An Investigation of Dimensionality Across Grade Levels and Effects on Vertical Linking for Elementary Grade Mathematics Achievement Tests

Samantha S. Burg, Ph.D.

AN INVESTIGATION OF DIMENSIONALITY ACROSS GRADE LEVELS AND EFFECTS ON VERTICAL LINKING FOR ELEMENTARY GRADE MATHEMATICS ACHIEVEMENT TESTS

ABSTRACT

It is a widely held belief that mathematical content strands reflect different constructs which produce multidimensionality in mathematical achievement tests for Grade 3-8. This study analyzes the dimensional structure of mathematical achievement tests aligned to NCTM content strands using four different methods for assessing dimensionality. The effect of including off-grade linking items as a potential source of dimensionality was also considered. The results indicate that although mathematical achievement tests for Grades 3-8 are complex and exhibit some multidimensionality, the sources of dimensionality are not related to the content strands or the inclusion of several off-grade linking items. The complexity of the data structure along with the known overlap of mathematical skills suggest that mathematical achievement tests could represent a fundamentally unidimensional construct.

INTRODUCTION

The psychometric models used in the context of many achievement tests assume a unidimensional construct is being measured. That is, in the context of measuring student achievement, most tests are considered to measure one latent trait, construct or ability (i.e., *unidimensional*). Other tests are designed to measure a combination of abilities (in which it is referred to as *multidimensional*). In either context, the dimensional structure of a test is intricately tied into the purpose and definition of the construct to be measured. However, it is sometimes the case that a test that is intended to be unidimensional may unintentionally be measuring more than one latent variable.

The consequences of violating the assumption of unidimensionality have important implications on many facets of the test development process including parameter estimation, vertical scaling, and gathering validity evidence. Test items and student performance are analyzed using mathematical models such as IRT or MIRT which assume a certain dimensional structure. Therefore, misdiagnosis or misrepresentation of the dimensional structure can impact model parameter estimates including person ability estimates (i.e., student scores). The dimensional structure of a test is also used to provide one type of validity evidence based upon the internal structure of a test. Because validity refers to the degree to which evidence and theory support the interpretations of test scores, it is a fundamental consideration in test development. Modeling student growth and adequately yearly progress have also become important considerations in a testing program. This has necessitated the use of vertical scales that model the mathematical developmental continuum across grades and content standards. While previous research on the consequences of violating the assumption of unidimensionality has been inconclusive due to differences about definitions and evidence of dimensionality, it seems that eliminating any error is advantageous with so many high-stakes associated with the test results.

Unintentional sources of multidimensionality may exist particularly in a complex subject like mathematics due to the subject matter as well as the typical curriculum standards. Most states have adopted the National Council of Teachers of Mathematics' (NCTM) guidelines presented in *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics, 2000, hereafter NCTM *Standards*). The NCTM *Standards* highlight the growth of expectations in five content areas (called “strands”): Number Sense and Operations, Algebra, Geometry, Measurement, Data Analysis and Probability. It is not expected that every topic would be addressed to the same extent instructionally each year as shown in Figure 1; rather, students would develop a certain depth of understanding of concepts and acquire certain levels of fluency in a curriculum so that subsequent instruction can build on this understanding. For example, the curriculum for students in earlier elementary school would have a heavier focus on Number Sense and would introduce the simple ideas of Algebra. As the students progress through elementary school toward middle school, the curricular emphasis changes; instructional time spent on Number Sense and Operations would decrease while the focus on Algebra would increase.

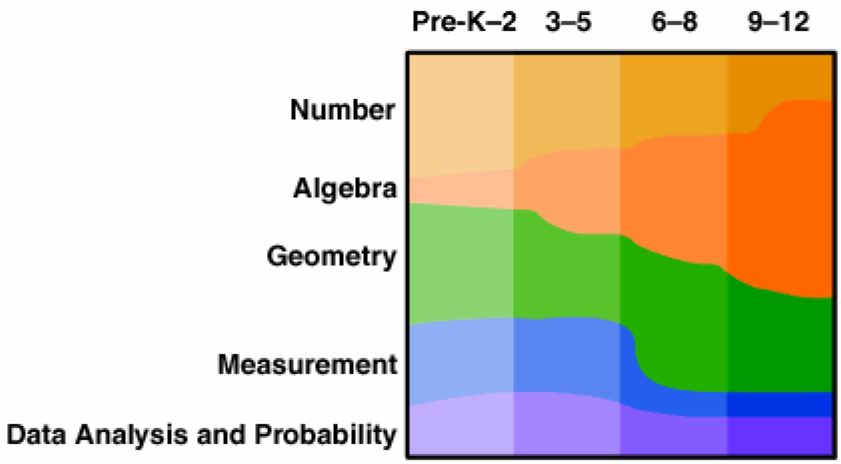


Figure 1. NCTM Content Standards Across the Grade Bands (National Council of Teachers of Mathematics, 2000)

Source: (National Council of Teachers of Mathematics, 2000)

While the instructional emphasis of the different mathematics strands changes over a typical mathematics curriculum, standardized tests report a single mathematics achievement or proficiency score at each grade. Because "achievement tests that are constructed with an emphasis on content specifications are likely not to be unidimensional" (Reckase, Davey, & Ackerman, 1989, p.2), further research is needed to explore the unintentional sources of multidimensionality that may arise due to mathematics test construction traditions that follow the NCTM *Standards* and explore whether test dimensionality changes with the grade appropriate curriculum.

In addition to the mathematical subject matter, unintentional sources of multidimensionality may also be introduced when developing a vertical scale and the inclusion of off-grade-level items on a test. Therefore further understanding of test dimensionality, sources of multidimensionality, assessment of dimensional structure, and consequences of violations of dimensionality assumptions is warranted. The purpose of this study was twofold; 1) to examine the stability of the dimensional structure across elementary grades mathematics achievement tests; and 2) to investigate the dimensional structure of these mathematics achievement tests in situations where vertical linking items (below and above grade level) are included in on-grade level tests. In addition, due to the multi-method analysis of the study, a comparison of the different methods of assessing dimensionality was also explored.

METHODOLOGY

This study used data collected in February 2004 as part of a large-scale field study. Several school districts across the country agreed to participate in the study which resulted in a large and diverse sample of elementary and middle schools students. Data were collected on a total of 9,165 students in grades 2 through 9 in 34 schools from 14 districts across six states (California, Indiana, Massachusetts, North Carolina, Utah, and Wisconsin). The participants were diverse in their geographical location as well as the size and type of community (e.g., suburban; small town, city or rural communities; and urban).

Field tests were administered at each grade level. All items in each form were multiple-choice format and dichotomously scored. Each field test form consisted of 30 multiple choice items which included both on-grade level items and a common block of items from out of grade level (below- and above–grade level) for vertical scaling. For example, the grade 4 form included Grade 4 (on-grade) items, and items from Grade 3 and Grade 5 (off-grade) as well. Each of the mathematics achievement tests was developed in the same way including attention to content specification, item writing and review, and field testing. The content specifications required that the items be aligned with the five content strands suggested in the NCTM framework (NCTM, 2000) which are as follows: (1) Numbers and operations, (2) Geometry, (3) Algebra/Patterns and functions, (4) Data analysis and probability, and (5) Measurement. All items were written and reviewed by trained item writers who were experienced mathematics educators and item-development specialists and therefore familiar with mathematical achievement of students at various grade levels. Items were then reviewed by content and psychometric experts to ensure quality of the response options and sensitivity issues.

