FLORIDA COMPREHENSIVE ASSESSMENT TEST

Technical Report on Vertical Scaling for Reading and Mathematics

R. Gene Hoffman, Lauress L. Wise, and Arthur A. Thacker
Human Resources Research Organization
(HumRRO)
Alexandria, Virginia

Under subcontract to

Harcourt Educational Measurement

San Antonio, TX

Harcourt
Educational Measurement
A Harcourt Assessment Company

San Antonio, TX

October, 2001

TABLE OF CONTENTS

List of Tables

List of Figures

INTRODUCTION AND OVERVIEW

This report presents technical information on the construction of measurement scales – for the Florida Comprehensive Assessment Test's (FCAT) Reading and Mathematics assessments – designed to place all Grade 3 through 10 results on a common scale. Currently, operational FCAT reporting scales are completely independent of each other so that assessing gains in achievement across grades is not possible. While the existing operational scales will remain in use, the new vertical scales will provide a basis for comparing students' achievements across grade levels. This report includes an overview of the purpose and limitations of the vertical scales and the steps used to construct the vertical scales, including checks on the process, and proposed final scales.

The report is technical in nature; however, an attempt has been made to make it accessible to an audience with working familiarity with testing concepts and a willingness to tackle some algebra. The intent is to provide an intuitive understanding of vertical scaling, while still including enough detail for psychometricians to recognize what we have done. Technical details are inserted in footnotes as appropriate. Also note that in addition to describing results for different grade levels, some of the following discussion refers to testing in different school years.

Before proceeding with a description of the FCAT Vertical Scale, a short introduction to the concept is presented. A more complete review of vertical scaling, written in preparation for this effort, is also available (McBride & Wise, 2000).

Current FCAT assessments and reporting

The FCAT currently includes operational assessments for Grades 3 through 10 in Reading and Mathematics. Each of the separate grade/subject tests (e.g., Grade 4 Mathematics, Grade 8 Reading) in the FCAT (http://fcat.fldoe.org/) is designed to concentrate on content and skills defined for each grade by the Florida Sunshine State Standards (http://www.fldoe.org/bii/curriculum/sss/). Analyses are conducted each year (see *FCAT Technical Report*, 2000, 2001) that allow students to receive scale scores and performance level designations separately for each grade/subject test. The reporting scale for each grade and subject runs from a low score of 100 to a high score of 500. Scores on these 100 to 500 scales are divided into five Achievement Levels according to performance standards set by Florida educators for each grade and subject.

In addition to providing all students with individual feedback, students' scores are a major component of the School Accountability Report (http://schoolgrades.fldoe.org/reports/index.asp). This report judges the quality of education in elementary, middle, and high schools throughout Florida. Goals are set for schools in terms of both the level of student achievement and in terms of a reduction from one year to the next in the percentage of students scoring in the lowest Achievement Levels. The general intent is for schools throughout Florida to improve their educational processes and raise the FCAT scores of their students.

The Florida Department of Education has been reporting FCAT scores to students and schools annually for Reading in Grades 4, 8, and 10 and for Mathematics in Grades 5, 8, and 10 since 1998. Throughout the remainder of this report, we will simply refer to these six grade and subject combinations as the “old” grades. Reporting FCAT scores for Reading and Mathematics in the remaining Grades 3 though 9 began in 2001. We will refer to these as the “new” grades for each subject.

What is a vertical scale and why do we need one?

Missing from the current reporting system is a direct estimate of the year-to-year growth for individual students. Certainly, a student’s relative standing can be monitored with the current data, that is, whether a students has maintained Level 2 or 3, etc. from one year to the next. On the other hand, there is no way to decipher the amount of achievement that students are gaining from one year to the next. A vertical linking of the grade-specific, operational scales is needed to create a means for more directly assessing achievement growth for individual students. Vertical linking provides the means for translating operational, grade-level test scores to a common measurement scale. For each subject (Reading and Mathematics), a vertical scale would provide separate equations that would translate each grade and subjects’ operational 100-to-500 scale scores to a common measurement scale. With operational test scores translated to a common scale, comparisons in relative achievement can be made across grades.

Limitations on the interpretation of vertical scales

There are two important caveats about vertical scaling that must be recognized, both of which stem from the fact that each FCAT is constructed with items that are most appropriate to content standards for the grade level being assessed. For example, results on the fourth-grade test indicate achievement on fourth-grade skills and content, while fifth-grade FCAT scores indicate the achievement of fifth-grade skills and content. However, it is likely that some fifth grade content may be learned in the fourth grade, but not tested until the fifth grade. The fifth grade test may show that the content has been learned, but would not necessarily be a correct index of when the content was learned. In other words, to measure achievement gains precisely, there needs to be a pre-test (given before a grade) which tells what students know and do not know about that grade’s content. At the end of the grade, a post-test needs to be administered that covers the same material. Vertical scaling depends on there being enough similarity between grades in the skills and content taught so that performance on a lower-grade test can serve as a pre-test for the next higher grade, providing a reasonable estimate of the starting point for students’ achievements in the next higher-grade skills and contents. Of course, each test must also serve as the post-test for its own grade. A single test can serve as both a post-test for its grade and a pre-test for the next grade if there is sufficient overlap in content and skills between grades. Only then can the difference between a lower grade test score and higher-grade test score be interpreted as an estimate of learning during the higher grade. Dampening our concern, however, is the fact that – at least in the case of mathematics – items are actually shared across grades during normal operational testing.

The distribution of scores from any one grade expressed on a vertical scale may overlap scores on the vertical scale from several adjacent grades. It becomes tempting to make projections that high scoring students in a low grade know the content achieved by lower scoring students in grades two or three levels higher. Because the skills and content shift between grades, making inferences about knowledge of specific content across multiple grades is increasingly inexact. While any overlap in scores across grades is instructive in a general sense, projections about specific content knowledge across grades should not be over interpreted

The FCAT is intended to stimulate the academic capacity of Florida's education system, thereby improving students' year-to-year achievement gain. As a result, achievement gains are likely to be dynamic, changing from year to year. This dynamic creates a second caution that is important to recognize before viewing the results that follow. The FCAT vertical scales reported here were built by a special comparison of FCAT performance across grades during Spring 2001. As a result, the scales were based on the cross-sectional (cohort-to-cohort) differences that existed between grades in Spring 2001. As the Florida education system changes and student learning is accelerated in future years, gains in achievement between grades may change as students in each grade learn increasingly more content and skills. Such changes may alter our expectations for growth, but would not necessarily change the common measurement scale. If, however, instruction and even content standards were modified to give different emphasis to topics or to emphasize them in different grades, the relative difficulty of test questions tied to these topics may change and the nature of the common scale would be altered. Consequently, vertical scaling should be periodically checked to ensure that differences between grades are not being distorted by changes over time in content standards or in the nature and effectiveness of instruction at each grade.

In a dynamic environment, the vertical scale itself cannot be interpreted as an index of how much students *should* learn. Rather, it can only provide an index of how much students have learned based on a measurement scale that is calibrated to differences between grades captured during Spring 2001. We will add to the discussion of this topic later in the report.

Finally, because the vertical scale will be used to make comparisons from one grade to the next, there will be inherent reliability problems. Such comparison will involve computing differences in scores on the vertical scale. Difference scores are significantly less reliable then the individual scores from which they are computed (Cronbach & Furby, 1970). We will explore this issue later as well.

In spite of these uncertainties, vertical scaling is a common practice among test publishers (McBride & Wise, 2000). It provides a metric for reporting students' achievement growth. However, the drawbacks can be sufficient for us to know that caution will be required in interpreting scores and score gains on the vertical scale.

DATA COLLECTION FOR CONSTRUCTING THE VERTICAL SCALE: SPECIAL TEST FORMS DESIGN

In order to link an achievement scale from one grade to the next, a special data collection scheme was devised which incorporated the use of common items administered to students in more than one grade. These common items were the basis for translating the separate operational tests to a common scale. McBride and Wise (2000) reviewed the variety of options available and identified some specific conditions placed on gathering responses to common items in the context of FCAT administration. These conditions are summarized in Table 1.

Table 1.
Conditions for Vertical Scaling Data Collection Design

1. There would be no special vertical scaling test administration.
2. There would be a separate operational test for each grade/subject.
3. All students would be administered their on-grade operational test.
4. All students would be administered a single operational test form.
5. There would be space provided – normally field-test item locations – for additional, off-grade vertical scaling items in the test form.

