

**Equivalent Test Structure across Grades:
A Multi-Group Confirmatory Factor Analysis Approach**

Peng Lin

Hua Wei

Robert W. Lissitz

University of Maryland

Paper presented at the annual meeting of the American Educational Research Association,
April 12, 2007, Chicago.

Introduction

An important assumption, as required in the construction of vertical scales is that tests designed for different grade levels are measuring the same construct(s). This requires that the content taught and measured be highly homogeneous across grades. It has been argued that this assumption is hardly satisfied in some subject areas such as science and social studies in which the content covered at each grade varies so much that “a vertical scale may not be able to capture the grade-specific dimensions that may be of considerable importance” (Huynh & Schneider, 2005, p. 101). In other words, if the content changes significantly from one grade to the next, no single underlying construct exists that would lend support to linking performance across grades.

Even in subject areas such as reading and mathematics which are continuously taught over the entire span of elementary grades and have substantial content overlap across grades, it is hard to make an argument that the construct(s) measured in the tests across grades is strictly equivalent. In statewide testing programs, assessments are deliberately aligned with state content standards, which are organized under a number of different topics or content strands. For example, a state level mathematics accountability test may be developed to assess knowledge in data analysis, geometry, measurement, number sense, and algebra. The number of content strands may vary from grade to grade when new topics are introduced or old topics are removed from the curriculum.

In addition, the content strands may not receive equal amounts of curricular emphasis at different grade levels. Nor are they equally represented in the assessments across grades. If we assume that the test construct for each grade is a weighted composite of the content strands covered in that grade, the weights of the content strands will determine the nature of the overall test construct. We believe that differences in the

weights lead to differences in the test construct. For example, suppose that the mathematics tests for grades 3 and 4 in a test battery measure the abilities of data analysis and number sense. Suppose the test for grade 3 puts equal emphasis on both abilities, while the test for grade 4 lays more stress on data analysis. As a result, the test for grade 4 contains more items developed to measure the ability or achievement related to data analysis than that for grade 3. This example is illustrated in Figure 1, where θ_1 represents the ability of number sense, θ_2 represents the ability of data analysis, θ_{3c} represents the composite of θ_1 and θ_2 in grade 3, and θ_{4c} represents the composite in grade 4. In grade 3, items I_1 , I_2 , and I_3 are more sensitive to θ_1 , and items I_4 , I_5 , and I_6 are more sensitive to θ_2 . In grade 4, only two items, I_7 and I_8 , are more sensitive to θ_1 , and the other four items, I_9 , I_{10} , I_{11} , and I_{12} are more sensitive to θ_2 . As can be seen from the graph, θ_{4c} leans more toward θ_2 , the ability of data analysis, than θ_{3c} .

 Insert Figure 1 about here

Changes of content standards are observed in statewide testing programs. Martineau and Plank (2005) reviewed the California content standards in reading and discovered that there were shifts in the construct taught and measured across grades. These shifts, as he argued in his study, also existed in the content standards for other subject areas and in other states. Even if the major content strands remain the same across grades, the objectives or sub-strands within each content strand are typically subject to changes from grade to grade. Take the Maryland Voluntary State Curriculum in mathematics as an

example. The content domains covered in the assessment for each grade are the same, but the learning objectives within each content strand vary from grade to grade. Shifts in construct dimensions and changes of substance of the content strands as reflected in the assessments at different grade levels make it unreasonable to assume common constructs across grades.

Vertical scales are usually obtained by using a common-item scaling design with unequal groups. Within a test at each grade level, there are items designed specifically for that grade as well as items for the adjacent grade(s). The common items between two adjacent grades, also known as linking items, allow for statistical manipulations of the raw scores of each grade and yield vertical scale scores that are directly comparable, if the test construct measured at the two grade levels is homogeneous or invariant and the linking items are representative of the test construct. In this way, students' performance on the linking items is used to indicate the amount of growth that occurs from one grade to the next (Kolen & Brennan, 2004).