Researchers do not agree on a single method for assessing test dimensionality and therefore, assessment of dimensionality was analyzed using four popular approaches. The approaches included two parametric approaches (item factor analysis: NOHARM and principal component analysis: WINSTEPS and two nonparametric approaches (assessment of essential dimensionality:DIMTEST and conditional covariance: DETECT). All four approaches have been shown to be effective indices of dimensional structure.

RESULTS

Using real test data and applying a variety of popular dimensionality assessment methods, the test structures of mathematical achievement tests were examined across Grades 3-8. Both exploratory, confirmatory or a combination of both approaches were used when appropriate. The first research question required analyses using on-grade items only. Therefore, only Grade 3 items were considered for the assessment of the Grade 3 test structure, only Grade 4 items for the Grade 4 test, etc. The second research question included off-grade level items which is typical of a vertical scale linking design. The results related to the first research question (on-grade items) are presented first, followed by those for research question two (off-grade items). The final section in this chapter offers a comparison of the different solutions and approaches as stated in research question three.

Results for Dimensional Structure across Grades

To explore data pertaining to Research Question 1 (the dimensional structure across mathematical achievement tests), each set of on-grade items were analyzed for possible sources of dimensionality related to five mathematical content strands. The analyses were also used to compare test structures across grades. The original expectation was the tests would be essentially unidimensional or would exhibit only modest amounts of multidimensionality due to the different strands.

Conditional Item Covariance and On-Grade Items

The results for applying a conditional covariance analysis approach using the software program DETECT are shown Table 1. The results include the DETECT Index (D_{max}) which indicates the amount of multidimensional simple structure; (2) the r_{max} index which indicates whether the data are displaying simple or complex structure; and (3) the number of clusters needed to maximize D_{max} where the number of clusters is theoretically equal to the number of dominant abilities or dimensions of the test. However, one condition must be noted about the relationship between dimensions and clusters: the number of dominant abilities measured by the test is indicated by the number of clusters *only* in the optimal partition of items for a test that is essentially multidimensional and exhibits simple structure. Overall, the results indicated that the on-grade items exhibit weak to moderate amounts of multidimensionality and a complex structure (i.e., some item responses are effectively determined by more than one ability).

Table 1. *Results of Conditional Covariance Analysis (DETECT) of On-Grade Items*

Grade	D_{max}	r_{max}	No. of Clusters
Grade 3	0.4558	0.5534	5
Grade 4	0.4905	0.6032	4
Grade 5	0.4148	0.4998	5
Grade 6	0.4550	0.5204	5
Grade 7	0.6536	0.6119	4
Grade 8	0.5631	0.6197	5

Zhang and Stout (1999) found that while the clusters partitioned by DETECT are more accurate when r_{max} is greater than 0.80 (i.e., approximate simple structure), DETECT is still very informative when approximate simple structure fails to hold. Therefore, the clusters were examined further but caution should be exercised when interpreting the cluster results. The clusters for the Grade 3 on-grade items are shown in Table 2. The last row displays the total number of items per cluster. The subsequent rows show the number of items per strand in each cluster. For example, Cluster 1 consisted of 12 items (out of 26 items on the form). Four of those items were from the Numbers and Operations strand, one item from the Geometry strand, five items from the Algebra and Pattern Recognition strand, and one item each from the Data Analysis and Probability strand and the Measurement strand. Recall that each item was written to a specific content strand and the test specifications required items from all five strands. The clusters however do not match item designated strands indicating that the item clusters do not appear to be based on the content strands. For example, as can be seen in the table, the eight items that were designated as being in the Numbers and Operations content strand were identified by DETECT as failing to cluster together as intended, but were distributed across three clusters: Cluster 1, Cluster 2, and Cluster 3. The clustering of items for the other grades were similar to the clusters for Grade 3 in that item clusters did not appear to be strand-based.

Table 2. *Distribution of Strand-Designated Items by Cluster and Content Strand for Grade 3*

	Distribution of Strand-Designated Items by Cluster					
Content Strand	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Numbers & Operations	4	3	1	0	0	8
Geometry	1	0	0	1	1	3
Algebra & Patterns	5	0	1	0	0	6
Data Analysis & Probability	1	2	0	0	0	3
Measurement	1	2	0	2	1	6
Total Number of Items in Cluster	12	7	2	3	2	26

Assessment of Essential Dimensionality of On-Grade Items

The second method used to assess potential changes in dimensional structure across the grade levels studied was an assessment of essential dimensionality. DIMTEST uses Stout's *T* statistic for a nonparametric test of unidimensionality. The *T* statistic is used to test the null hypothesis that a set of items is essentially unidimensional. The p-values from applying confirmatory DIMTEST (based on strands) are presented in Table 3. In summary, when a set of items based on content strand was compared to the items on the rest of the test, the null hypothesis of essential unidimensionality could not be rejected for all strands in Grades 3, 4, 6, 7 and 8. In other words, subsets of items based on content were not dimensionally different from the remaining items suggesting that the data are essentially unidimensional. However, Grade 5 results displayed a slightly different story. The items designated as Numbers and Operations for Grade 5 suggest a potentially different dimension than the remaining Grade 5 items from the other four strands.

Table 3. *P-values from DIMTEST Using On-grade Items*

Content Strand	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Numbers & Operations	0.8497	0.1290	0.0660	0.1961	0.2605	0.1218
Geometry	0.2742	0.3558	0.0154	0.3299	0.6492	0.2822
Algebra and Patterns	0.1133	0.1122	0.3674	0.4419	0.0354	0.6955
Data Analysis & Probability	0.9863	0.4373	0.1655	0.8989	0.6453	0.1827
Measurement	0.1038	0.6310	0.4281	0.4253	0.0243	0.9407

Nonlinear Item Factor Analysis of On-Grade Items

The third method used to answer Research Question 1 (that is, potential changes in dimensional structure across grade levels) was a nonlinear item factor analysis approach as employed by the software program NOHARM. NOHARM computes the residual covariances of the items after fitting a model (the user specifies the number of dimensions) and calculates the root mean square of the residual covariances as an overall measure of misfit of the model to the data. In other words, the residual matrix offers an indication of how well the principle of local independence has been satisfied given the prescribed model.

Initially, a confirmatory analysis was conducted in NOHARM. The hypothesis of five dimensions (based on content strands) was tested. The results for each grade are shown in Table 4. The root mean square residual (RMSR) is an indicator of model fit; RMSR=0 indicates a perfect model fit and increasingly higher values indicate worse fit (Kline, 2005). The RMSR values were relatively small across the grades, ranging from 0.0089 to 0.0174, signifying very little misfit of the data to a five-dimensional model. Tanaka's index is another fit index and it ranges from 0 to 1; while there are no specific interpretive guidelines, better fit is indicated by values closer to 1 (Tanaka, 1993). Tanaka's index was higher in Grades 3 and 4 than Grade 5-8 indicating a better fit for a 5-dimensional model in the lower grades than the higher grades.