Given these conditions, common items were introduced in data collection for vertical scaling by using the field-test item locations for each grade as a place to administer items that were operational in adjacent grades. For example, some of the students in Grade 4 were administered a sample of Grade 3 items and some were administered a sample of Grade 5 items. Operational FCAT testing includes approximately 45-50 items (depending on grade level and subject) and space for an additional five or six field-test items. In order to field test more items, operational FCAT employed 10 forms for each grade/subject combination in 2001 with the only difference being in the field-test items. For the special vertical scaling data collection, five additional forms were constructed for each grade and subject.

Figure 1 depicts the design and the resulting overlap in tested items created by this design. The design depicts operational items being "borrowed" by both the higher adjacent grade and by the lower adjacent grades. These borrowed items were placed in field-test positions that were spread across five forms per grade/subject.

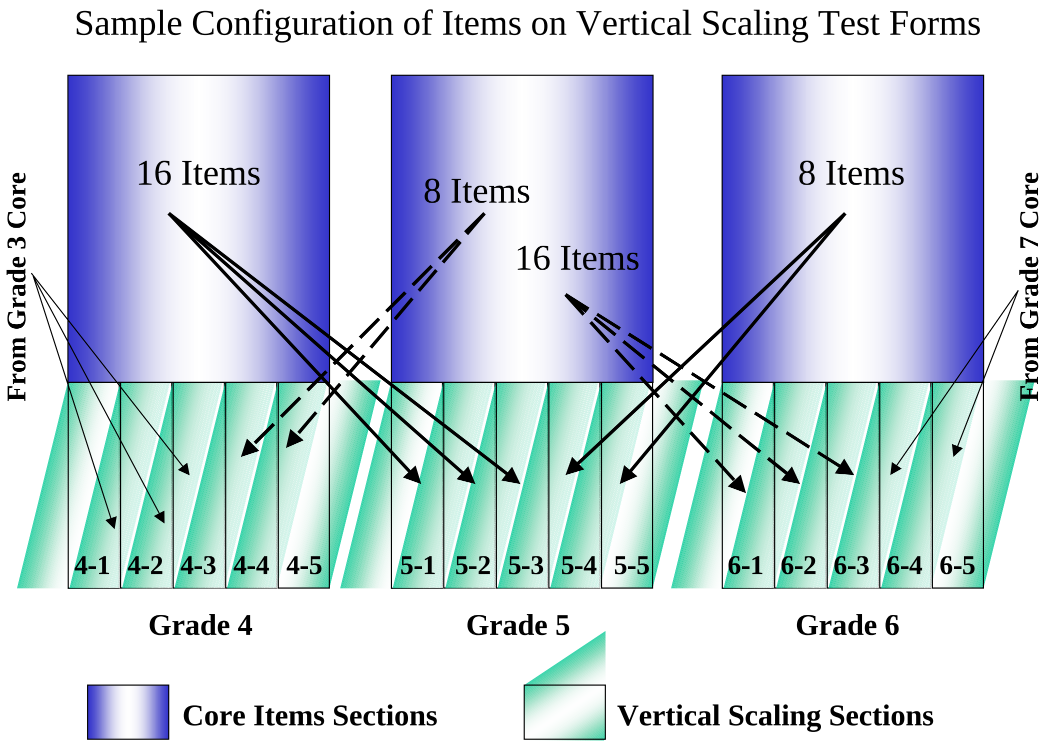

Figure 1. Basic vertical scaling forms design. Exact numbers of items vary slightly by grade and subject.

Three of the five vertical scaling forms contained items borrowed from the next lower grade with the remaining two forms containing items borrowed from the next higher grade. More items were included from the lower grade in order to reduce the possibility that higher-grade items would be so difficult as to preclude a useful number of lower-grade students from passing the items.[1] In addition, relatively easy items from the higher grade were selected. These were items that primarily discriminated among students in the lower half of the distribution of students in the higher grade.[2] Items from the full range of difficulty for the lower grade were included in the special vertical scaling data collection.

[1] For an item to be useful, students' performance must vary. Our typical rule of thumb is that items with less than 20% passing (near chance performance) or more then 90% passing would have had little utility for scaling.

[2] FCAT multiple-choice and gridded-response items are analyzed with 3PL and 2PL Item Response Theory models using Multilog™ (Thissen, 1991). Item difficulty was judged by the "b" parameters with selected higher-grade items having "b" values below the mid-point of the higher grade's operational scale.

Operationally, initial grades use machine-scored multiple-choice and gridded-response (for mathematics) items plus hand-scored open-response performance tasks. On the other hand, the “new” grades use only machine-scored items. Therefore, only machine-scored items could be used as common items across grades. Finally, items were selected from the higher and lower grades to assess the full breath of content in those grades.

ITEM RESPONSE THEORY

Students’ operational FCAT scores are constructed using an Item Response Theory (IRT, Lord & Novick, 1968) model. As further explained in the FCAT Technical Report (Human Resources Research Organization and Harcourt Educational Measurement, 2001), IRT processing defines an achievement scale for each grade/subject assessment based on the overlapping content of the items included in the test. The achievement scale has two independent properties, one related to the attribute being measured, and a second related to the range of numbers used to quantify the amount of the measured attribute.

Items and attributes being measured

Each Reading and Mathematics test focuses on its own set of Sunshine State Standards. These standards describe the content and skills applicable to each grade. Test items are written to target these standards. While similarities clearly exist in content across grades, each grade-level collection of items is unique, and therefore each IRT achievement scale uniquely reflects its target content. In a sense, each scale is the best statistical representation of the corresponding Sunshine State Standards. In measurement terms, the collection of test items defines the achievement “construct” being assessed by the test. Within the general topic of Reading or the general topic of Mathematics, the specific content of test items changes to some degree from grade to grade. As a result, the achievement construct changes as well.

The IRT model uses information on the pattern of responses to all of the test questions. This results in differential emphasis across test questions. Missing an easy question may be more significant than getting a difficult question correct, if the possibility of guessing is taken into account for multiple-choice items. As a result, there is not an exact relationship between the number of questions answered correctly and the achievement construct or scale. Instead, the relationship between an item and its achievement scale is expressed as a non-linear function that relates the probability of correctly responding to the test item to the student’s true level of achievement as measured by the IRT scale. These curves are called “item characteristic curves.” Figure 2 (on page 8) depicts a sample relationship between achievement and the probability of responding correctly to a single item. This particular curve is for a multiple-choice item and shows that for even low-achieving students there is the possibility of getting the item correct by guessing. With increasing levels of achievement, the probability of answering the item correctly improves but not in a straight line. There are a variety of standard non-linear, frequently logistic, equations used in IRT to describe this curve. When achievement reaches a high enough level, the probability of answering the item correctly becomes almost 1.

Specific curves such as in Figure 2 are created by estimating so-called “item parameters.” These are parameters in the probability function that allow the curves to bend and slide back and forth in order to model actual student response data. These parameters are something like slope and intercept coefficients for simply linear equations that determine the placement of a straight line on a graph. Because IRT uses particular types of curves, item characteristic curves for multiple-choice and gridded-response items are similar to Figure 2. However, because items on a test will relate differently to the achievement scale, these response probability functions will vary with where they start (the lower asymptote) and where and how steep the upward slope is. We will have more to say about item parameters later.

There are two ways of interpreting the horizontal scale (Y-axis) of the item characteristic curve in Figure 2. As labeled, the curve shows the probability of getting an item correct, given some level of achievement. For example, from the figure it appears that persons with an Achievement Level of 1 on this scale have about an 80 percent chance of getting the item correct. The alternative way to think about the scale is as the expected number of points persons with a given ability would receive for this item, on average, if they got 1 point for a correct answer and 0 points for an incorrect answer. From this perspective, all students at Achievement Level 1 would, on average, get 0.8 points for the item. With this alternative perspective, we can imagine using similar curves for all items in a test, and, for any given Achievement Level, adding up the expected points for students at each Achievement Level. The sum is the average number of points expected on the whole test for persons with the given Achievement Level. If the expected total points for each Achievement Level are plotted, the resulting curve is called the “test characteristic curve.” Test characteristic curves will have an S-shape similar to the example item characteristic curve; however, the horizontal axis will describe total test points.

Numeric scale

We indicated earlier that IRT modeling creates an achievement scale with two characteristics: a construct that it represents and a numeric expression of that construct. In Figure 1, the achievement scale is depicted on a numeric scale from –3 to +3, centered on 0. Traditionally, IRT analyses produce similar scales which fix the distribution of achievement to a standardized metric with a true score mean of 0 and a true score standard deviation of 1. Often, this metric is inconvenient, particularly the negative numbers, and so the numeral aspect of the scale is transformed by a linear adjustment. The construct captured by the IRT scale remains, but the number scale is changed.