Students' performance on the linking items is the basis for deriving a function for score conversions, provided that "the statistical properties of each linking item remain unchanged from one grade to the next" (Haertel, 2004, p. 2). As Haertel stated, "if the statistical properties of items are unchanged, then any systematic difference in the proportion of examinees responding correctly from one grade to the next may be attributed to grade-to-grade differences in the ability distribution" (p. 2). However, as discussed above, in some subject areas where there are changes of topics or content emphasis across grades, the statistical properties of the linking items may change from grade to grade. Besides, if content coverage differs between two grades, the linking items

may not represent both grades equally well or bear the same relationship with the rest of the items at each grade.

As discussed above, assuming that the test constructs measured in two adjacent grades are homogeneous, a good linking item should be able to represent those constructs equally well. In this study, we assess the equivalence of the test constructs of pairs of adjacent grade levels by constraining their relationship with the linking items. Whether the linking items bear the same relationship with the two constructs is evidence about the homogeneity of the constructs. A multi-group confirmatory factor analysis is applied to the data matrices of adjacent grades, with constraints set on the linking items.

Data

The data examined in this study came from the mathematics test results of a statewide testing program. The test at each grade level from grade 3 to grade 8 was designed to cover five content strands: Data Analysis, Statistics, and Probability (D); Geometry and Spatial Sense (G); Measurement (M); Number Sense, Properties, and Operations (N); and Pattern, Algebra, and Functions (P). For each grade level, there were twelve different test forms (A-L), each of which was composed of items that were common across forms and items that were unique to each form. Students' total scores were based on their performance on the common items only. The unique items were field test items and not included in the computation of total scores. The last two forms (K and L) were specially developed to facilitate the construction of a vertical scale. The unique items of form K at a particular grade were items designed for its lower adjacent grade while the unique items of form L were items designed for its higher adjacent grade. In other words, the unique items of form K for a grade were chosen from the common items across forms with the lower grade adjacent to the grade in question, and

the unique items of form L were chosen from the common items across forms with the higher grade adjacent to the grade in question. In this study, only the students who took either Form K or Form L of the math test in each grade were examined. Table 1 summarizes the sample sizes for each grade.

Insert Table 1 about here

The test for each grade consisted of multiple-choice and constructed response items, which were distributed across the five content domains. In each of the two test forms we examined in the study, there were 45 common items and 7 field test items. The distribution of the items among the content domains and the number of items of each item type are summarized in Table 2.

Insert Table 2 about here

Methodology

As discussed above, when a test covers more than one content strand, the unidimensional test construct can be considered as a weighted composite of the content strands being covered. Even if the tests of different grade levels cover the same content, differences in the weights of the content strands and changes of substance contained in the content strands across grades may result in heterogeneity of the test constructs. In the mathematics test we examined here, although all grade levels covered the same five content domains, and the content domains appeared to be uniformly represented in the

tests across grades, we can not be certain that the test constructs measured at different grade levels are invariant.

In this study, the multi-group confirmatory factor analysis method was used to test the assumption of construct invariance. For two adjacent grade levels, responses to Form L of the lower grade and responses to Form K of the higher grade were analyzed concurrently since the two forms had 14 items in common (12 multiple-choice and 2 constructed response items), which could link the test performance of the two grades. A three-step procedure was followed in this study.

Step 1: Construct invariance with and without field test items at each grade level

As discussed above, the two test forms we examined consisted of common items and field test items. Students' total scores were computed based on their performance on the common items only. Therefore, the test construct of interest in each grade was represented by the common items only. Before we test the construct invariance assumption by making use of students' responses to the linking items, an assumption needs to be verified that the field items measure the same construct as the common items in each grade. In this step, the analysis was conducted to test if this assumption was true.

The multi-group confirmatory factor analysis technique was applied to the response matrices of each grade. Responses to the common items were treated as Group 1 and the test construct represented by the common items was indicated by θ_{c1} . Responses to both the common and field test items were treated as Group 2 and the corresponding test construct was indicated by θ_{c2} . If the field test items measure the same construct as the common items, θ_{c2} should be equivalent to θ_{c1} . If that is true, the correlation between each common item and the test construct should be invariant between the two groups.