Table 4. *Confirmatory Nonlinear Item Factor Analysis Results (NOHARM) for On-Grade Items (Five-Dimensions)*

Grade	RMSR	Tanaka's Index
Grade 3	0.0101	0.9568
Grade 4	0.0089	0.9556
Grade 5	0.0174	0.8950
Grade 6	0.0151	0.9098
Grade 7	0.0174	0.8931
Grade 8	0.0142	0.9159

Note. The five dimensions were based on the five mathematical content areas.

To further investigate the structure of the tests, the resulting factor loadings produced by NOHARM in an exploratory five-dimensional case were examined for patterns among the factor loadings and content strands. Varimax rotation was used after factor extraction to maximize high correlations and minimize low ones. This orthogonal rotation was selected to explore distinct, uncorrelated dimensions that would be expected if the content strands represented different constructs or abilities. Correlated factors make interpretation of the factor loadings difficult. Furthermore, in a recent study using a Monte Carlo simulation, Finch (2006) compared the factor recovery performance for Varimax and Promax methods of rotation using NOHARM. His results suggested the two approaches were equally able to recover the underlying factor structure, regardless of the factor correlations.

A summary of the number of items by content strand and factor for Grade 3 is shown in Table 5. The last row displays the number of total items that load on each factor obtained in these analyses. The other rows in each table show the number of items that load on each factor by the intended content strand. For example, the last column of Table 5, indicates that 8 of the 26 items on the test were intended to measure the Number and Operations strand. However, as can be seen in the first row of the table, three of those items loaded on Factor 1, three items loaded on Factor 2, and two items loaded on Factor 4. Overall, the results were similar for Grades 4-8. That is the items do not tend to load according to the content strands as expected if a potential source of multidimensionality was due to differences in strand or content specificity.

Table 5. *Summary of NOHARM Factor Loadings by Content Strand*

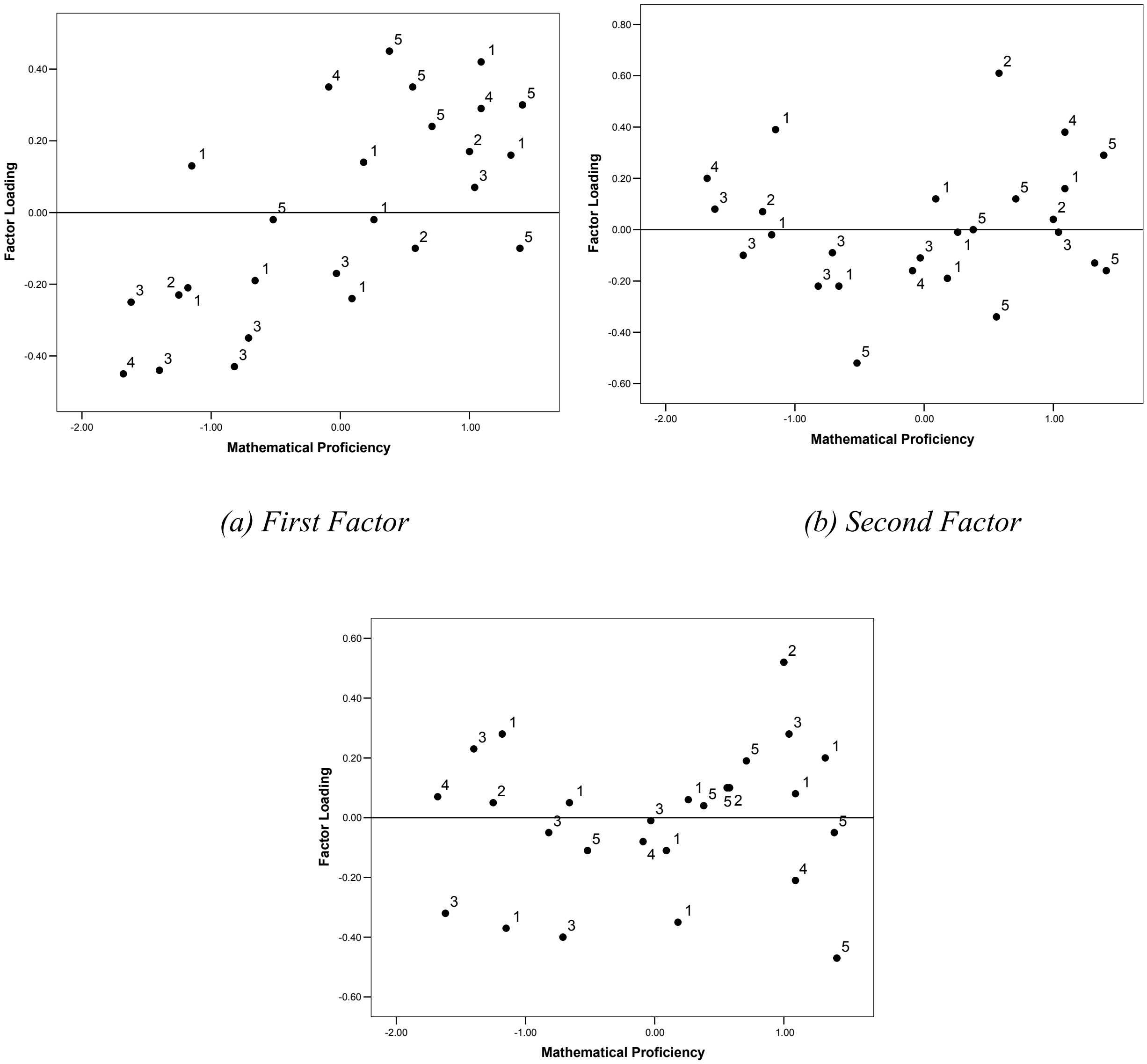
Grade 3	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Total
Numbers & Operations	3	3	0	2	0	8
Geometry	0	1	0	1	1	3
Algebra & Patterns	0	6	0	0	0	6
Data Analysis & Probability	1	1	0	0	1	3
Measurement	2	0	2	2	0	6
Total	6	11	2	5	2	26

Principal Components Analysis of On-Grade Items

The final method used to explore potential changes in dimensional structure across grades (that is, Research Question One) was a principal components analysis (PCA) of residuals. Using PCA, the software program WINSTEPS identifies secondary dimensions in the data by the decomposition of the observed residuals. Residuals are the deviations of the observed data from the predicted values based on the Rasch model which is a one-dimensional measurement system. High correlation of residuals for two items indicates that they may not be locally independent. That is, both items may be measuring some other shared dimension.