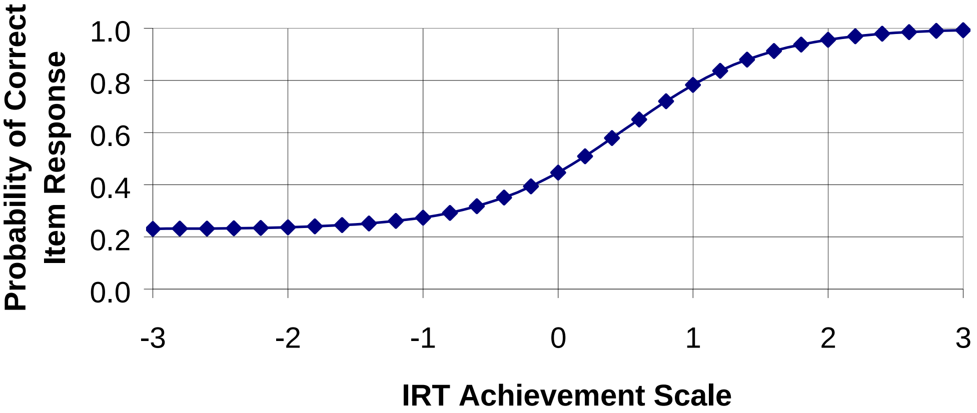

Figure 2. Example of an item characteristic curve.

Linking scales across years or grades

For any given grade and subject, the item and test characteristic curves produced by IRT are initially created using a numeric scale that places average achievement at 0 with an expected score standard deviation of 1 for the sample of students used to estimate the item parameters. From year to year or from grade to grade, however, different samples of students are assessed. Achievement means and standard deviations for students in the new year or grade will be different from achievement means and standard deviations in the base year or grade. We need to adjust the initial scale for the new year or grade to reflect these achievement differences appropriately. This is accomplished using a common set of items administered to both the base and new samples. If achievement is higher for the new (e.g., higher grade) sample, the probability that an average student (at initial score level 0) will pass common items will be greater than for the base sample. In fact, the whole item characteristic curve will be shifted to the left in the new sample. In addition, the scale for the new sample may be expanded or compressed as a function of differences in standard deviations. Stocking and Lord (1983) provide an algorithm for identifying a linear adjustment to the new scale, of the form

$$Y = a X + b \qquad (1)$$

where X is a score on the initial scale for the new group and Y is the corresponding score on the adjusted scale and “a” and “b” are “slope” and “intercept” parameters respectively. The algorithm finds slope and intercept parameters for the linear adjustment that make the test characteristic curve of the new scale numerically as similar as possible to the base scale.

Vertical scaling uses this approach to analyze the parameters of items repeated in adjacent grades in order to capture the otherwise hidden differences in achievement between grades. With mathematical links, such as this between all pairs of adjacent grades, it becomes a simple algebraic task to string together relationships among all grades.

PROCESSING DETAILS AND RESULTS

The following section is technical by nature and includes some details that are necessary for the technical audience.

Item IRT parameters

The first step in vertical scaling was acquisition of item parameters. Recall that every item used in linking was an operational item for one grade and a “borrowed” item for an adjacent grade. Parameters for the operational use of the items were available from operational FCAT scaling, except for one detail created by the Spring 2001 reporting schedule. Scaling for the borrowed use of the items required separate analyses.

Operational item parameters

Parameters for the operational usage of the linking item were estimated by IRT processing which produced a true score scale centered at 0 with a standard deviation of 1.[3]

FCAT 2001 was administered in early March, and student score reports had to be completed by early May. With this compressed timeline, achievement scores for initial grades were computed without using the hand-scored performance tasks. Performance tasks were subsequently scored and became available for analysis in June. For all initial grades/subjects, IRT parameter estimation was repeated with the performance tasks included. Grade 10 Reading and Mathematics scores were reissued because these assessments were high school exit requirements. Although scores were not reissued for the remaining initial grades, the re-analyses were

[3] Operationally, these standardized IRT parameters are converted to the FCAT 100-to-500 scale for use in computing students’ scale scores. For the initial grades, converting items from the initial 2001 IRT standardized scale to the 100-500 scales was conducted by a Stocking/Lord process which simultaneously converted them to the 100-500 scale and adjusted them to be equivalent to the 1998 scale. The new grades were first centered on 100-500 in 2001, so only a simple mathematical transformation was required for conversion of the IRT standardized values. If we were to have used the operational 100-to-500 parameters for vertical linking, we would have been confounding grade-to-grade differences with within-grade, year-to-year differences. This was the first of two reasons that contributed to a decision to conduct vertical scaling using unadjusted (1/0) item parameters. The second reason is presented later.

conducted since future operational FCAT scores for the initial grades are expected to include the performance tasks. Therefore, "operational" parameters for the initial grades actually refer to the reanalyzed parameters that included performance tasks.

Since the tests for the new grades did not include performance tasks, it was not necessary to recompute operational parameters.

Borrowed item parameters

Parameters for items "borrowed" from a lower or higher grade were computed separately for each grade/subject. The process we used created parameters for these items that were on the operational scale for the grade that borrowed the items.[4]

Student samples

Given the timing of the data analyses, sample sizes for computing IRT item parameters varied between new grades and initial grades and between operational and borrowed items. In all cases, however, only "standard curriculum" students who received operational scores were used in the analyses. This excluded special needs students, home-school students, exempt English language learners, students who

[4] McBride and Wise (2000) outlined several alternative approaches, and this was the one which fit the needs of the FCAT vertical scales. For example, we rejected using an anchor test that would have included items from the full range of mathematics or reading content from Grade 3 to Grade 10 for two reasons. One reason was that we thought that the content domain would have been too broad to scale with a single IRT analysis. The second reason was that this design was not feasible for FCAT administration. Within the administration design that was adopted, we could have conducted special IRT analysis for each grade that included both operational and borrowed items, simultaneously estimating borrowed item parameters and re-estimating operational item parameters. This procedure would have created a hybrid scale (construct) with a broader content and skills base. The resulting hybrid achievement scales would have allowed linking adjacent grades with achievement scales more closely aligned on content. On the other hand, this process would also have created an uncertainty about the relationship between the operational scales and the hybrid scales. Since students will receive scores on both the existing operational scales and on the new vertical scale, using a hybrid scale as an intermediary step would have created an uncertainty about the relationship between the operational scales and the vertical scale. Therefore, parameters for borrowed items were estimated by "fixing" the ability scale.

As a result of operational processing, operational achievement scores had already been computed for the students who were included in the scaling of borrowed items. These achievement scores provide the means for "fixing" the ability scale. The operational scale scores, however, were computed on the FCAT 100-to-500 scale. In order to analyze the off-grade item parameters, Multilog™ (Thissen, 1991), the IRT software we used, required the fixed scale to be centered on zero. This was the second reason for maintaining vertical scaling analyses on the unadjusted parameters. Therefore, students' operational scale scores were standardized prior to the IRT analysis. In this way, item parameters for the off-grade items were also centered at 0 with a standard deviation of 1 on the operational achievement scale for the grades in which they were used.

failed to meet an “attemptedness” criteria, and others with processing identification problems.

Chronologically, parameters for the new-grade operational items were computed first. These operational parameters were based on a special “early return” sample (described in the FCAT technical report) and included approximately 4,000 students per grade/subject.

For the initial grades, the repeated operational scaling occurred after all students’ performance tasks had been scored. To keep the analyses manageable, students were selected by first sorting all students by school district, ethnicity, and gender, and then selecting every fourth student. Standard curriculum students with matched data for the machine-scored and performance task sections of the test were then identified. This resulted in sample sizes for these grades between 32,000 and almost 40,000 students.

For both initial and new grades, the parameters for the borrowed use of the items were computed after item responses for all students’ responses became available. Borrowed items were in the field-test positions and were spread across five of the fifteen forms. This led to approximately 1/15 of the standard curriculum students per grade, or approximately 10,000 students being used to estimate the parameters for each of these items.

Table 2, on the following page, presents sample sizes for each of the IRT analyses.

Parameter files

The next step in the analysis was essentially a clerical sorting and ordering of the operational and borrowed item parameters into matching files that could be used by the Stocking/Lord routine to link adjacent grades. For each grade, two files were created: one to link with the next higher grade and one to link with the next lower grade. The specific numbers of items used in each linking can be found in Table 4 that appears in a later section of the report (page 18).