A one-factor confirmatory factor analysis was conducted on both groups simultaneously with constraints set on the factor loadings of the common items. Specifically, the factor loading of each common item in Group 1 was set equal to that in Group 2. Figure 2 illustrates the models that were compared in this step. Since the models with and without constraints were nested, the χ^2 difference between the two models can be tested to evaluate the invariance of test construct with and without the field test items. The analysis was done on both Form L and Form K at each grade level.

 Insert Figure 2 about here

Step 2: Construct invariance between two adjacent grade levels

In this step, invariance of test constructs between two adjacent grade levels was tested using the multi-group factor analysis method that we used in step 1. For a pair of adjacent grades, for example, grade 3 and grade 4, Form L of grade 3 was treated as Group 1 and Form K of grade 4 was treated as Group 2. The test construct being measured was indicated by θ_{3c} in grade 3 and θ_{4c} in grade 4. If the test constructs measured at the two grade levels were invariant, the correlation between a linking item and the test construct measured in grade 3 should be equal to the correlation between that item and the test construct measured in grade 4. This rationale was implemented in the multi-group factor analysis by setting the loadings of the linking items on the test construct equal in both groups. Since the models with and without constraints were nested, the χ^2 difference between the two models can be tested to evaluate the invariance of test construct between adjacent grade levels. Figure 3 illustrates the models

compared in this step. As we mentioned before, in the concurrent analysis of two adjacent grades, there are 14 linking items in total. Please note that the red arrows represent the loadings of the linking items that are common items of the lower grade but are field test items of Form K of the higher grade, while the blue arrows represent the loadings of the linking items that are common items of the higher grade but are field test items of Form L of the lower grade.

 Insert Figure 3 about here

If the test result in Step 2 indicates that the test constructs of grade 3 and grade 4 are invariant, the test scores from the two grades can be transformed onto a vertical scale. If the two constructs differ, we will go on to Step 3 to study the nature of the changes in the test construct.

Step 3: The nature of construct shift between adjacent grade levels

If the test constructs of two adjacent grade levels differ, the changes in the loadings of the on-grade items across grade levels might reflect the nature of changes in the test construct being measured.

All the analyses were conducted with EQS 6.1 (Bentler, 2005). One thing that needs to be mentioned is that, in this study, all the factor analyses were performed on the tetrachoric or polychoric correlations among the items. Theoretically, EQS can estimate the tetrachoric/polychoric correlation matrix, but it has a high requirement for computing resources. When the number of variables gets large, the huge amount of required CPU time makes the computation infeasible. In this study, all the correlation matrices were estimated by SAS.

Results

Results from Step 1

Table 3 shows the result of the χ^2 difference test for each form at each grade level. Since there were 45 constraints (45 on-grade items) in the restricted model, the degrees of freedom of the χ^2 difference statistic was equal to 45. Results indicated that the linking items in either Form K or Form L measured the same test construct as the on-grade items in each grade. Therefore, we proceeded to the next step of analysis.

 Insert Table 3 about here

Results from Step 2

Table 4 summarizes the results from the multi-group factor analyses conducted to test the construct invariance assumption between adjacent grade levels. Since there were 14 constraints (14 linking items) in the restricted model, the degrees of freedom of the χ^2 difference statistic was equal to 14. The test results indicated that none of the pairs of adjacent grade levels shared exactly the same construct. The largest difference seemed to exist between grade 3 and grade 4, where the χ^2 statistics reached 624.34 with a p-value far less than the nominal level (.05). For the other four pairs of adjacent grade levels, the χ^2 difference ranged from 132.682 to 203.594. Therefore, we proceeded to the next step of analysis.

