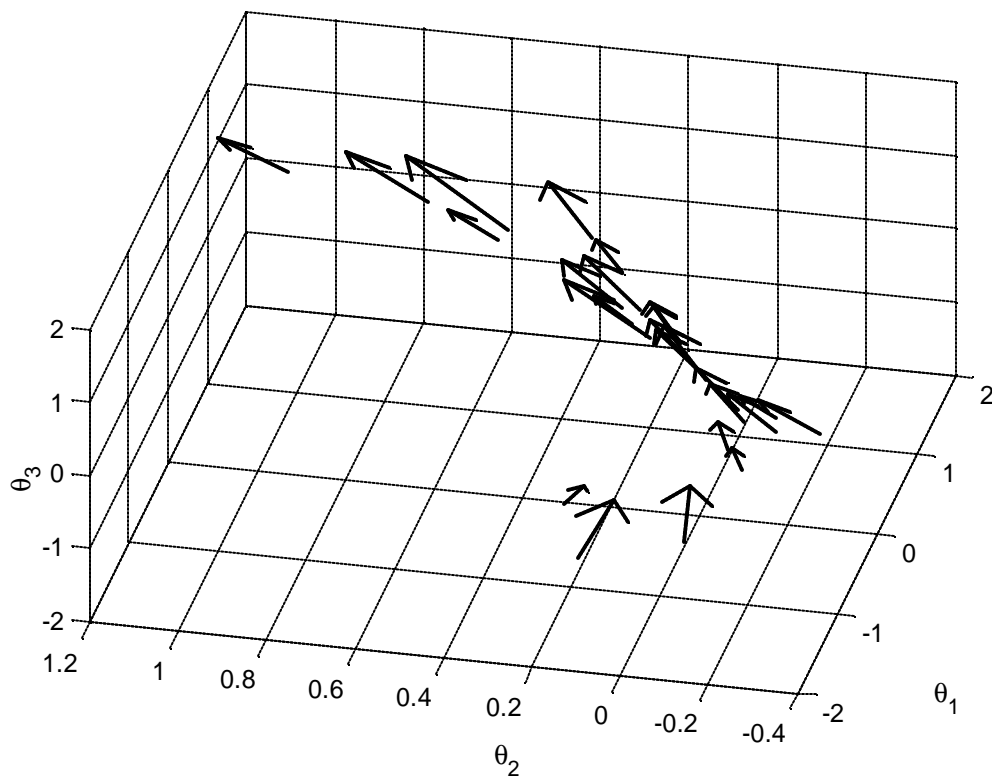
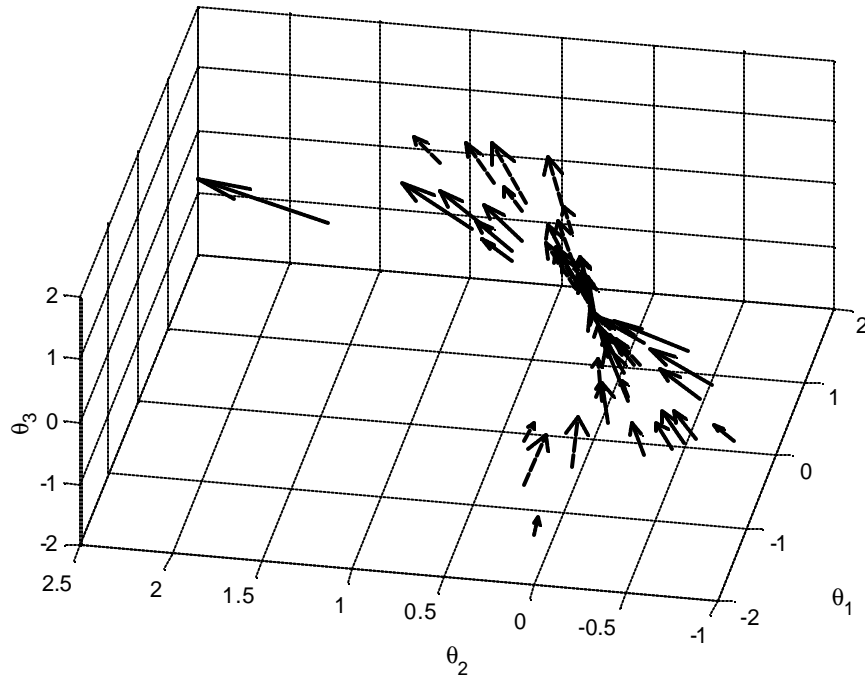


Figure 7 shows the items from Level 16 with dashed lines and those from Level 17 with solid lines. From a comparison of the dashed vectors to the solid vectors, it is clear that the Level 17 has items that are more aligned with the second dimension than those from the test at Level 16. This implies that that construct is shifting somewhat with the increase in level of the test.