Initial Stocking/Lord linking

The product of Stocking/Lord linking between adjacent grades was the slope and intercept constants (labeled “a” and “b” in Equation 1) which can be used to transform parameters from one grade onto the achievement scale of the adjacent grade. We chose to place each higher grade on the scale of the lower grade.[5]

[5] As an additional check on our Stocking/Lord routine, we also linked each adjacent grade to the higher grades. With the reverse linking, we would expect to obtain 1/a as the slope and –b and the intercept.

Table 2.
Sample Sizes for IRT Item Parameter Estimation

	Operational Items	Forms with Lower Grade Items			Forms with Higher Grade Items	
		1	2	3	1	2
Reading						
Grade 3 *(n)*	4631	n/a	n/a	n/a	10845	10658
Grade 4 *(i)*	39677	10701	10715	10644	10725	10263
Grade 5 *(n)*	4290	9948	9985	10034	10020	9960
Grade 6 *(n)*	5094	10404	10350	10384	10318	10240
Grade 7 *(n)*	5247	10164	10168	10211	10165	10114
Grade 8 *(i)*	36798	9754	9655	9710	9651	9554
Grade 9 *(n)*	5467	10931	10861	10874	10919	10773
Grade 10 *(i)*	32148	8418	8399	8333	n/a	n/a
Mathematics						
Grade 3 *(n)*	4623	n/a	n/a	n/a	10825	10671
Grade 4 *(n)*	4631	10255	10277	10293	10357	10244
Grade 5 *(i)*	38644	10470	10422	10418	10492	10242
Grade 6 *(n)*	5086	10393	10332	10381	10318	10211
Grade 7 *(n)*	5265	10177	10171	10185	10148	10108
Grade 8 *(i)*	35811	9751	9669	9696	9672	9563
Grade 9 *(n)*	5439	10937	10831	10826	10867	10751
Grade 10 *(i)*	31655	8369	8339	8285	n/a	n/a

Note: Letters in parentheses indicate new *(n)* and initial *(i)* grade/subjects.

Table 3 presents the initial Stocking/Lord adjustment constants. Immediately noticeable were two intercept values less than zero: the Grade 5 and 6 link for Mathematics and the Grade 8 and 9 link for Reading. This means that the average performance on the IRT achievement scale for the higher-grade students was lower than average performance on the achievement scale for the lower-grade students. On average, the higher-grade students did not perform as well as the lower-grade students on the linking items. In each case, the higher grade was a new grade for which operational testing began in 2001 and the lower grade was an initial grade for which testing has been on-going since 1998. Three other grade pairs showed low intercept values: Grades 4 and 5 for Reading and Grades 8 and 9 for Mathematics followed the

The resulting differences from this additional check were trivial, well within the values that might be expected due to computational rounding.

same pattern of the higher grade being a new grade and the lower grade an initial grade. The exception to this pattern was the low intercept for the Reading Grade 5 and 6 link, both of which were new grades.

Table 3.
Initial Stocking/Lord Results

Link	Reading		Mathematics	
	a (Slope)	b (Intercept)	a (Slope)	b (Intercept)
4 on 3	0.922	0.714	0.957	0.640
5 on 4	1.030	0.151	0.970	0.811
6 on 5	1.021	0.117	1.044	-0.104
7 on 6	0.972	0.210	0.957	0.605
8 on 7	0.865	0.456	0.861	0.594
9 on 8	1.271	-0.076	1.016	0.127
10 on 9	0.882	0.492	0.942	0.454

Because of the unexpectedly low and negative intercept values, we conducted some additional investigations of actual proportions of students with correct responses, i.e., item *p*-values.

Item *p*-values

Stocking/Lord results were based on analyses of IRT item parameters. A more straightforward way of examining the unexpectedly low performance by students in some of the higher grades was to check *p*-values for the linking items. We explored *p*-values by separately analyzing items that were operational in the lower grade and items that were operational in the higher grade because these items differ in two ways: (a) their source, operational or borrowed, and (b) their placement in the test forms. The source of the items may have curriculum implications in terms of the timing of content coverage. For example, fifth graders may have performed better on fifth grade content than sixth graders because they learned it more recently. In addition, borrowed items were placed in field-test positions, and in all cases, the field-test positions were at the end of the test form. Item placement can have a noticeable adverse effect on students' responding (Diaz & Wise, 2000) when items are placed at the end of the test.

Table 4 (on page 15) presents mean *p*-values for all items in each link, and separately based on the operational source for the items. Differences in mean *p*-values are also presented.

Mean *p*-values for all items mirrored the Stocking/Lord results. In the same two links as above, Grade 8 to 9 Reading and Grade 5 to 6 Mathematics, the higher-grade students averaged lower on the items than the lower-grade students. The pattern of results, however, was systematically different for items that were operational in the higher grade versus those that were operational in the lower grade. For items operational in the higher grades, mean *p*-values were always higher for the higher-grade students than for the lower-grade students. On the other hand, for items operational in the lower grade, in four cases mean *p*-values were actually lower for students in the higher grade than for students in the lower grade. In all cases, the difference in *p*-values between lower-grade students and higher-grade students was greater for items operational in the higher grade than for items operational in the lower grade.

The pattern is consistent with the expected performance differences based on test form position. The items that were operational in the lower grade were at the end of the higher grade's test form. This could have depressed correct responding for the higher-grade students and could have led to an underestimate of achievement gains from grade to grade. Likewise, items that were operational in the higher grade were at the end of the lower grade test form. This could have depressed correct responding on the higher-grade items for the lower-grade students and exaggerated grade-to-grade changes in achievement.

We also considered a curriculum explanation for the pattern of these mean p-values. The explanation was that there might be more instructional emphasis in the higher grade on the content of the higher-grade operational items than on the content of the lower-grade items. As a result, achievement would have improved more on the higher-grade operational items than achievement from remedial instruction on the lower-grade operational items. A content review of the items by the staff at the Florida Department of Education (FDOE) could not substantiate this alternative explanation.

We also created scatterplots of *p*-values from the lower grade against *p*-values from the higher grade in order to look for other problems for linking adjacent grades. Generally, the scatterplots defined a thin oval such that the items more difficult for the lower grade students were also the more difficult for the higher-grade students as well. However, these plots also revealed the above differences in means. Within the oval defined by the scatter of all items, the higher-grade operational items and the lower-grade operational items defined thinner, parallel ovals. An example, using the Grades 7 and 8 link for Reading appears as Figure 3. Based on the pattern of all items, higher-grade items were harder than expected in the lower grade, and lower-grade items were harder than expected in the higher grade. The pattern was consistent with fatigue effects for items placed at the ends of the tests. Given that many of the lower-grade items were actually harder than the higher-grade items for students in both the higher- and lower-grade, the item placement explanation seemed more likely than the curriculum emphasis explanation for the overall pattern of *p*-values. We will revisit this later.

Table 4.
Mean *p*-values for Vertical Scaling Items

	All items				Items operational in the higher grade				Items operational in the lower grade				
Link	No of Items	Mean P-value in Lower Grade	Mean P-value in Higher Grade	Mean Difference in P-values	No of Items	Mean P-value in Lower Grade[1]	Mean P-value in Higher Grade[2]	Mean Difference in P-values	No of Items	Mean P-value in Lower Grade[2]	Mean P-value in Higher Grade[1]	Mean Difference in P-values	Difference in Mean Differences
Reading													
3 to 4	27	0.62	0.76	0.15	12	0.61	0.79	0.18	15	0.62	0.74	0.12	0.09
4 to 5	28	0.67	0.69	0.02	10	0.68	0.78	0.10	18	0.66	0.64	-0.02	0.09
5 to 6	29	0.61	0.63	0.02	11	0.65	0.73	0.08	18	0.59	0.57	-0.02	0.04
6 to 7	29	0.62	0.66	0.04	12	0.63	0.73	0.10	17	0.62	0.62	0.00	0.01
7 to 8	27	0.60	0.68	0.09	12	0.59	0.72	0.13	15	0.60	0.65	0.05	0.07
8 to 9	28	0.73	0.72	-0.01	10	0.73	0.79	0.06	18	0.74	0.68	-0.06	0.05
9 to 10	27	0.61	0.70	0.09	12	0.60	0.75	0.15	15	0.62	0.67	0.05	0.08
Mathematics													
3 to 4	31	0.58	0.70	0.12	13	0.54	0.71	0.17	18	0.61	0.69	0.08	0.06
4 to 5	30	0.55	0.70	0.16	12	0.56	0.78	0.21	18	0.53	0.65	0.12	0.12
5 to 6	29	0.64	0.62	-0.02	11	0.64	0.66	0.01	18	0.63	0.60	-0.03	0.10
6 to 7	30	0.44	0.55	0.11	12	0.50	0.63	0.12	18	0.39	0.50	0.11	0.10
7 to 8	30	0.47	0.60	0.13	12	0.58	0.75	0.17	18	0.40	0.50	0.10	0.08
8 to 9	31	0.52	0.56	0.04	13	0.63	0.70	0.07	18	0.44	0.46	0.02	0.12
9 to 10	31	0.59	0.68	0.09	13	0.66	0.80	0.14	18	0.53	0.59	0.06	0.10

[1] Items placed at the end of the test forms.