WINSTEPS provides two ways to look at model fit: eigenvalues and principal components factor plots of the standardized residuals. Overall the first residual factors do not show much strength; the subsequent factors show even less strength. The first residual factor of Grades 4 and 7 accounted for the most unexplained variance (2.2 eigenvalue units), followed by Grades 3 and 8 (2 eigenvalue units) and Grades 5 and 6 (1.6 eigenvalue units). Previous simulation studies have shown that random data (i.e., noise) can have eigenvalues of size 1.4 therefore WINSTEPS and PCA analysis use 1.4 as a cutoff value (Linacre, 2005). That is, a residual factor with an eigenvalue greater than 1.4 could potentially be a valid factor (i.e., enduring or repeatable structure) but if its eigenvalue is less than 1.4 then it most likely noise, random error, etc. Generally, the results indicate that after the unidimensional model has been applied to the data, there is little evidence of structure--that is, additional dimensions--in the residuals for Grades 3 – 8.

The principal components factor plots of the standardized residuals were also analyzed. Figure 2 a-c shows the first, second and third residual factor plots respectively for Grade 3 on-grade items. The X-axis is the measurement axis (i.e., the posited single dimension). This dimension has been extracted from the data prior to the analysis of the residuals. The items are labeled with their content strand designation: (1) numbers and operations, (2) geometry, (3) algebra and patterns, (4) data analysis and probability and (5) measurement. The trend in Figure 2 (a) shows a positive correlation between Rasch item measures and factor loadings. However, notice that this trend disappears as the second and third factors are analyzed (Figure 2 b and c). The WINSTEPS results for Grade 4 - 8 were similar to Grade 3 results shown in Figure 2.



(a) First Factor

(b) Second Factor

(c) Third Factor

Figure 2. Principal Components (Standardized Residual) Factor Plots of Grade 3 On-Grade Items

Summary of Dimensionality of On-grade Items across Grades 3-8

In summary, nonlinear tem factor analysis using NOHARM and principal component analyses using WINSTEPS show some evidence of multidimensionality but the results from the assessment of dimensionality employed in DIMTEST purport that the multidimensionality does not appear to be related to the five mathematical content strands. The number of potential dimensions seemed to vary slightly and randomly across Grades 3-8. That is, there does not seem to be relationship among the number of potential dimensions and grade level. However, the results suggest that overall the five content strands are not possible sources of dimensionality of mathematics achievement tests for Grades 3-8.

Results for Inclusion of Linking Items

The second research question considered the possible change in dimensional structure within a grade level test due to the inclusion of off-grade (above and below grade) level linking items. The inclusion of off-grade items is a widely used method for developing a vertical scale to span two or more grades. The number of off-grade items included in the grade level forms examined in this study was very small (typically two to four items), although this, too, is typical of vertical scaling designs in K-12 educational achievement testing.

Conditional Item Covariance and Inclusion of Linking Items

The first method used to assess potential changes in dimensional structure due to the inclusion of linking items was an analysis of conditional item covariances. Exploratory DETECT was applied to off-grade item data using two different runs. First, on- and below-grade items were explored and then data for on- and above-grade items were examined. The results are presented in Table 6. When below-grade items were included in the DETECT analyses, the D_{max} values indicate weak to moderate multidimensionality and the r_{max} values signifying complex structure. The number of clusters ranged from 4-6. These results were similar to the findings for the on-grade items alone as shown previously in Table 1.

Table 6. Comparison of Test Structure for Including On-Grade and Off-Grade Items Using Conditional Item Covariances (DETECT)

	Below- and On-Grade Items			Above- and On-Grade Items		
Grade	D_{max}	r_{max}	Number of Clusters	D_{max}	r_{max}	Number of Clusters
Grade 3	0.4381	0.5569	5	0.4238	0.5188	5
Grade 4	0.4599	0.5524	5	0.4582	0.5638	5
Grade 5	0.3794	0.4843	4	0.4222	0.4965	4
Grade 6	0.4432	0.5003	4	0.4199	0.4760	5
Grade 7	0.6595	0.6074	4	0.5683	0.5572	4
Grade 8	0.4770	0.5424	6	0.5724	0.5976	4

Assessment of Essential Dimensionality When Off-Grade Items Are Included

A second approach to answering Research Question 2 regarding the potential affects on dimensionality of including linking items involved assessing the essential dimensionality of the data via the computer program DIMTEST. The results of applying DIMTEST when off-grade items are included are shown in Table 7. The results for below- and on-grade items are shown in the shaded rows; results for the above- and on-grade items are presented in the unshaded rows. The first column in the table provides the grade and item combinations and the second column specifies the number of off-grade items included in each grade level form. The last column provides the p-values associated with the T statistics that DIMTEST calculates. As seen in the last column, the p-values generated by DIMTEST do not permit the null hypotheses of unidimensionality to be rejected. That is, for none of the grade levels does the inclusion of off-grade items result in a test that is dimensionally distinct from one that is constructed of on-grade items only.

Table 7. *Assessment of Essential Unidimensionality (DIMTEST) Including Off-Grade Items*

Item Levels	No. of Off-Grade Items	p-value
Grade 3: G2 & 3 Items	2	0.3278
Grade 3: G3 & 4 Items	2	0.5584
Grade 4: G3 & 4 Items	3	0.212
Grade 4: G4 & 5 Items	2	0.1075
Grade 5: G4 & 5 Items	4	0.5300
Grade 5: G5 & 6 Items	2	0.6125
Grade 6: G5 & 6 Items	4	0.9924
Grade 6: G6 & 7 Items	2	0.4672
Grade 7: G6 & 7 Items	4	0.4349
Grade 7: G7 & 8 Items	2	0.5921
Grade 8: G7 & 8 Items	4	0.3675
Grade 8: G8 & 9 Items	2	0.8157

Nonlinear Item Factor Analysis When Linking Items Are Included

Another approach to examining the presence of linking items on dimensional structure (i.e., Research Question 2) is nonlinear item factor analysis. It was hypothesized that there would be two dimensions related to the grade level: one dimension representing on on-grade items and a second dimension resulting from the off-grade level items. Therefore, confirmatory factor analyses using the software program NOHARM and a priori specification of two dimensions was applied to the datasets containing on- and off-grade items. The results for the two-dimensional analyses are presented in Table 8. The results for below- and on-grade items are shown in the shaded rows and the above- and on-grade items are presented in the unshaded rows. The RMSR were small, ranging from 0.0092 to 0.0122 for below- and on-grade items and from 0.0093 to 0.0117 for the above-and on-grade items. Tanaka's Index ranged from 0.9475 to 0.9598 and 0.9414 to 0.9609, respectively. Recall that interpretation is rather limited because currently there are no specific guidelines for RMSR or Tanaka's Index. In general, a good model fit is indicated by a small RMSR (i.e., close to zero) and a high Tanaka's index (closer to 1).

Table 8. *Confirmatory Nonlinear Item Factor Analysis (NOHARM) for Off-Grade Items (Two Dimensions)*

Item Levels	RMSR	Tanaka's Index
Grade 3: G2 & 3 Items	0.0103	0.9541
Grade 3: G3 & 4 Items	0.0106	0.9503
Grade 4: G3 & 4 Items	0.0092	0.9511
Grade 4: G4 & 5 Items	0.0093	0.9513
Grade 5: G4 & 5 Items	0.0101	0.9598
Grade 5: G5 & 6 Items	0.0104	0.9609
Grade 6: G5 & 6 Items	0.0112	0.9455
Grade 6: G6 & 7 Items	0.0114	0.9462
Grade 7: G6 & 7 Items	0.0122	0.9414
Grade 7: G7 & 8 Items	0.0118	0.9476
Grade 8: G7 & 8 Items	0.0110	0.9475
Grade 8: G8 & 9 Items	0.0117	0.9408

Exploratory analyses were also conducted to determine where the off-grade items would load on a two-factor solution if NOHARM selected the factor loadings. The off-grade items do not appear to form a separate factor in either the below- or above-grade items and even appear to load on different factors. The clusterings appeared to be random and no observable pattern in the item types was distinguished. Thus, the results for Grades 3-8 indicate that the presence of a small number of linking items do not appear to change the dimensional structure of the test forms.