Insert Table 4 about here

Results from Step 3

Table 5 shows the average of the loadings of the on-grade items within each content strand at each grade level. These results are also plotted in Figure 4, which makes it easier to identify the pattern of changes in the loading averages. The loading average of content D decreased from grade 3 to grade 6, increased a little in grade 7, and decreased sharply in grade 8. The loading average of content G did not change much from grade 3 to grade 4. It decreased from grade 4 through grade 6 and then increased in grade 7 and grade 8. The change in the loading average of content P exhibited a trend opposite to that found with content G. It started to increase from grade 4, reached a peak at grade 6, and then decreased in grade 7 and grade 8. The changes in the loading averages of content M and N followed a similar pattern. Both increased from grade 3 to grade 4 and then decreased from grade 4 to grade 5 and leveled off in grade 5 through grade 7. The only difference was that the loading average of content M increased from grade 7 to grade 8, while that of content N did not. Detailed information about the individual item loadings within each grade level are provided in Table 6.

Insert Table 5 about here

Insert Figure 4 about here

Insert Table 6 about here

Conclusion and Future Work

Vertical scales are usually constructed through the creation of linking items, which are intended to be equally representative of the constructs measured at adjacent grades. This is hard to achieve in practice, since the constructs assessed at different grade levels are, more often than not, different. In this paper, the assumption of construct invariance is tested by making use of the relationship between the linking items and the test constructs. If the same relationship exists between the linking items and the test constructs for different grades, we have some evidence to say that the test constructs are invariant across grades. Otherwise, the assumption of construct invariance does not hold.

In this study, the multi-group factor analysis technique is applied to test the construct homogeneity across grade levels with the common item design. Results indicated that construct shift existed in the tests across grades. The changes in the loadings of the items indicated the content domains in which differences of constructs might be attributed to and the nature of the differences across grade levels.

In this study, we assessed the equivalence of test construct between two adjacent grade levels. We made an assumption that the test construct measured in each grade was a composite of the same content strands. How the composite was formed at each grade level was not studied. Specifically, we were not interested in investigating the internal structure of the test construct measured in each grade. Nor did we care about the number of dimensions that made up the composite construct. Further studies can be conducted to explore the structural differences between the test constructs across grades.

References

- Bentler, P. M. (2005). *EQS Structural Equations Program Manual*. Multivariate Software, Los Angeles.
- Haertel, E. H. (2004). *The behavior of linking items in test equating* (CSE Technical Report 630). Los Angeles, CA: CRESST/CSE, University of California, Los Angeles, Graduate School of Education and Information Studies.
- Huynh, H., & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education, 18*, 99-113.
- Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: methods and practices (2nd)*. Springer Science & Business Media, Inc.
- Martineau, J. A., & Plank, D. N. (2005). *Promise, betrayal, and redemption: Using vertical scales in growth-based value-added assessment*. Unpublished manuscript.