Figure 8 includes the item vectors from the Level 18 test. Now these items are the solid vectors and those from the previous two levels have dashed vectors. The Level 18 vectors are also printed darker. Two characteristics of the Level 18 vectors compared to the other levels can be observed. First, the Level

18 vectors seem to have two distinct clusters. This suggests that here are two major content domains in the test. The second observation is that many of the vectors are very short indicating that they are less discriminating than those from the other levels. This could be caused by items measuring a fourth dimension that is being projected down into this three dimensional representation.

Figure 7
Item Vector Plot for Levels 16 and 17

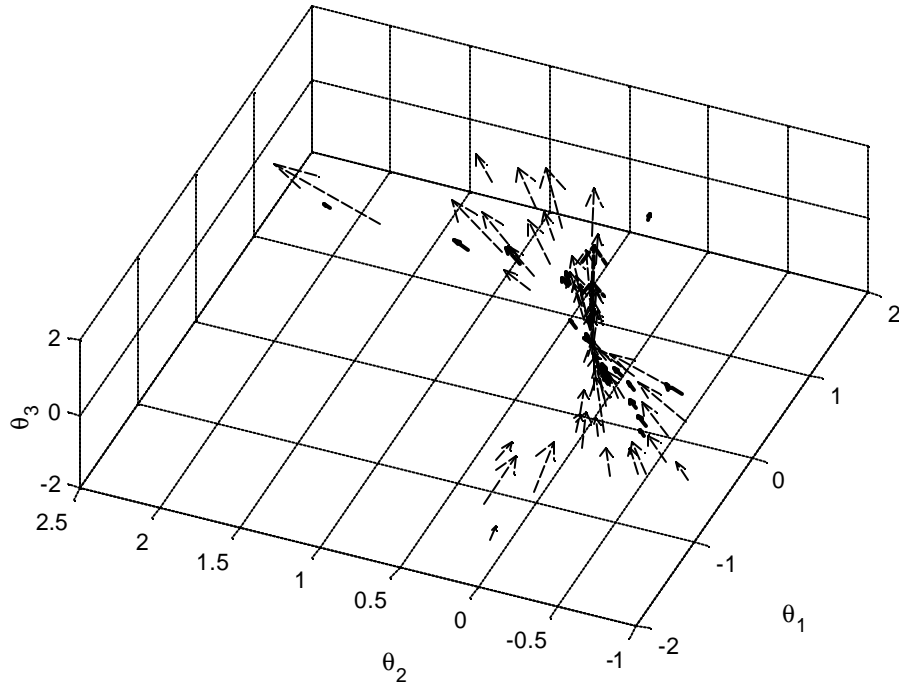


Another way of determining the shift in the construct with change in level of the test is to plot the change in skills for groups taking each test. Students who took each level of the test were sorted into ten groups based on the unidimensional IRT scaling of the test. Then, the mean on each dimension was determined for each of the ten groups. The vectors of means for each group was then plotted to show the increase in skills and knowledge represented by the single unidimensional score.

The results of this analysis are shown in Figure 9. Three lines are shown in that figure -- one for each of the levels. The lowest level used 6th grade student data. It is shown using a dotted line. The second level used the 7th grade student data. That line has dots and dashes. The third level used the 8th grade student data and it is represented by a solid line. The lines in the three dimensional space are projected onto each of the planes of the space so their orientation can more easily be seen. One obvious feature of these lines is that

they are not straight lines. They wind through the space showing that students do not improve on all dimensions in a consistent way.

Figure 8
Item Vector Plots for Levels 16, 17, and 18



Traditional unidimensional equating methods must connect these convoluted scales using linear approximations to them. If the general trends of the lines are linear and in the same direction, the traditional equating methods may work fairly well. The result is a vertical score scale that is based on a constant composite of skills and knowledge from many dimensions. However, if the composite of skills is related to the level of the test, then using a single unidimensional model will not adequately fit the data.

Once the scales are linked together using multidimensional procedures, the change in performance over grades six through eight can be determined. Figure 10 shows the change in performance in the three dimensional performance space on the linked scale over the three levels of the test. The projections on the coordinate planes also include ovals that are proportional to the standard error of measurement for the score points in the graph. From this graph, it is clear that the growth in skills and knowledge is not uniform along the different dimensions. At the lowest point on the scale, the improvement is mainly along Dimension 1 and then there is a big jump along Dimension 3. Then growth progresses along a combination of Dimensions 1 and 2 and growth on

Dimension 3 is relatively flat. When the actual test items for the levels are available, information on the skills related to the dimensions will be provided.