Principal Components Analysis for Inclusion of Off-Grade Items

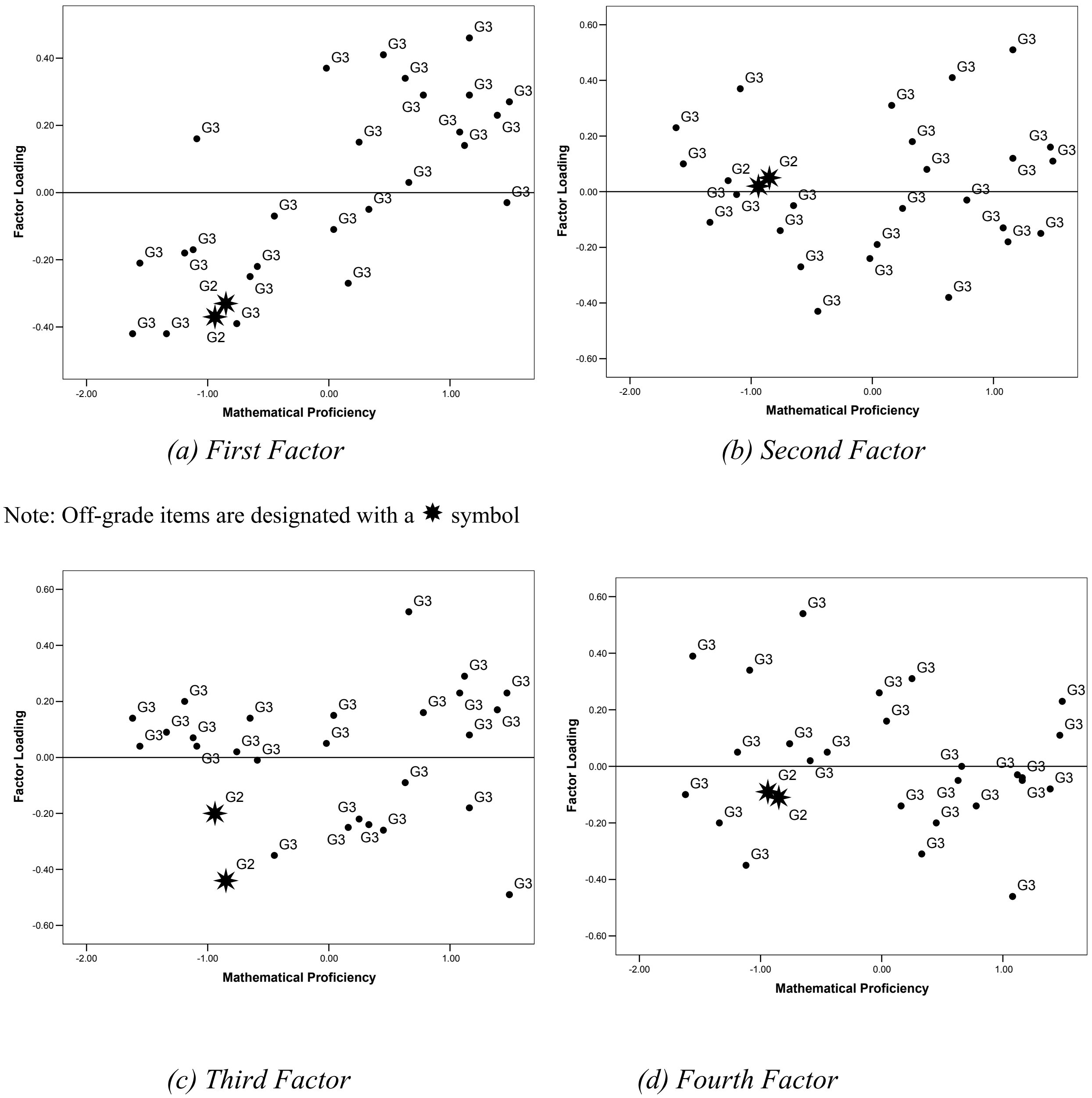
The final method used to assess Research Question 2 (potential influence of off-grade level items on the dimensional structure) was a principal components analysis. The principal components analyses for the off-grade items using WINSTEPS with Grade 3 items is shown in Table 9. For comparison purposes, the first row contains the results from on-grade items only. The next two rows show the eigenvalue units for off-grade items. Note that the amounts of unexplained variance explained by additional factors are similar to the corresponding results for the on-grade items. The residuals from the Grade 2 and 3 items displayed a fourth factor but it the eigenvalue is very small.

Table 9. *Principal Components Analyses Results for Grade 3 On- and Off-Grade Items*

Grade	Total Unexplained Variance (Eigenvalue units)	1st Residual Factor (Eigenvalue units)	2nd Residual Factor (Eigenvalue units)	3rd Residual Factor (Eigenvalue units)	4th Residual Factor (Eigenvalue units)
Grade 3: G3 Items Only	26	2	1.5	1.4	na
Grade 3: G2 & 3 Items	28	2.1	1.4	1.5	1.4
Grade 3: G3 & 4 Items	28	2	1.5	1.5	na

The factor plots of the residuals based on the inclusion of Grade 2 items on the Grade 3 form are shown in Figure 3 a-d. The item labels show the grade level of the item (G2 or G3). The Grade 2 items are also marked with an asterisk (✱) in the figures. These plots were basically identical to the plots for the on-grade items only presented previously in Figure 2 a-c. The first factor (after extracting the primary dimension) plot shows a positive correlation between the mathematical proficiency and the factor loading (Figure 3 a). The other plots of the residuals in Figure 3 (b-d) display residuals that are more random and do not appear to follow a trend which suggests that there is no further important or enduring structure in the data. That is, a unidimensional model appears to fit the data well.

Analyzing Grades 3 and 4 items on the Grade 3 form using WINSTEPS produced the factor residual plots shown in Figure 4 a-c. Notice that the positive trend seen in the first factor of both on-grade and below/on grade items does not appear when items from Grades 3 and 4 are used (Figure 4 a) and the residuals are more dispersed. This random pattern is also seen in the second and third factors.