Figures and Tables

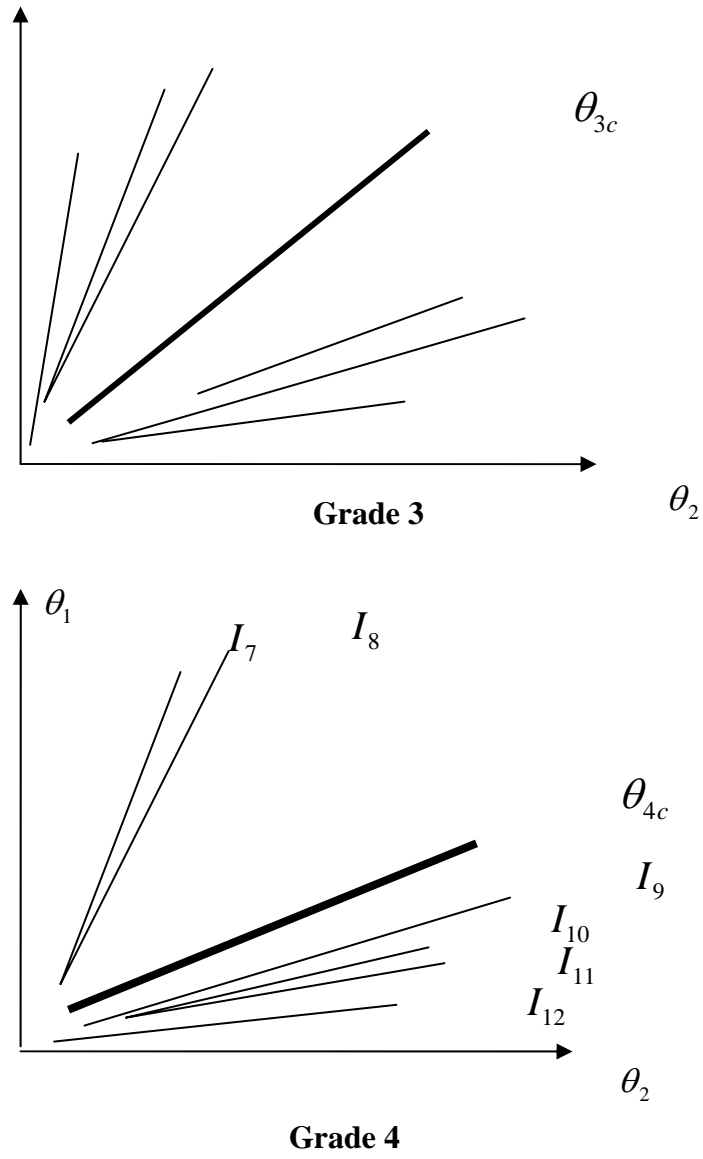


Figure 1: Tests of two grades that cover the same content but have different test constructs

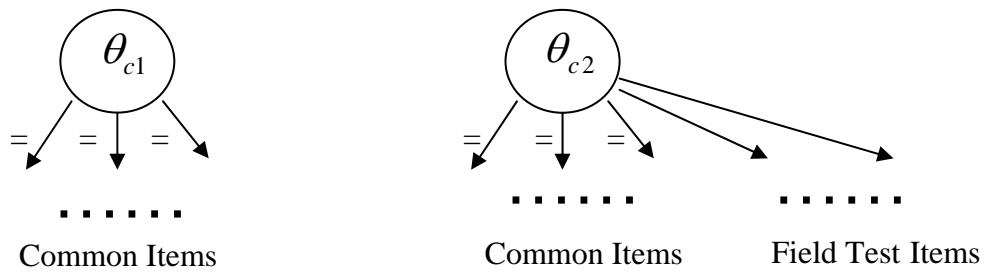


Figure 2: Test of construct invariance with and without field test items within each grade level