[2] Items placed in the operational portions of the test forms.

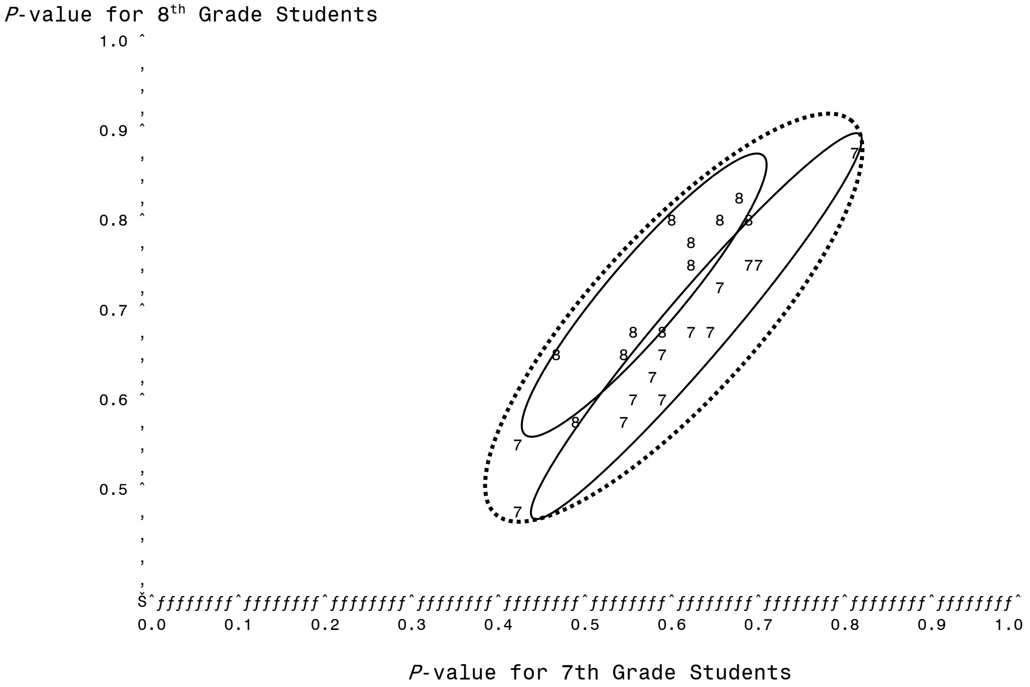

Figure 3. P-values for linking items administered to both 7th and 8th Grade students. Items are identified by the grade for which they were operational.

Stocking/Lord revisited

Armed with these new data, we revised our approach to linking adjacent grades. The vertical linking test form design incorporated more lower-grade items than higher-grade items. With the strong possibility of an artifact affecting item responses, we deemed it necessary to balance the weight of higher-grade and lower-grade operational items. The Stocking/Lord routine starts by summing, across all linking items, estimates of correct item responding with the sum yielding an estimate of an expected total number correct score for those items. Sums were computed at systematically varying achievement levels. The result was a test characteristic curve that describes the relationship between achievement and expected total score for the linking items only. As a result of the straightforward summation, each item contributes equally to the test characteristic curve. The Stocking/Lord routine actually computes a test characteristic curve from item parameters obtained from the lower-grade students and a second test characteristic curve from item parameters obtained from the higher-grade students. The required transformation constants are then derived to reduce the difference (squared) between these curves. The important point here is that normally each item contributes equally to the test characteristic curve, and therefore, equally to the linking solution.

The *p*-value analyses suggested a systematic bias such that high-grade operational items were overestimating grade-to-grade achievement gains, while lower-grade operational items were underestimating grade-to-grade achievement gains. With more lower-grade than higher-grade operational items included in the linking analyses, it became undesirable to have each item contributing equally. To balance any bias, it became more desirable to have the set of higher-grade items make a contribution equal to the set of lower-grade items.[6] To achieve this revised weighting, the Stocking/Lord routine was modified to give extra weight to the higher-grade items in the test characteristic curve summation. That is, the contribution to the test characteristic curve of the higher-grade items was increased by the ratio of the number of lower-grade items to higher-grade items. The weight approximated 1.5 for all links, but varied slightly depending on the exact numbers of higher- and lower-grade operational items.

The new Stocking/Lord linking constants are presented in Table 5. Again, we are working with initial standardized IRT output which fits each grade's achievement scale to a true score mean of 0 and standard deviation of 1 for the corresponding estimation sample.

Table 5.
Revised Stocking/Lord Results – Higher-Grade and Lower-Grade Operational Items Equally Weighted

	Reading		Mathematics	
Link	a (Slope)	b (Intercept)	a (Slope)	b (Intercept)
4 on 3	0.919	0.726	0.946	0.673
5 on 4	1.003	0.261	0.956	0.864
6 on 5	0.991	0.197	1.030	-0.066
7 on 6	0.944	0.254	0.937	0.621
8 on 7	0.859	0.480	0.835	0.624
9 on 8	1.206	0.056	1.007	0.157
10 on 9	0.878	0.525	0.925	0.489

The grade-to-grade gain pattern appeared somewhat stronger, although one link, Grade 5 to 6 Mathematics, still showed a decline on tested achievement.

[6] If there remains any credibility to a curriculum explanation for the differences in performance on the different items, then equal weighting by operational source of the items creates a balance in estimating gains based on new learning on higher-grade operational items versus remediation on lower-grade operational items.

Cumulative linking

The slopes and intercepts in Table 5 were used to create an interim vertical scale that was based on Grade 3, where Grade 3 was centered at 0 with an expected true score standard deviation for standard curriculum students of 1. Straightforward algebra was used. If

$$Y_3 = a_{34}X_4 + b_{34} \text{ and}$$

$$Y_4 = a_{45}X_5 + b_{45}$$

are the linking equations for Grades 3 and 4 and Grades 4 and 5, respectively, then

$$Y_3 = a_{34}\, a_{45}\, Y_5 + (a_{34}\, b_{45} + b_{34})$$

is the link between Grades 5 and 3. The product of two slopes (a_{34} and a_{45}) is the slope that relates the Grade 5 achievement scale to the Grade 3 achievement scale, and the terms in parenthesis are the intercept for this function. By repeatedly applying this algebraic manipulation, functions were created that numerically placed each of the grades onto the Grade 3 achievement scale.[7]

The new compound slopes and intercepts are presented in Table 6. These represented the interim vertical scale. Because we were operating with standardized IRT values and basing the interim scale at Grade 3's true score mean of 0 and standard deviation of 1, the intercepts also represented each grade's standard-curriculum, true-score mean on the interim vertical scale and the slopes represented each grade's standard curriculum true score standard deviation as expressed on the interim vertical scale. These means and standard deviations did not represent actual observed scores. Observed score means for standard curriculum students will tend to be close to the projected true score means; however, observed score standard deviations will be larger due to measurement error. Observed score data will be presented later.

[7] The achievement constructs, of course, are not altered by this transformation. Each grade retains its own construct.

Table 6.
Interim Vertical Scale Defined by Relationships between Operational Scales and Grade 3 (Mean 0/Standard Deviation 1) Base Scale.

	Reading		Mathematics	
Link	b (Intercept) & Mean	a (Slope) & S. D.	b (Intercept) & Mean	a (Slope) & S. D.
3 on 3	0.000	1.000	0.000	1.000
4 on 3	0.726	0.919	0.673	0.946
5 on 3	0.966	0.922	1.490	0.904
6 on 3	1.148	0.914	1.431	0.931
7 on 3	1.380	0.863	2.009	0.872
8 on 3	1.794	0.741	2.553	0.729
9 on 3	1.836	0.894	2.668	0.734
10 on 3	2.306	0.784	3.027	0.679

With these data, there was enough information to plot the relationship in achievement across grades. Figures 4 and 5 present expected true score means and standard deviations for each grade projected onto the Grade 3-based vertical scale. Several notes and observation are required.