Figure 9
Plots of Student Performance Levels on the Three Test Level

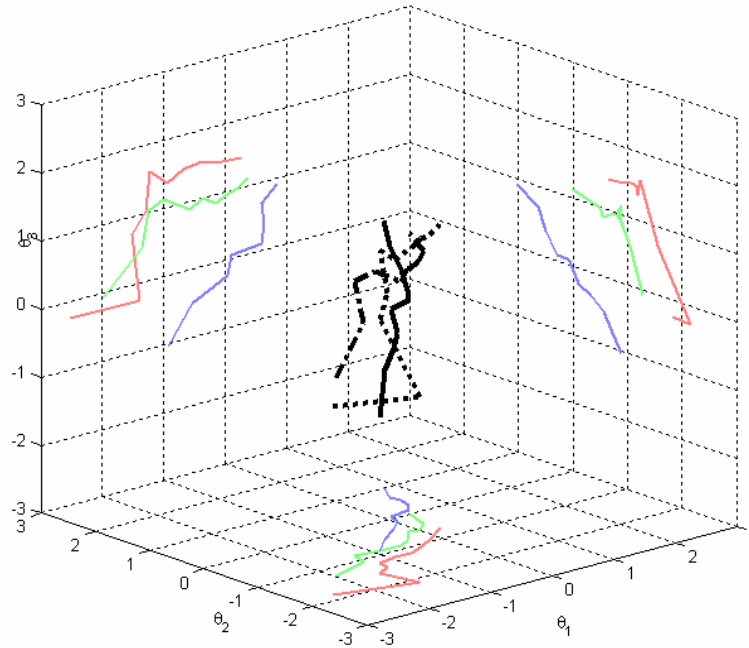
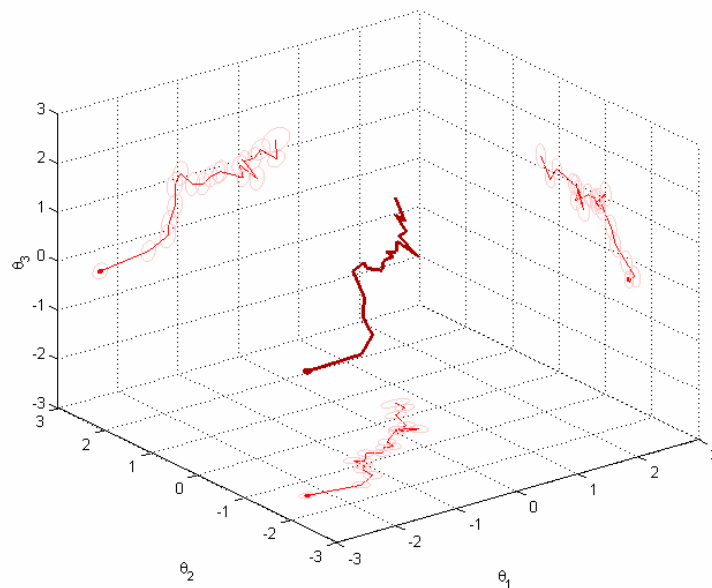


Figure 10
Representation of the Linked Unidimensional Scale
Over Levels 16, 17, and 18 for Grades 6 through 8.



Fifteen Dimension Analysis

The results presented above are for only three test levels and they are shown in three dimensions so that convenient graphic representations could be used. Working with more test levels so changes from Grade 3 to Grade 8 can be represented is more challenging. More dimensions are needed because there is more variation in content. As a result, these simple graphic representations can no longer be used. For the purpose of this analysis, 15 dimensions were assumed. This is not meant to imply that there are really 15 distinct dimensions in the data. That number was used because earlier work (Reckase & Hirsch, 1991) indicated that over estimating the number of dimensions did not cause too many difficulties, but underestimating the number of dimensions caused dimensions to project on top of each other resulting in misinterpretation of test structures.

The data for this study came from the norming/scaling sample for national grade level science test battery. The data from grades 3 to 7 were used. The test forms were developed for use with grades 3 to 8. The data is structured such that students in each grade took the appropriate grade-level test plus the next level up, as shown in Table 1. Thus, the equating design is a common-items non-equivalent groups design, with the fourth through seventh grade items being taken by students from two grades. The third grade test had 20 items, and the fourth through eighth grade tests had 25 items. The sample sizes for the groups of students are displayed in Table 1.