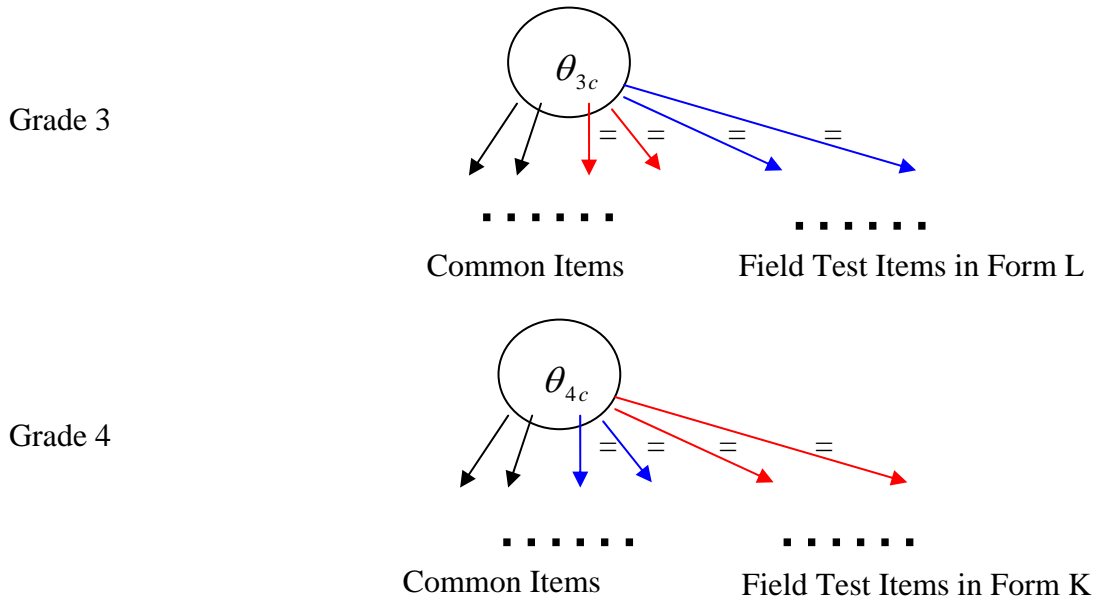


Figure 3: Test of construct invariance between adjacent grade levels

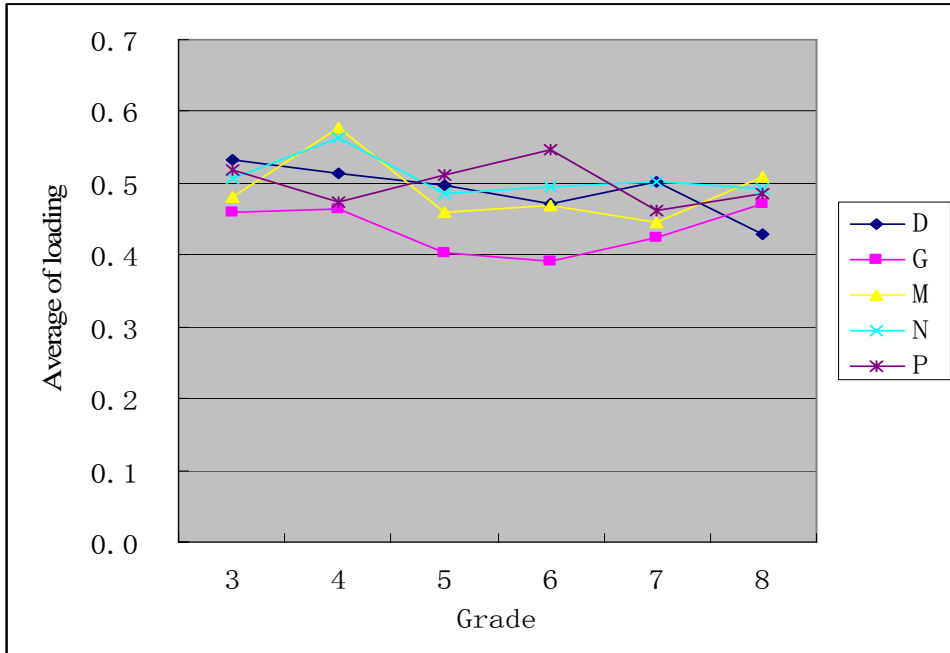


Figure 4: The average of loadings of the on-grade items for each content strand at each grade level

Equivalent Test Structure across Grades 18

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Total	27,141	26,867	27,386	27,573	29,285	29,288
Form K	2,204	2,232	2,280	2,287	2,433	2,428
Form L	2,263	2,239	2,271	2,285	2,419	2,450

Table 1: Number of Students tested in each grade

Content	Common Items		Field Test Items	
	MC	CR	MC	CR
Data analysis, Statistics, and Probability (D)	8	1	2	or 1
Geometry and Spatial Sense (G)	8	1	2	or 1
Measurement (M)	8	1	2	or 1
Number Sense, Properties, and Operations (N)	8	1	2	or 1
Patterns, Algebra, and Functions (P)	8	1	2	or 1
Total	40	5	7	

Table 2: Distribution of items in each grade

Equivalent Test Structure across Grades 20

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Form K		36.638	2.544	1.158	1.303	3.336
Form L	1.542	1.426	0.511	1.248	0.757	