With the assistance of linear trend lines through the means, two observations seem obvious. First, grade-to-grade progress is remarkably linear.[8] However, where achievement is higher than expected from the trend line, the history of testing in the initial grades appears to be operating. That is, for Reading, initial Grade 4 is above expectation, with some residual effect in Grade 5. Achievement then shows another jump at initial Grade 8, followed by a slowing of achievement in Grade 9 and another rise in Grade 10. For mathematics, Grade 5 is the lowest initial grade and shows achievement higher than the trend. Then, like reading, there is another spike at initial Grade 8, followed by a slowing of achievement in Grade 9, and another rise in initial Grade 10. Although there may be other explanations, the initial grades certainly have had the FCAT assessments longer than the new grades. The data may be showing the consequences of a longer period of pressure to improve instructional effectiveness in the initial grades.

[8] The linear trend lines may be regarded as about 96% accurate, as defined by the R^2's between trend-predicted means and actual means for both Reading and Mathematics.

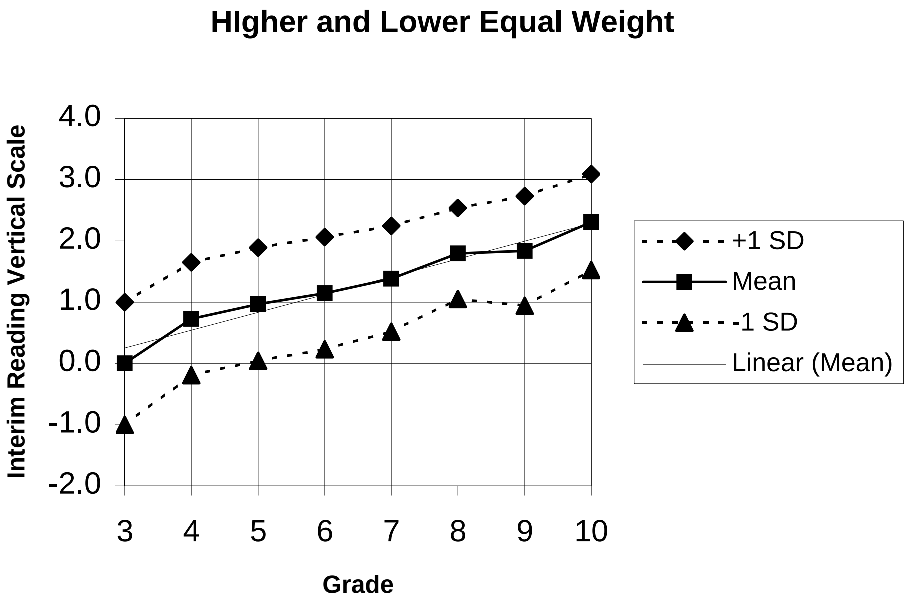

Figure 4. Reading interim vertical scale with true score means and standard deviations by grade.

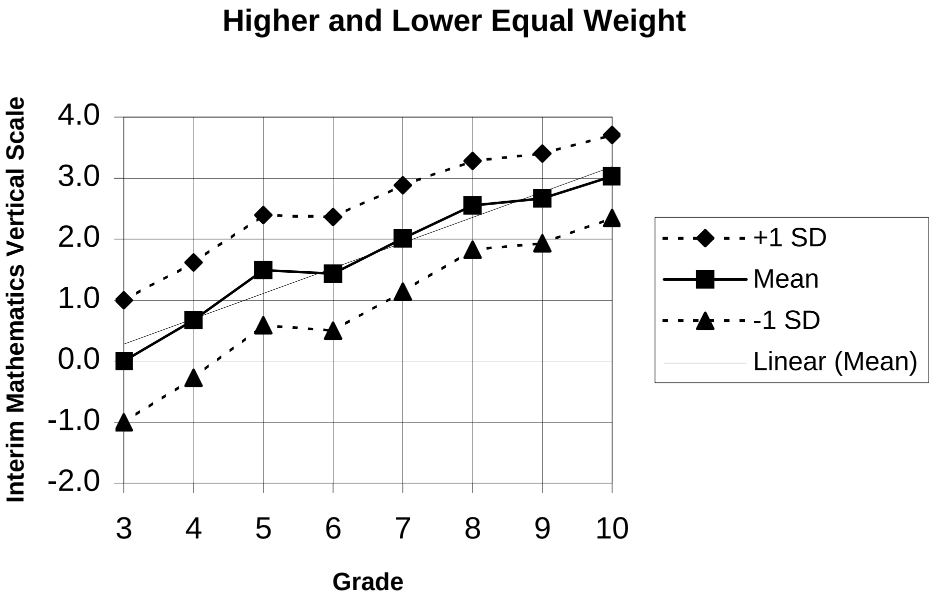

Figure 5. Mathematics interim vertical scale with true score means and standard deviations by grade.

The figures also show gain patterns for higher (+1 Standard Deviation) and lower (-1 Standard Deviation) students. For Reading, within grade (true-score) standard deviation shrinks about 22 percent from Grade 3 to Grade 10. For Mathematics, within grade (true-score) standard deviation shrinks about 32 percent for Grade 3 to Grade 10. As a result, lower achieving students appeared to be gaining slightly more, grade-to-grade, than higher achieving students.

Finally, a comment about the amount of overlap among the grades is in order. The overlap in achievement can be quantified by the relationship between grade-to-grade gains compared to within-grade variation. Based on the data in Table 6, grade-to-grade differences in means for Reading averaged 36 percent of the lower grade's standard deviation. For Mathematics, grade-to-grade gains averaged 48 percent of the lower-grade's standard deviation. Assuming normal distributions, this would mean that approximately 36 percent of the lower-grade students outperformed the average higher-grade students in Reading and approximately 48 percent of the lower-grade students outperformed the average higher-grade students in Mathematics.

For enlightenment, we compared this overlap to the overlap for the SAT 9's national normative sample. We used differences at the 50th percentile to represent grade-to-grade average gains and estimated standard deviations for each grade as one-half of the difference between the 16th and 84th percentiles. As a result, grade-to-grade gains for Reading averaged 38 percent of the lower grade's standard deviation (compared to 36 percent for FCAT) and for Mathematics grade-to grade gains were 39 percent of the lower grade's standard deviations (compared to 48 percent for FCAT). Thus, given the SAT 9 data, which is from a different type of test on a different sample of students, one should not be surprised at the amount of overlap in FCAT achievement distributions across grades.

Supplemental analysis

An important step during Stocking/Lord linking is to examine the relationships from one year to the next or from one grade to the next between the IRT parameters that define the item characteristic curves. Students may become more proficient in successive years or grades so that items become easier. As a result, item characteristic curves may shift. The shift, however, should be about the same for all items. Examining scatterplots of item parameters allows a search for items that deviate from the common pattern. Anomalous items can then be reviewed for content and curriculum differences that may make the items inappropriate for making comparisons in achievement across years or grades.

This type of analysis was conducted for our vertical scaling with a couple of modifications based on our *p*-value analyses and the initial Stocking/Lord results. The *p*-value plots showed a distinction between the items that were operational in the lower grade and the items that were operational in the higher grade. The same sort of distinction occurred for item parameter plots, particularly for the "b" parameters, which may be interpreted as item difficulty. Like the *p*-value analysis, "b" parameters showed that lower-grade operational items tended to be more difficult for higher-grade students than expected based on the difficulties of all items of lower-grade students.

Given this systematic shift in item parameters, we focused our outlier analysis on deviations in item parameters when items were compared to other items from the same operational grade. We essentially ran the Stocking/Lord routine two more times for each grade/subject link: once with items that were operational in the higher grade and once with items that were operational in the lower grade. We also augmented our outlier identification with a statistical computation (described in Appendix A) of each item's deviation from the pattern of grade-to-grade shifts exhibited by all similar (higher-grade operational or lower-grade operational) items. A number of items were detected as potential outliers and flagged for review to determine if there might be curriculum differences or test form cueing that could be causing item parameter differences. We also reran complete Stocking/Lord linkings for all grades/subjects without the most extreme outlier items. When cumulative results (such as Table 5) were obtained, we found that the impact of removing the items was negligible.[9] Content review and test form placement review revealed no obvious reasons for the outlier items to have unusual shifts in parameters. Without context reasons for removing the items, all items were retained in the final solution, as presented above in Table 5 and 6 (on pages 17 and 19, respectively).