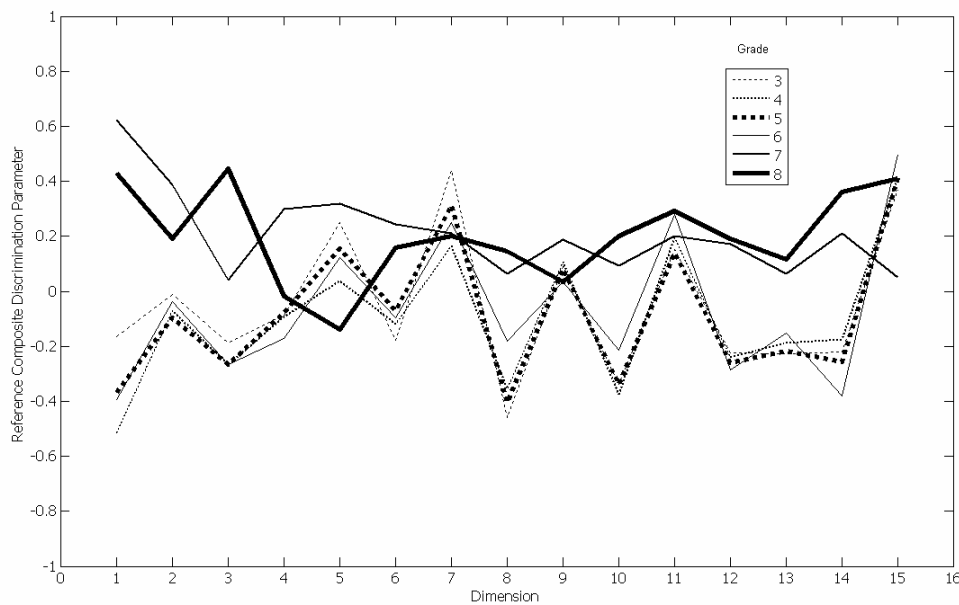
Table 6
Structure of the Science Achievement Data

Test level	# of items	Number of students in grade...				
		3	4	5	6	7
Grade 3	20	1855				
Grade 4	25		2074			
Grade 5	25			2167		
Grade 6	25				2044	
Grade 7	25					1734
Grade 8	25					

The data from each grade level were calibrated using NOHARM (Fraser, 1988) with 15 dimensions specified for the solution. The results from each grade level were then linked together using an extension of the procedure developed by Min (2003). The technical details of the process are given in the appendix to this report.

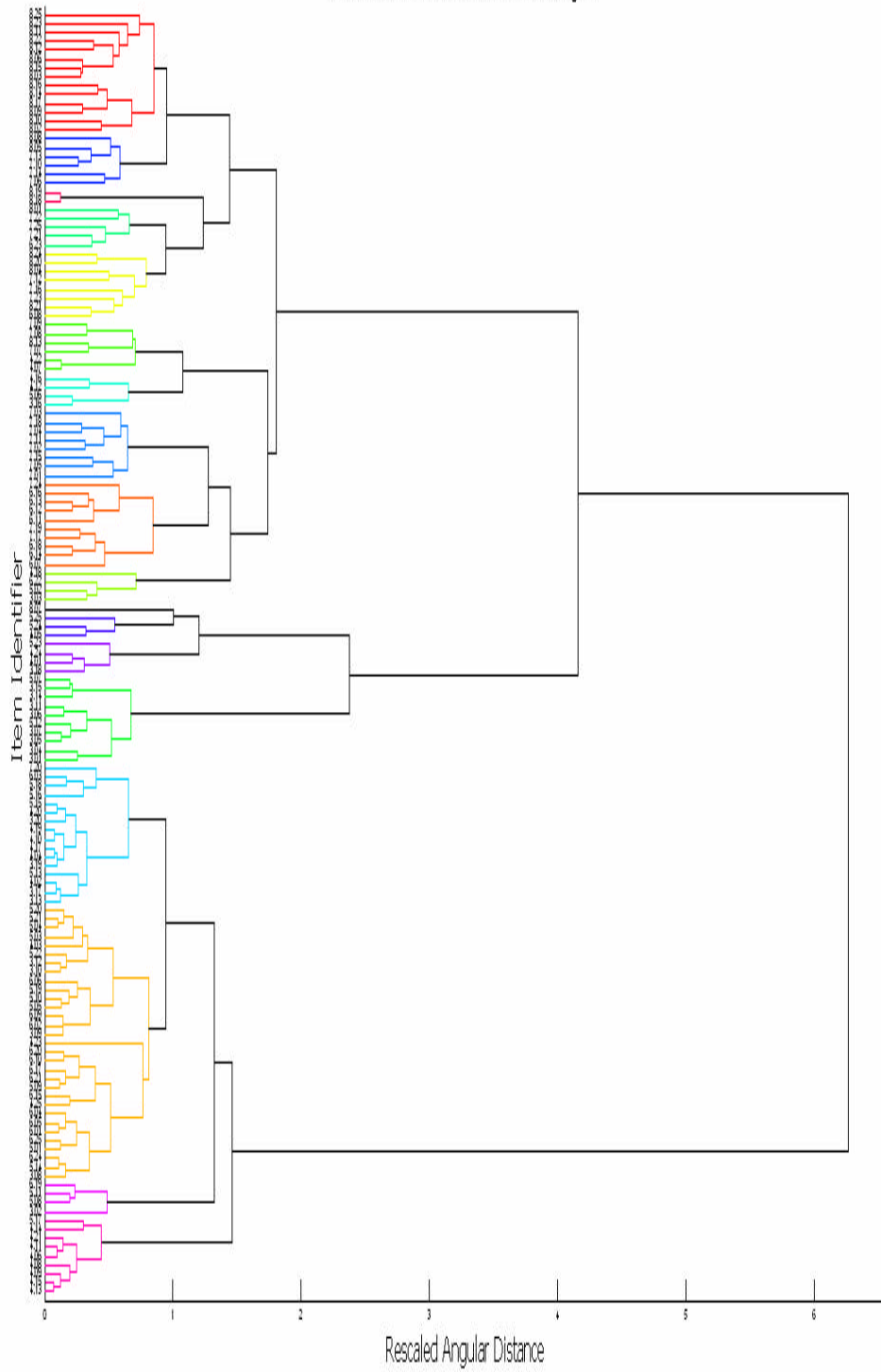
To determine if there were shifts in dimensions in 15 dimensions that were of the type shown in the three-dimensional plot in Figure 10, the major dimension measured by each test level was determined. This dimension is called the reference composite. It is essentially the average direction in the 15-dimensional space for the set of items in that test level. If the reference composites have the same angles with all of the coordinate axes, the test levels are measuring the same composite of skills. If the angles with the coordinate axes change over levels, that means that there is a shift in the composite of skills measured by the tests. The reference composite is determined from the matrix of a -parameters for the items in the test. Specifically, it is the first eigenvector of the $\vec{a}'\vec{a}$ matrix. Figure 10 shows a plot of the first eigenvectors for each of the grade level tests from the battery. The figure shows that the tests for grades three through six were almost indistinguishable from each other. However, the tests for grades seven and eight were quite different.