(a) First Factor

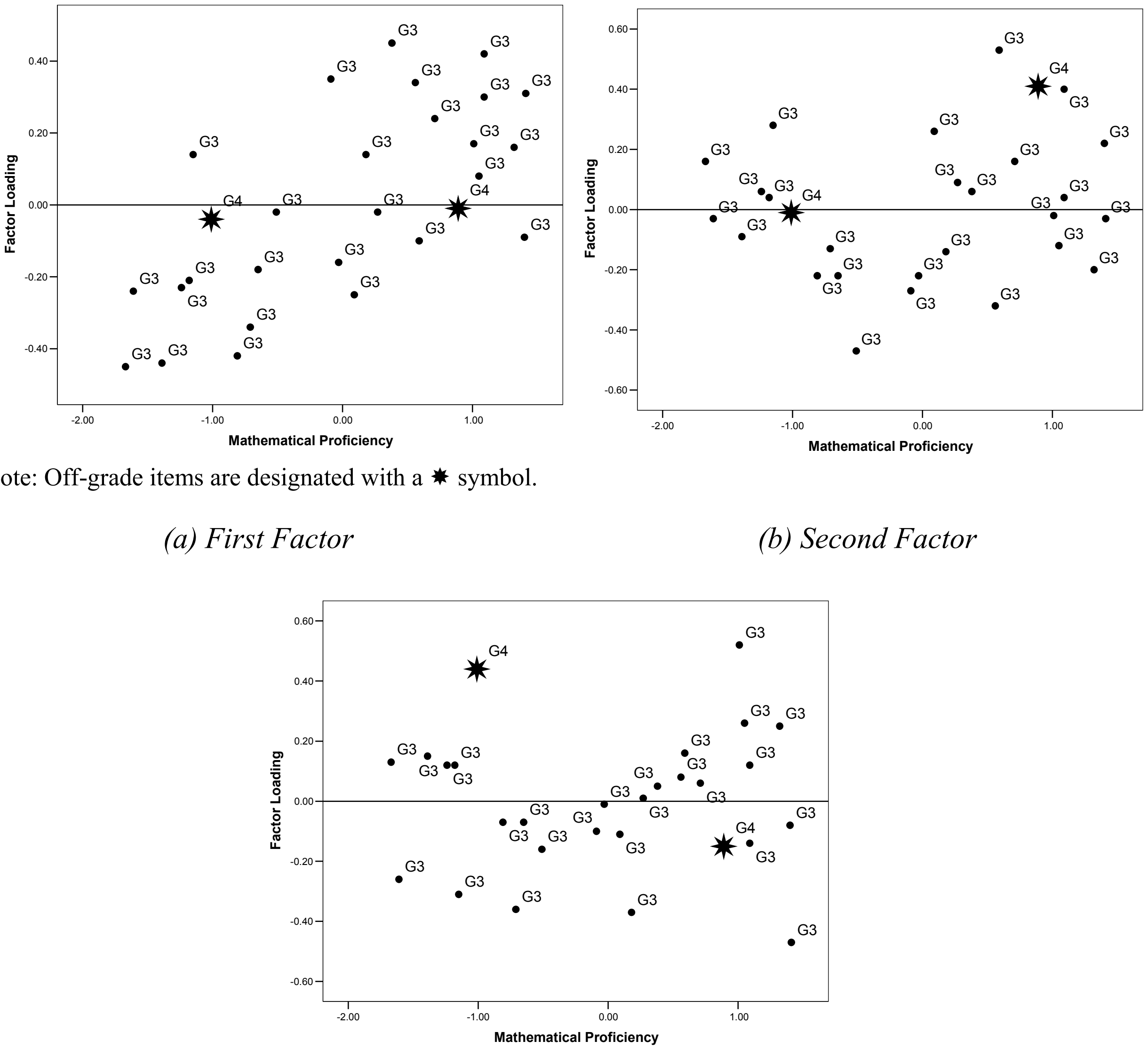
(b) Second Factor

Note: Off-grade items are designated with a ✸ symbol

(c) Third Factor

(d) Fourth Factor

Figure 3. Principal Components (Standardized Residual) Factor Plots of Grade 3: Grade 2 and 3 Items



Note: Off-grade items are designated with a ✷ symbol.

(a) First Factor

(b) Second Factor

(c) Third Factor

Figure 4. Principal Components (Standardized Residual) Factor Plots of Grade 3: Grade 3 and 4 Items

Overall, the inclusion of off-grade items in the test structure analyses did not appear to change the dimensionality results. As in the analysis previously reported regarding the dimensional structure of on-grade items (i.e., research question 1), the software used to gauge dimensionality (DETECT) again identified weak to moderate multidimensionality and complex structure. The inclusion of off-grade items tended to change the clustering of items compared to the clustering that was obtained from analysis of on-grade items alone. According to the results produced by the software program designed to assess essential unidimensionality (i.e., DIMTEST), off-grade items were not dimensionally different from on-grade items which was evidenced by the factor loadings obtained by the nonlinear item factor analysis approach using NOHARM. The principal components analysis of residuals found little structure in the residuals that would suggest the presence of multidimensionality when off-grade items are included in a grade level form.

Comparison of Methods

Research question 3 concerned possible differences in the results of dimensionality analyses yielded by the various approaches and software programs. As expected, the different methods and programs lead to different conclusions about the test structure not only regarding the number of dimensions (as shown in Table 10) but also regarding the items that comprise those dimensions. In addition, the unique pieces of information offered by each program can be combined together to better understand the data structure.

Table 10. *Summary of Overall Exploratory Analyses Using On-Grade Items*

Grade	Conditional Item Covariance (DETECT)	Assessment of Essential Unidimensionality (DIMTEST)	Nonlinear Item Factor Analysis (NOHARM)	PCA Analysis of Residuals (WINSTEPS)
Grade 3	5	>1	2 or 3	1
Grade 4	4	~1	2 or 3	1
Grade 5	5	>1	1	1
Grade 6	5	~1	2 or 3	1
Grade 7	4	>1	4	1
Grade 8	5	>1	5+	1

Summary

Three research questions were explored using data from typical mathematics achievement tests for Grades 3-8. The exploration was conducted using four different approaches: conditional item covariances, assessment of essential unidimensionality, nonlinear factor analysis, and principal components analysis. Research question 1 considered possible influence of five mathematical content areas on the dimensional structure. While the data did display small to moderate amounts of multidimensionality and was complex in nature, this did not appear to be generated by the five content areas. Research question 2 explored the use of off-grade items in a linking project. The scope was rather limited with so few off-grade items but the available data did not appear to be influenced by the inclusion of off-grade items. In regards to Research Question 3, each of the software programs designed to provide information relevant to assessment of test structure appears to offer a unique piece of information to the bigger picture of dimensionality. For example, DETECT estimates the amount of multidimensionality and complexity of the data structure and this information is helpful in interpreting the NOHARM factor loadings where each item loads on each factor (implying a complex structure).