In the process of running the outlier analysis, we computed two additional sets of Stocking/Lord linkings: one using only the higher-grade items and one using only the lower-grade items. Plots similar to Figures 4 and 5, constructed separately for higher- and lower-grade operational items, show the different impacts that the two kinds of items had on linking. (See Figures 6 through 9 on the following pages.) As we expected from the previous data, grade-to-grade achievement gains appeared larger when using only items operational in the higher grade versus using items operational in the lower grade. In addition, we also observed a difference in variance. For each successively higher grade, variance appeared to decrease from Grade 3 to Grade 10, by 48 percent for Reading and 52 percent for Mathematics, when the higher-grade operational items were used to track gains. On the other hand, when the lower-grade operational items were used to track gains, variance systematically increased from Grade 3 to Grade 10, by 80 percent for Reading and 53 percent for Mathematics.

[9] The cumulative impact may be seen in the differences in the Grade 10 adjustment constants computed with and without extreme outliers. Differences in the slope parameters for Grade 10 were -0.015 for Reading and -0.016 for Mathematics. Differences in the Grade 10 intercepts were -0.077 for Reading and -0.036 for Mathematics.

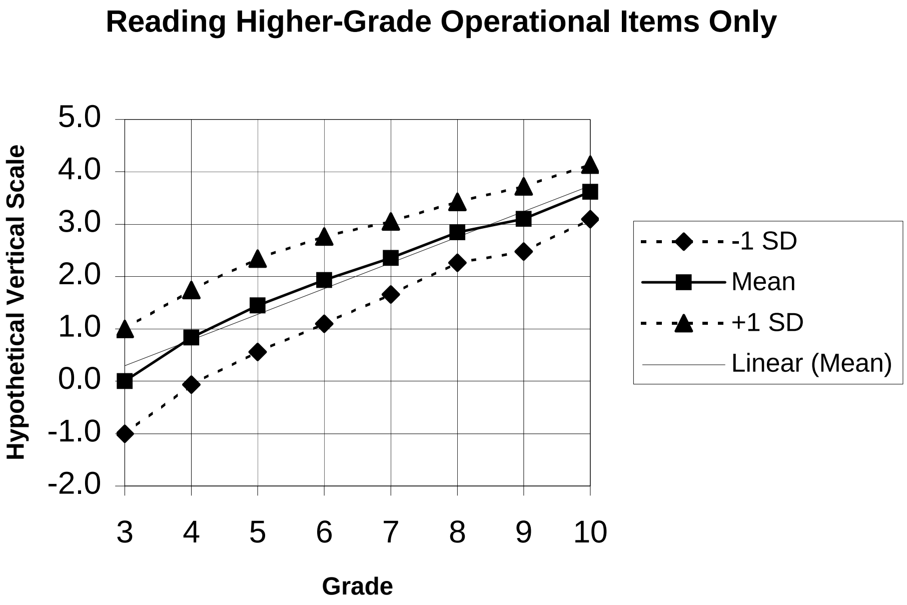

Figure 6. Exploration of vertical linking for Reading using higher-grade operational items only.

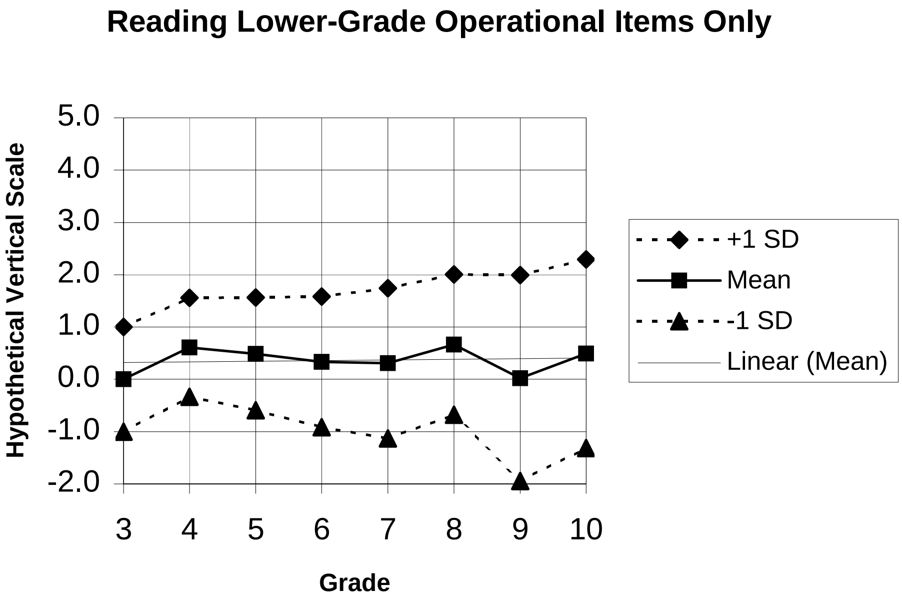

Figure 7. Exploration of vertical linking for Reading using lower-grade operational items only.

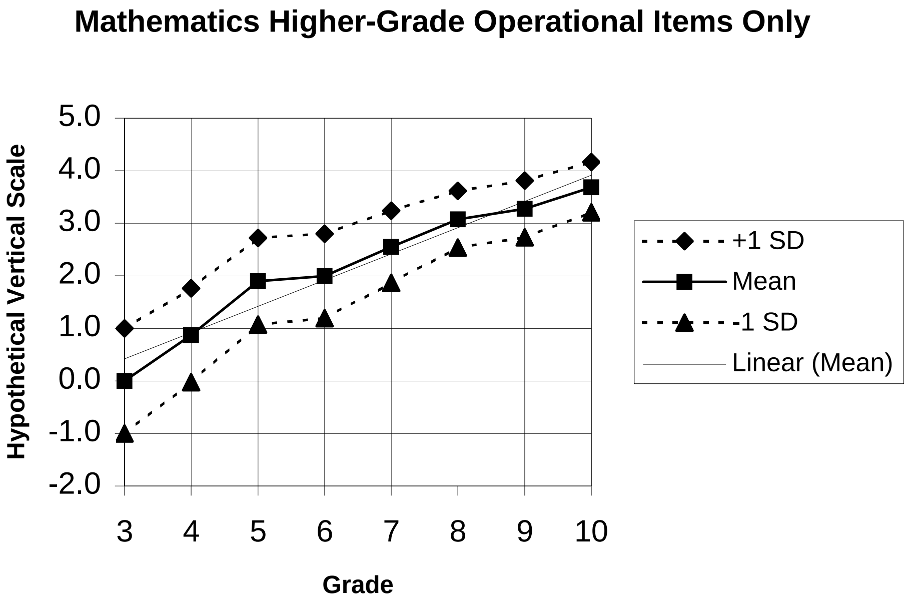

Figure 8. Exploration of vertical linking for Mathematics using higher-grade operational items only.

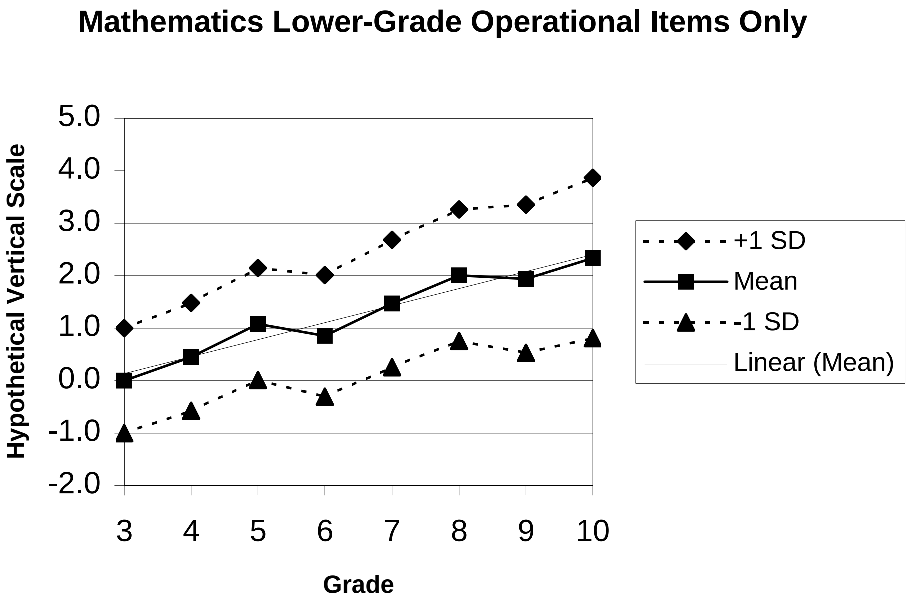

Figure 9. Exploration of vertical linking for Mathematics using lower-grade operational items only.