Figure 10
Reference Composites for the Grade Level Tests



To determine how these tests differed in a substantive way, the angles between the item vectors for all of the items in all of the test levels were determined and were used in a cluster analysis. Item vectors that had 0 degrees between them were measuring the same combination of skills. Those with large angles between them were measuring quite different skills.

Figure 11
Clusters from Forms for Grades 3 through 8



The cluster analysis resulted in three main clusters. The large bottom cluster contains mainly items from grades three, four, five, and six. Most of these items measured knowledge of scientific facts that would very likely be learned outside of the classroom and items that required general logical reasoning. For example, an item might ask the part of a plant that attracts insects and birds or the student would have to pick the picture of a thing that does not measure time – a thermometer. The middle cluster also contains mainly items from grades three through six, but they are of a different character than those from the first cluster. They ask mainly about knowledge and skills that would mainly be learned in the classroom. For example, an item might require that students know that the sun is closer than other stars and therefore it appears bigger in the sky.

The top cluster is dominated by items for grades seven and eight. These results are consistent with those from the reference composites. The subclusters in the large cluster are of similar type to those for the lower grades. In addition, there are clusters of items that require comprehending information presented in figures and graphs, and others that require knowledge of experimental procedures.

What is notable about the cluster analysis results is that there are no clusters that are related to the substantive content areas of science such as biology, chemistry, earth science and physics. From studying the items on the tests, it appears that these clusters do not exist because the items tend to be on general science topics rather than on topic specific topics. This is likely because the tests were designed to be usable in schools across the country that have different emphases and different schedules for the presentation of science content.

Discussion

The purpose of this report is to demonstrate how a multidimensional approach to the scaling of sets of tests that cover different grade levels of content can show the complexity of the content of science assessments and the complexity of patterns of growth that result from instruction. A series of levels of a grade level science assessment were analyzed to show how they can be linked together using multidimensional item response theory. Those procedures were first used assuming that three dimensions would be sufficient to model the data so that the results could be presented graphically. Even with the simplification of using three dimensions, the complexity of growth in science can be documented. Test items measure complex combinations of skills and knowledge and the combinations change with the level of the test. Growth in performance does not occur in a uniform way. There are jumps and pauses in the patterns of growth.

The 15-dimensional analysis shows that the major content emphasis of the tests shifts quite dramatically between grades 6 and 7. It also shows that the reference composites for the tests are fairly consistent between grades 3 to 6. This last result is probably due to a dimensional purification process that likely occurs when items are selected to support a unidimensional scaling process. The possibility that the tests were designed to support a vertical scaling based on a unidimensional model may explain why there are no content clusters in the solution. The items that were most content specific may not be used on the tests because they would not support the unidimensional model. Tests designed to give broad domain coverage to science content would likely give more extreme shifts in dimensions and in the tracking of growth.

Overall, the results show that the tracking of growth over grades is very complex. Students do not gain knowledge in multiple content areas in a uniform way. Rather, the growth is on different dimensions at different times. The tests also reflect different skills and knowledge at different grade levels. These results suggest that multidimensional models are needed to reflect the complexities of vertical scaling of science achievement.