CONCLUSIONS AND DISCUSSION

Before beginning a summary of the key findings of this research, it is important to review some limitations of the study sample, design, and analysis. One limitation of this study was the length of each test (24 -28 items). This limitation is particularly important in regards to Research Question 2 (i.e., the inclusion of off-grade level items on the dimensional structure). Due to the linking study design, each on-grade form contained only a few off-grade items (2-4 items). This linking design was a limitation because more off-grade items could potentially exhibit dimensionality due to content exposure, curricular and/or difficulty factors. In addition, the item format used for all of the mathematics items studies was limited to four-option multiple-choice items; therefore, the results can not be extended automatically to different item formats. In this study, four methods were used for investigating dimensional structure. Each of the four dimensionality assessment methods and programs introduces its own set of limitations as well. For example, two of the approaches (conditional item covariance and assessment of essential unidimensionality) are nonparametric approaches and two methods are parametric (nonlinear factor analysis and principal components analysis). Parametric methods assume a particular parametric model for the IRF while the nonparametric methods assume only that the IRF is monotonic. Therefore, assuming a particular parametric model might or might not fit the data well. One parametric model in particular, the Rasch model (1-PL), has additional limitations. Other IRT models include parameters for differences in item discrimination (2-PL) and guessing (3-PL) but WINSTEPS only employs the Rasch model. It is a possibility that some findings in the study would have differed or other interpretations been plausible had additional parameters been included in item calibrations (e.g., guessing, discrimination). Appendix A contains more information about each program used in this study.

These limitations notwithstanding, this study yielded insights into what is known about the dimensionality of mathematics achievement tests, how that dimensionality is affected when out-of-level linking items are embedded in mathematics achievement tests for the purpose of creating vertical (i.e., across-grade) scales, and how various procedures for assessing dimensionality perform in these contexts. These findings correspond to the three main research questions addressed in this study and the following summary of findings is organized according to those research questions.

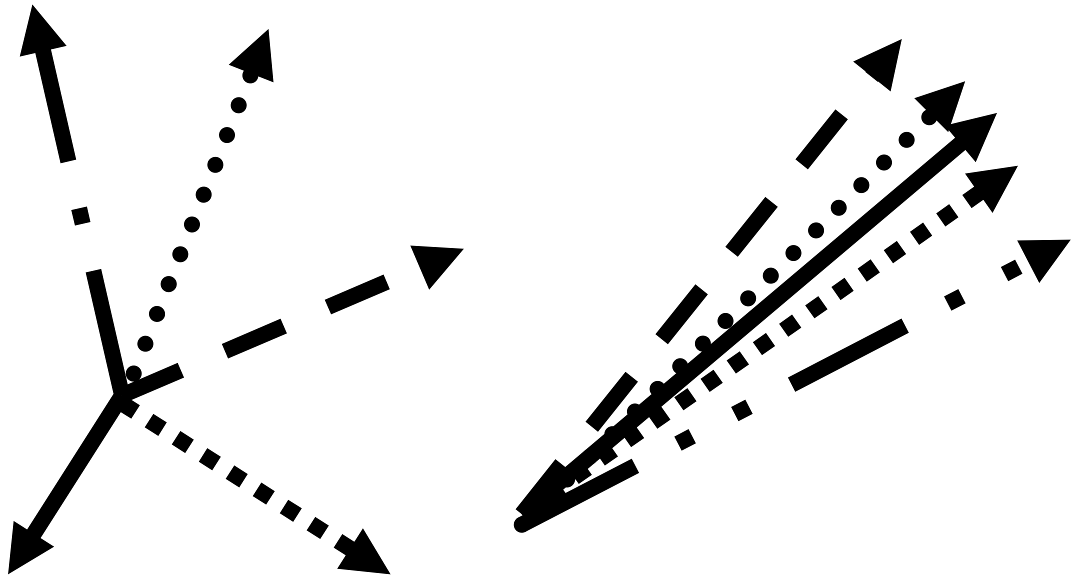
Complex Structure

The results of the conditional item covariance and DETECT's r_{max} index and the factor loadings yielded by the nonlinear item factor analysis operationalized by NOHARM suggested a complex test structure in the mathematics achievement tests across grades 3-8. If each item on a test measures one, and only one dimension, the test structure is labeled as exact or simple structure. If the items load highly on multiple dimensions, then the structure is referred to as a complex structure. When a test exhibits complex structure, some item responses are effectively determined by more than one ability or construct. When complex structure is observed, the type of test, the overall content, and the substantive and cognitive aspects of mathematics curriculum, instruction, language, and other assessment issues must be considered.

Many mathematical skills span content strands and are used in conjunction with other skills and/or in subsequent skills. Mathematics is often conceptualized as being made up of separate strands but this tends to be more an organizing principle for curriculums and textbooks rather than an indication of the structure of multidimensionality in the mathematical achievement construct. The results of this study did not show a relationship between dimensionality and the content strands. Additionally, these findings support the NCTM Connections Standard which proposed that all students (prekindergarten through Grade 12) should be able to make and use connections among mathematical ideas and see how the mathematical ideas interconnect. According to NCTM, "mathematics is not a collection of separate strands or standards, even though it is often partioned and presented in this manner" (National Council of Teachers of Mathematics, 2000, p. 64).

There is, however, a great amount of overlap and correlation in mathematical topics, skills and strands. For example, consider basic addition of whole numbers which is classified as a skill in the Numbers and Operations strand. Knowing addition facts leads to other skills such as (1) subtraction facts (also in the Numbers and Operations strand), (2) finding the mean of a set of data (Data Analysis and Probability strand) and (3) determining whether angles in a figure are complimentary or supplementary (Geometry strand). The last illustration (3) is particularly interesting. There tends to be more distinction or difference between algebra and geometry particularly when geometry involves learning basic shapes, properties of figures or spatial reasoning. However, at some point the content strands intertwine again, as geometry problems require students to use the four basic operations (addition, subtraction, multiplication and division) to find perimeters, areas and volumes or basic algebra skills and algebraic thinking to solve for a missing angle or side length. Thus, given the complex nature of mathematical skills and their correlations, the complex nature of the test structure is not surprising; indeed, it should be expected. The study results reflect the interconnectivity of the strands.

While the determination of complex structure in the data does not indicate the number of dimensions, it does suggest something about interaction of the dimensions. Figure 1 illustrates two possible relationships of factors of a complex structure. Figure 5(a) illustrates less correlation among five factors while Figure 5(b) displays five factors that are more correlated. Regarding the highly correlated factors observed in the mathematics achievement test data analyzed in this study, a relevant analogy, or image is that of a rope. A rope is made up of different fibers or strands that can be distinguished but are wound together to produce one rope as illustrated in Figure 6. If the constructs of a test are represented by fibers of the rope, this analogy shows how several dimensions might seem distinct and yet are woven together so tightly (i.e., correlated) that the minor dimensions blend into a single more prominent cable. Therefore, the complexity of the data structure along with the known overlap of mathematics skills perhaps suggest that mathematics achievement tests could represent a fundamentally unidimensional construct. Importantly, it should be noted here that the phrase, “essential unidimensionality”, is being avoided as it denotes a specific statistical model developed by Stout and Nandakumar (Nandakumar, 1991, 1993; Stout, 1987, 1990; Stout et al., 1996).