Table 3: Tests of χ^2 difference in Step 1

	3-4	4-5	5-6	6-7	7-8
χ^2 Difference	624.34	134.309	203.594	151.07	132.682
P	.000	.000	.000	.000	.000

Table 4: Tests of χ^2 difference in Step 2

Strand	Grade3	Grade4	Grade5	Grade6	Grade7	Grade8
Data analysis, Statistics, and Probability (D)	0.532	0.514	0.496	0.472	0.503	0.429
Geometry and Spatial Sense (G)	0.460	0.464	0.403	0.390	0.425	0.472
Measurement (M)	0.482	0.578	0.460	0.469	0.445	0.509
Number Sense, Properties, and Operations (N)	0.506	0.564	0.486	0.494	0.502	0.493
Patterns, Algebra, and Functions (P)	0.519	0.473	0.511	0.546	0.462	0.486

Table 5: The average of loadings of the on-grade items for each content strand at each grade level

Strand	G3	G4	G5	G6	G7	G8
D	0.304	0.283	0.33	0.28	0.258	0.3
	0.482	0.496	0.342	0.37	0.344	0.304
	0.486	0.504	0.463	0.375	0.473	0.336
	0.512	0.517	0.503	0.469	0.505	0.38
	0.538	0.519	0.513	0.472	0.549	0.416
	0.582	0.534	0.521	0.512	0.563	0.452
	0.607	0.547	0.54	0.58	0.566	0.454
	0.618	0.611	0.583	0.587	0.629	0.564
	0.657	0.619	0.671	0.599	0.636	0.656
G	0.304	0.333	0.157	0.042	0.189	0.318
	0.372	0.352	0.295	0.196	0.337	0.344
	0.405	0.437	0.318	0.306	0.344	0.376
	0.436	0.472	0.368	0.428	0.349	0.407
	0.44	0.476	0.448	0.428	0.352	0.457
	0.525	0.478	0.468	0.487	0.427	0.475
	0.542	0.485	0.476	0.52	0.489	0.536
	0.552	0.547	0.506	0.553	0.55	0.579
	0.568	0.599	0.59	0.553	0.785	0.758
M	0.314	0.365	0.216	0.291	0.288	0.219
	0.315	0.371	0.294	0.381	0.291	0.402
	0.333	0.403	0.359	0.404	0.359	0.461
	0.394	0.618	0.362	0.419	0.422	0.484
	0.47	0.644	0.429	0.446	0.448	0.519
	0.501	0.657	0.524	0.505	0.45	0.572
	0.561	0.676	0.588	0.522	0.494	0.573
	0.712	0.729	0.592	0.564	0.533	0.579
	0.735	0.738	0.773	0.69	0.718	0.768
N	0.279	0.419	0.23	0.405	0.335	0.331
	0.321	0.45	0.385	0.415	0.378	0.423
	0.392	0.503	0.456	0.429	0.399	0.434
	0.47	0.516	0.495	0.464	0.412	0.459
	0.501	0.556	0.495	0.466	0.48	0.489
	0.558	0.607	0.526	0.513	0.501	0.501
	0.616	0.637	0.526	0.522	0.631	0.529
	0.666	0.68	0.529	0.557	0.652	0.587
	0.749	0.708	0.732	0.676	0.732	0.688
P	0.303	0.338	0.37	0.401	0.328	0.331
	0.412	0.343	0.458	0.447	0.36	0.356
	0.467	0.385	0.465	0.471	0.383	0.383

Equivalent Test Structure across Grades 24

	0.487	0.492	0.481	0.475	0.417	0.457
	0.501	0.512	0.492	0.481	0.46	0.468
	0.549	0.52	0.54	0.622	0.463	0.475
	0.632	0.541	0.55	0.639	0.465	0.561
	0.637	0.547	0.587	0.686	0.512	0.645
	0.687	0.581	0.654	0.691	0.766	0.699

Table 6: The loadings of the on-grade items for each content strand at each grade level