The results that are presented are limited by the relatively small number of items analyzed at each level and the relative simplicity of the content of the tests. There were only 25 items in the links between levels and specific content knowledge questions were not on the test. This limits the amount of complexity that can be identified. Even with that limitation, the complex nature of growth in science knowledge over grade levels is evident.

References

- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W. and Hemphill, F. C. (Eds.) (1999). *Uncommon Measures: Equivalence and Linkage among Educational Tests*. Washington, DC: National Academy Press.
- Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. New South Wales, Australia: University of New England.
- Min, K.-S. (2003). *The Impact of Scale Dilation on the Quality of the Linking of Multidimensional Item Response Theory Calibrations*. Unpublished Dissertation, Michigan State University, East Lansing, MI.
- National Assessment Governing Board (1996). *Science Framework for 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: Author.

- Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and R. K. Hambleton (Eds.) *Handbook of modern item response theory*. New York: Springer.
- Reckase, M. D. & Hirsch, T. M. (1991, April). Interpretation of number correct scores when the true number of dimensions assessed by a test is greater than two. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Reckase, M. D. and McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 14(4), 361-373.
- Sympson, J.B. (1978). A model for testing multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). University of Minnesota, Department of Psychology, Minneapolis.
- Wang, M.-m. (1985). *Fitting a Unidimensional Model to Multidimensional Item Response Data: The Effects of Latent Space Misspecifications on the Application of IRT*. Unpublished manuscript.

Appendix

Mathematical Methods Employed in MIRT Equating

Various methods of equating exist for the MIRT model. Li & Lissitz (2000) identified three indeterminacies that must be resolved for MIRT equating to be accurately performed. These three indeterminacies are:

1. *Rotational indeterminacy*. Because the non-linear factor analysis employed in MIRT identifies the axes of the multiple dimensions in an arbitrary manner, the axes of comparison forms of an assessment must be rotated in the multidimensional space to match the axes of the base form.
2. *Unit indeterminacy*. Because the units of the scales on which various dimensions—or traits—are reported are arbitrary, the units of the various dimensions identified on comparison forms must be dilated or compressed to match the units of the base form.
3. *Origin indeterminacy*. Because the origins of the scales on which various dimensions—or traits—are reported are arbitrary, the origins of the various dimensions identified on comparison forms must be translated to match the origins of the base form.

Li and Lissitz (2000) resolve these three indeterminacies by using, respectively

1. An orthogonal Procrustes rotation matrix \mathbf{T} ,
2. A scalar dilation parameter k , and
3. A translation vector \mathbf{m} .

These quantities are used to transform the original item discrimination parameters (\mathbf{a}_{ci}), item difficulty parameters (\mathbf{d}_{ci}), and person trait parameters (θ_{ci}) estimated from the comparison forms to the corresponding metric on the base form (\mathbf{a}_{bi} , \mathbf{d}_{bi} , and θ_{bi}). These transformations are performed as follows:

$$\begin{aligned}\mathbf{a}_{\tilde{bi}} &= k\mathbf{a}'_{ci}\mathbf{T} \\ d_{\tilde{bi}} &= d_{ci} + \mathbf{a}'_{ci}\mathbf{T}\mathbf{m} \\ \theta_{\tilde{bi}} &= (\mathbf{T}^{-1}\theta_{ci} - \mathbf{m})/k\end{aligned}$$

where $\mathbf{a}_{\tilde{bi}}$, $\mathbf{d}_{\tilde{bi}}$, and $\theta_{\tilde{bi}}$ are the values of the items parameters from the comparison form transformed to match the metric of the base form.

Min (2003) identified a problem with the Li and Lissitz (2000) approach to MIRT equating in that the scalar dilation parameter is insufficient for compressing/dilating the scales of the multiple dimensions. A scalar dilation

constant dilates/compresses the scale of each dimension by exactly the same amount, but separate MIRT calibrations on multiple forms may dilate/compress the scales of the multiple dimensions to differing degrees. To address this flaw, Min (2003) developed an MIRT equating procedure that replaced the Li & Lissitz scalar dilation parameter with a diagonal dilation matrix that allows for differential dilation/compression of the scales of the various dimensions. The transformations then become:

$$\begin{aligned} \mathbf{a}_{\tilde{b}i} &= \mathbf{a}'_{ci} \mathbf{TK} \\ d_{\tilde{b}i} &= d_{ci} + \mathbf{a}'_{ci} \mathbf{Tm} \\ \tilde{b}_{ci} &= \mathbf{K}^{-1} (\mathbf{T}^{-1} \tilde{b}_{ci} - \mathbf{m}) \end{aligned}$$