(a) Distinct Dimensions (b) Highly Correlated Dimensions

Figure 5. Graphic Representations of Complex Structure and Multidimensionality

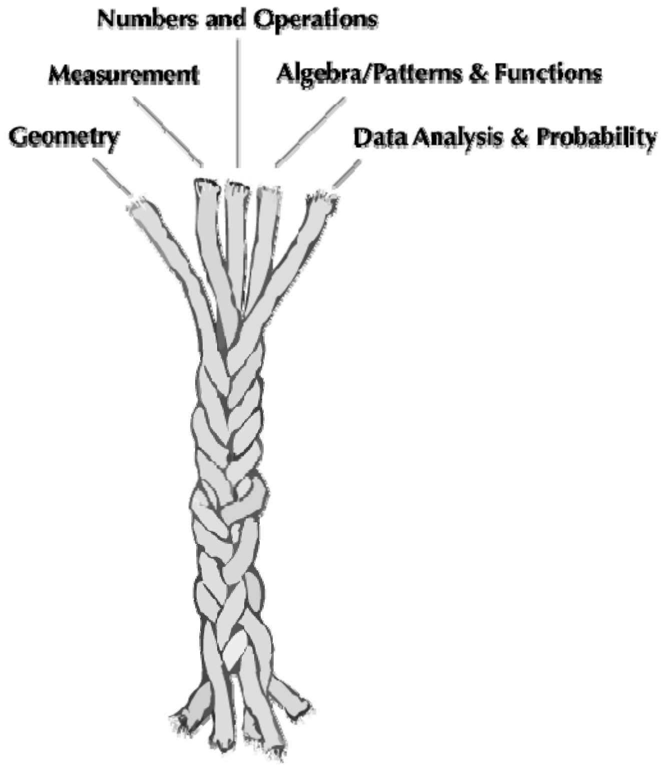


Figure 6. Relationships among Mathematical Strands

Interpretation of Multidimensionality

Although the complex nature of both the mathematical content and mathematical achievement test structure must be acknowledged, it is also important to evaluate the evidence of weak to moderate amounts of multidimensionality in the test data. The response to an item is often dependent upon several secondary dimensions in addition to the hypothesized primary dimension or proficiency (Traub, 1983). Dimensionality is a property of both the test and the examinee population taking the test (Hattie, 1985; Nandakumar & Stout, 1993; Reckase, 1990; Tate, 2002). There are several important features that are examinee-by-instrument interaction that can possibly confound dimensionality: namely, item difficulty and reading demand of mathematical items. Dimensionality can be confounded with item difficulty if the factors represent items with comparable difficulty levels as opposed to items that measure distinct dimensions (Ackerman, Gierl, & Walker, 2003). Dimensionality can also be confounded with reading demand added by the "application" or the desire to place mathematical assessment items within a context. These types of problem solving items contain more verbiage that could require an additional ability (i.e., reading) not essential for the solution of the more decontextualized mathematical computation items. Multidimensionality introduced by reading and language issues may have particular impact on English language learners.

Implications for Practice

The results of this study have several implications for test development and reporting. First, the results of this study support the use and development of vertical scaling. Inclusion of off-grade items used in the common item design does not appear to be potential sources of multidimensionality. Specifically, the results of this study showed that the inclusion of up to four common items, administered above or below one grade, does not substantially alter the dimensional structure of a test. In addition, dimensionality does not appear to be related to content strands for Grades 3-8. Thus, modest changes in the curriculum across grades, in test specifications for contiguous grade levels, or in content standards purposefully developed with the aim of vertical articulation (such as these characteristics were represented in the test development procedures for the tests studied here) should not present a major impediment to the ability to implement a vertical scale.

Second, the results of this study demonstrated a lack of relationship between dimensionality and the intended mathematical content strands. In terms of score reporting, this finding suggests that the common practice of reporting separate strand-based scores (i.e., a score for Numbers and Operations, another score for Measurement, etc.) does not have strong psychometric support. Alternatively, some researchers have recently suggested that accumulating information from items outside of those within an intended content strand shows promise as a means of enhancing the validity and utility of strand-based scores (Edwards & Vevea, 2006). Regardless of the eventual contribution of augmentation approaches, it is clear that content strands are useful for organizing curriculums and test specifications and therefore have utility independent of their dimensional structure.

The lack of relationship between dimensionality and the intended mathematical content strands suggest that the NCTM *Connections* standard may be functioning as intended. That is, the items developed for the mathematics tests used in this study appear to require students to make connections across the five different content strands. These results should encourage teachers, schools, and curriculum materials to continue to emphasize and build upon these connections to deepen students' mathematical reasoning skills and conceptual understanding. Rather than teach a skill one time and typically out of context, it should be reviewed when it comes up again and particularly when it is used in a context. For example, students learn how to add, subtract, multiple and divide integer numbers (numbers and operations strand) and are typically taught these as stand alone skills. However, working with integers becomes critical when learning to solve one- and two-step algebraic equations and integers are important when finding distances in the coordinate plane during a geometry lesson. It is important that the curriculum and textbooks work with teachers to build these connections for the students. It is also important teachers have a chance to explore these connections either with other mathematics teachers in group or lesson discussions or during professional development workshops which focus on the developmental, essentially unidimensional nature of mathematics.

The results of this study also emphasize the connectedness of mathematical topics such that knowing how mathematical skills build and relate to one another could be useful in other ways. Diagnostic information and determination of a potential need for early intervention strategies would be greatly aided by knowing how to approach mathematical skills and topics by bringing in related skills that a student better understands or feels more confident. It is vital to prevent students from falling behind in their mathematical proficiency, becoming frustrated or math anxious or a combination thereof.

Conclusions

This research study, like other studies involving educational data, shows how important the assessment of dimensionality is to a testing program and yet how intricate and complex the task is. It does not however preclude a testing program from periodically assessing "whether the test assembly process is producing tests that are in accord with the test construction blueprint" (Dorans & Lawrence, 1999, p.5) or from conducting periodic checks of the stability of a common scale over time as proposed in Standard 4.17 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Refining the definition of "*dimensionality*" to be considered as "*detectable dimensionality*" integrates two important characterizations of dimensionality: psychological meaning and statistical fit. It is only when these two components support one another that the true test structure can be assessed and interpreted and perhaps more importantly that the implications for the educational process be clarified.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53.

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

Dorans, N. J., & Lawrence, I. M. (1999). *The role of the unit of analysis in dimensionality of assessment*. Princeton, NJ: Educational Testing Service.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal for Educational and Behavioral Statistics, 31*, 241-260.

Finch, H. (2006). Comparison of the performance of Varimax and Promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*, 39-52.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17*, 29-38.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18*(1), 41-68.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston,VA: Author.

Reckase, M. D. (1990, April). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Reckase, M. D., Davey, T., & Ackerman, T. A. (1989, April). *Similarity of the multidimensional space defined by parallel forms of a mathematics test.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimates. *Psychometrika, 55*, 293-325.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariances-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.

Tanaka, J. S. (1993). Multifacted concepts of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.

Tate, R. (2002). Test dimensionality. In D. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181-211). Mahwah, NJ: Lawrence Erlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, Canada: Educational Research Institute of British Columbia.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 34*, 213-249.

MetaMetrics®

LINKING ASSESSMENT WITH INSTRUCTION

MetaMetrics, Inc.
1000 Park Forty Plaza Drive, Suite 120
Durham, North Carolina 27713

Phone: 919–547–3400/1–888–539–4537
Fax: 919–547–3401
Web site: www.Quantiles.com