This study identified an additional important weakness in the Min (2003) approach to MIRT equating. When the dimensionality modeled is low, both procedures perform well, but when the dimensionality modeled is high, the computational burden of MIRT equating becomes infeasible. This is because of an interaction between dimensional dominance across forms and a particular characteristic of orthogonal Procrustes rotation:

1. The relative dominance of the various dimensions may change across forms of an assessment. Therefore, the ordering of the dimensions as they are estimated in MIRT software may change from one form of the assessment to another.
2. Orthogonal Procrustes rotation assumes that the ordering of the dimensions is the same for each form of the assessment.¹

This incompatibility results in the need to perform a Procrustes rotation of every permutation of the dimensions on each comparison form to a single specified permutation of the dimensions on the base form. This allows for determining which permutation of the comparison form has the best fit to the specified permutation of the base form. Because the number of permutations of n dimensions is $n!$, this is a relatively easy problem in low-dimensional spaces, but becomes prohibitive in high-dimensional spaces.

One solution to this problem is to employ a non-orthogonal Procrustes transformation (see Mulaik, 1972). Non-orthogonal Procrustes transformation automatically aligns each dimension of the comparison matrix with the dimensions of the base matrix depending upon the best fit of the dimensions of

¹ This is not explicitly stated anywhere, but the MDEQUATE program (Li, 1996) uses the technique of permuting the dimensions to obtain the best-fit solution, and my empirical experience verifies that markedly different fit measures result from different permutations of the comparison matrices.

the comparison matrix with the dimensions of the base matrix². At the same time that the non-orthogonal transformation reduces the computational load of equating every permutation of a comparison matrix, it does introduce another complication: by using an oblique transformation, the decision of which dimension in the comparison matrix is to be matched with which dimension in the base matrix is left entirely up to the mathematical procedure. When too many dimensions are modeled, random sampling error may result in the wrong dimensions being matched to each other by the oblique Procrustes transformation.

An additional benefit of the oblique Procrustes transformation is the elimination of the need for a dilation parameter/vector, since the oblique Procrustes rotation includes scale dilations for each dimension. From Mulaik (1972), the oblique Procrustes procedure results in the rotation matrix

$$\mathbf{T} = (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}'\mathbf{B}$$

where \mathbf{T} is the rotation matrix, \mathbf{A} is the matrix of the comparison form, and \mathbf{B} is the matrix of the base form. This resolves the rotational and unit indeterminacy of MIRT equating, resulting in transformations of:

$$\mathbf{a}_{\tilde{b}i} = \mathbf{a}'_{ci} \mathbf{T}$$

$$d_{\tilde{b}i} = d_{ci} + \mathbf{a}'_{ci} \mathbf{T} \mathbf{m}$$

$$\tilde{\mathbf{b}}_i = \mathbf{T}^{-1} \mathbf{b}_i - \mathbf{m}$$

However, this leaves unresolved the origin indeterminacy. In this portion of the MIRT equating procedure, this study also uncovered a weakness of previous methods of MIRT equating. Li and Lissitz (2000) and Min (Min, 2003) solve for the translation vector via a least-squares solution, minimizing the value Q , where

$$Q = \sum_{i=1}^{n_j} (d_{bi} - d_{\tilde{b}i})^2,$$

or Q is the sum of squared differences between the shared-item difficulty parameters on the base form of the assessment and the transformed difficulty parameters on the comparison form of the assessment.

² This is not explicitly stated anywhere, so it was verified empirically by equating all 24 column permutations of a comparison discrimination matrix with four columns (dimensions) to a base discrimination matrix. The result was equivalent fit and rotated discrimination matrices for all permutations within rounding error (differences from one permutation to another were near the computational accuracy of my computer, or 2.22×10^{-16} , for every value in every fit and discrimination matrix).

Li and Lissitz' (2000) solution to the translation vector \mathbf{m} is obtained by differentiating this expression with respect to each element of \mathbf{m} , setting the resulting expressions equal to zero, and solving the multiple equations simultaneously for the various elements of \mathbf{m} . Again, this approach to solving for \mathbf{m} functions well when working in a low-dimensional space. However, when working in a high-dimensional space, \mathbf{m} is considerably affected by rounding error, producing nonsensical results in some transformed item difficulty parameters.

Instead of taking the derivative of this expression with respect to each element in \mathbf{m} , this problem is solved by taking the derivative of this expression with respect to the entire vector \mathbf{m} , as follows:

$$Q = \sum_{i=1}^{n_i} (d_{bi} - d_{\tilde{b}i})^2 = (\mathbf{d}_b - \mathbf{d}_{\tilde{b}})' (\mathbf{d}_b - \mathbf{d}_{\tilde{b}}).$$

But,

$$d_{\tilde{b}i} = d_{ci} + \mathbf{a}'_{ci} \mathbf{T} \mathbf{m} = d_{ci} + \mathbf{a}'_{\tilde{b}i} \mathbf{m} \Rightarrow \mathbf{d}_{\tilde{b}} = \mathbf{d}_c + \mathbf{A}_{\tilde{b}} \mathbf{m}.$$

Therefore,

$$Q = (\mathbf{d}_b - \mathbf{d}_c - \mathbf{A}_{\tilde{b}} \mathbf{m})' (\mathbf{d}_b - \mathbf{d}_c - \mathbf{A}_{\tilde{b}} \mathbf{m}).$$

Expanding this expression,

$$Q = (\mathbf{d}'_b \mathbf{d}_b + \mathbf{d}'_c \mathbf{d}_c - \mathbf{m}' \mathbf{A}'_{\tilde{b}} \mathbf{A}_{\tilde{b}} \mathbf{m} - 2\mathbf{d}'_b \mathbf{d}_c - 2\mathbf{d}'_b \mathbf{A}_{\tilde{b}} \mathbf{m} + 2\mathbf{d}'_c \mathbf{A}_{\tilde{b}} \mathbf{m}).$$

To simplify this expression, the quantity

$$\mathbf{d}_{c-b} \equiv \mathbf{d}_c - \mathbf{d}_b$$

is defined, and

$$Q = (\mathbf{d}'_b \mathbf{d}_b + \mathbf{d}'_c \mathbf{d}_c - \mathbf{m}' \mathbf{A}'_{\tilde{b}} \mathbf{A}_{\tilde{b}} \mathbf{m} - 2\mathbf{d}'_b \mathbf{d}_c + 2\mathbf{d}'_{c-b} \mathbf{A}_{\tilde{b}} \mathbf{m}).$$

Taking the derivative of Q with respect to \mathbf{m} gives

$$\frac{\partial Q}{\partial \mathbf{m}} = \frac{\partial (2\mathbf{d}'_{c-b} \mathbf{A}_{\tilde{b}} \mathbf{m} - \mathbf{m}' \mathbf{A}'_{\tilde{b}} \mathbf{A}_{\tilde{b}} \mathbf{m})}{\partial \mathbf{m}},$$

since the terms that have disappeared do not contain \mathbf{m} . Evaluating this derivative gives

$$\frac{\partial Q}{\partial \mathbf{m}} = 2\mathbf{d}'_{c-b}\mathbf{A}_{\tilde{b}} - 2\mathbf{m}'\mathbf{A}'_{\tilde{b}}\mathbf{A}_{\tilde{b}}.$$

Setting this expression equal to zero, and solving for \mathbf{m} ,

$$0 = 2\mathbf{d}'_{c-b}\mathbf{A}_{\tilde{b}} - 2\mathbf{m}'\mathbf{A}'_{\tilde{b}}\mathbf{A}_{\tilde{b}}$$

$$0 = \mathbf{d}'_{c-b}\mathbf{A}_{\tilde{b}} - \mathbf{m}'\mathbf{A}'_{\tilde{b}}\mathbf{A}_{\tilde{b}}$$

$$\mathbf{m}'\mathbf{A}'_{\tilde{b}}\mathbf{A}_{\tilde{b}} = \mathbf{d}'_{c-b}\mathbf{A}_{\tilde{b}}$$

$$\mathbf{m}' = \mathbf{d}'_{c-b}\mathbf{A}_{\tilde{b}} \left(\mathbf{A}'_{\tilde{b}}\mathbf{A}_{\tilde{b}} \right)^{-1}.$$

This method of solving for \mathbf{m} resolves the origin indeterminacy of MIRT equating without the considerable rounding errors of the Li and Lissitz (2000) approach.

All of this, however, still leaves the meaning of the dimensions unidentified. Identifying the meaning of MIRT dimensions is a difficult task because of the empirical nature of meaning identification. A typical method of identifying the meaning of dimensions is to cluster-analyze a proximity matrix of angles between item discrimination vectors in the multidimensional space Kim (2001). Kim suggests that more clusters are needed than dimensions because items may cluster based upon particular mixes of the dimensions measured by the clusters of items, and this mixing of dimensions may produce more clusters than dimensions. Kim suggests hierarchical cluster analysis using the Ward method to identify clusters of similar items, and the identification of the characteristics of the items in each cluster that cause those items to cluster together.

After identifying the clusters of items and their meaning, a reference composite is obtained for each cluster (see Wang, 1985). A reference composite can alternately be obtained for the items on each grade level of the assessment to determine the degree to which the emphases on the various dimensions change over grade levels. The reference composite is obtained by taking the first eigenvector (or first principal component) of the quantity $\mathbf{A}'\mathbf{A}$, where \mathbf{A} is the item by dimension matrix of discriminations for the group of items needing a reference composite. This reference composite represents the composite

multidimensional direction in which a group of items best discriminates on student achievement.