Post hoc, these differences could be explained as either a consequence of end-of-test fatigue or systemic curriculum effects. This is most easily seen in the linking with the lower-grade items and the difference between low scoring students (-1 standard deviation) and higher scoring students (+1 standard deviation). Test "fatigue" effects typically appeared stronger among less proficiency students than among more highly motivated, highly proficiency students. As a consequence, we expected to see high achieving students appear to be showing more achievement gains than low achieving students. This was the case for Reading and for Mathematics as exhibited by the differences in the lines labeled +1 S.D. and –1 S.D.

A similar pattern appeared in the linking using lower-grade operational items, rather dramatically for Reading. Using these items, higher achieving students appeared to be gaining in Reading proficiency while lower achieving students appeared to be losing ground. Again, this difference might be attributed to motivational differences that tend to depress the test-taking persistence of lower achieving students compared to higher achieving students.

To what extent such hypothesized motivational differences are a transitory phenomenon affecting test performance or part of a broader academic motivational pattern to perform well in general remains an unanswered question. Academic motivation, ability, and test-taking performance are integrally intertwined (see Covington, 2000, for an integrated overview). Therefore, it seems impossible to attribute all of the end-of-test fatigue effects to a temporary state. We can only deduce that the apparent gains are a mix of true gain and test-taking persistence. Our best course of action for vertical linking was to accept that our lowest estimate of gains and our highest estimate of gains bracketed true gains.

Proposed final vertical scales

The remaining step for creating final Reading and Mathematics vertical scales was the translation of the interim scales to final scales. Essentially the goal was to shift and stretch the horizontal, Y-axes of Figures 4 and 5 to create a more "user friendly" range of numbers while maintaining the relative standing among the grades.

An initial preference was expressed by the Florida Department of Education to attempt an adjustment such that mid-points (means) for the grades fell in the pattern 350, 450, 550, etc., up to 1050 for grades 3 through 10. This would also have created a 100-point difference between grades. Two interrelated difficulties immediately surfaced. The first was the amount of overlap between grades and the second was the unevenness in achievement gains. If we attempted to fix grade mid-points as indicated we would have had to stretch and shrink the interim scale differently along different parts of the scale, distorting the metric created by the IRT analyses. We tried relaxing the target mid-points (e.g., letting them vary somewhat from the 350, 450, etc. targets), particularly for the new grades. However, because of the amount of variation within grades, setting the scale this low would have led to a number of the grades having the range of their vertical scale fall below zero. Since avoiding negative numbers was one of the major reasons for not using the interim scale, we began searching for other alternatives.[10]

With the understanding that a constant, i.e., linear, adjustment for the whole range of the scale was needed, it was also clear that essentially all linear adjustments were psychometrically/mathematically acceptable. So, the effort turned to creating a tool to facilitate a search for a linear transformation of the interim values to an alternative that would be meaningful to teachers, students, and parents.

Creating the adjustment became a multi-step process that was implemented in an Excel spreadsheet. The spreadsheet also became an exploratory tool. The goal was to create a figure showing the relationship between each grade's current operational FCAT scores and final vertical scores. On such a figure, meaningful benchmarks could be added to aid interpretation. In order to orient the reader, we will describe the final figures and then explain the process.

Figures 9 and 10 present the proposed vertical scales for Reading and for Mathematics. The figures depict each grade's operational scale as a series of vertical dots representing scale scores from 100 to 500. Dashes indicate the dividing points between the five FCAT Achievement Levels. In addition, we have plotted means and standard deviations on the figures, but with an important difference. Figures 9 and 10 show the actual means of actual scores for all students, in contrast to the expected, true-score means for standard curriculum students that have been presented previously. The +1 and –1 standard deviation scores in Figures 9 and 10 are also based on reported scores for all students, again in contrast to previously presented expected true score standard deviations for standard curriculum. Non-standard curriculum students tend to have lower scores, so the means are lower and the standard deviations higher for all students than standard curriculum only students.

Arriving at these figures began with linking each grade's operational scale to the interim vertical scales. Table 6, in a previous section, shows standard curriculum students' true score means and standard deviations on the interim vertical scale as captured by IRT scaling and linking across grades. Table 7, on the following page, shows standard curriculum students' true score means and standard deviations on each grades' operational scale.[11] The new grades all have true score means of 300 and standard deviations of 50, because 2001 was their base years. The means and standard deviations of the initial grades vary based on changes in students' performance since 1998. Having these two points of information (mean and standard

[10] We may also note that we briefly explored using a non-linear transformation in order to achieve the 350, 450, mid-point targets and at the same time avoid negative numbers. We abandoned further attempts after realizing that avoiding negative numbers at the lower grades required a positively accelerating curve (i.e., a power or exponential function) that greatly exaggerated scores and variance of scores for the upper grades.

[11] These values were the scaling adjustments used to transform IRT standardized parameters to the operational scales for these grades. For the initial grades, these constants, derived by Stocking/Lord equating, placed the 2001 scales on the 1998 base-year scale.

deviation) for each grade, we could compute the linear relationship between each operational scale and the interim vertical scale. The resulting transformation constants are also shown in Table 7.

Table 7.
Operational Scale to Interim Vertical Scale Transformation Based on Operational Means and Standard Deviations

	Operational Data		Transformation	
Grade	Mean	S.D.	Slope	Intercept
Reading				
3	300.000	50.000	0.0200	-6.000
4	309.220	47.636	0.0193	-5.242
5	300.000	50.000	0.0184	-4.566
6	300.000	50.000	0.0183	-4.334
7	300.000	50.000	0.0173	-3.797
8	301.364	47.262	0.0157	-2.931
9	300.000	50.000	0.0179	-3.526
10	313.623	44.016	0.0178	-3.283
Mathematics				
3	300.000	50.000	0.0200	-6.000
4	300.000	50.000	0.0189	-5.000
5	327.945	44.298	0.0204	-5.201
6	300.000	50.000	0.0186	-4.153
7	300.000	50.000	0.0174	-3.226
8	319.135	42.695	0.0171	-2.895
9	300.000	50.000	0.0147	-1.737
10	327.113	38.266	0.0177	-2.776

The second step was to create a transformation between the interim vertical scale and a proposed final vertical scale. In order to do so, we needed to identify two points on the interim scale, match them to two points on the proposed scale, and then compute the transformation equation. This was done by way of the operational scores and the transformation data in Table 7. That is, two points were picked on the operational scales, one on the Grade 3 operational scale and one on the Grade 10 operational scale. Although a variety of options were examined, the recommended points were the respective Grade 3 and Grade 10 means from Table 7. These two points were translated to the interim vertical scale using the Grade 3 and Grade 10 transformation constants in Table 7. Next, two points were picked for the final scale to mate with the two points already identified for the interim scale. Among a variety of possible values that were examined, the values 1300 and 2000 are recommended, with the intent of trying to match values around 1300 with Grade 3 and values around 2000 with Grade 10 and create an average difference of 100 between grades. Table 8 shows the computations.

Table 8.
Computation of Transformation Constants for Proposed Vertical Scale

Grade	Anchor Operational Score	Anchor Values on Interim Vertical Scale	Target Proposed Vertical Scale	Slope	Intercept
Reading					
3	300.0	0.00	1300	303.59	1300
10	313.6	2.31	2000		
Mathematics					
3	300.0	0.00	1300	231.25	1300
10	327.1	3.03	2000		

The final step was to algebraically combine the operational-to-interim scale transformations for each grade (Table 7) with the interim-to-final scale transformation in Table 8. The results are shown in Table 9.

Table 9.
Operational to Proposed Vertical Scale Transformations

	Reading		Mathematics	
Grade	a (Slope)	b (Intercept)	a (Slope)	b (Intercept)
3	6.072	-521.569	4.625	-87.499
4	5.860	-291.417	4.373	143.659
5	5.598	-86.090	4.719	97.196
6	5.547	-15.777	4.305	339.511
7	5.239	147.116	4.035	554.057
8	4.761	410.068	3.948	630.602
9	5.426	229.605	3.396	898.322
10	5.410	303.295	4.102	658.073

To create Figures 10 and 11, the transformations in Table 9 were applied to the various benchmarks used to map the operational scales onto the proposed vertical scale. The Excel spreadsheet developed for this analysis allowed exploration of how altering operational anchors and proposed vertical scale targets would affect the range of vertical scale values.

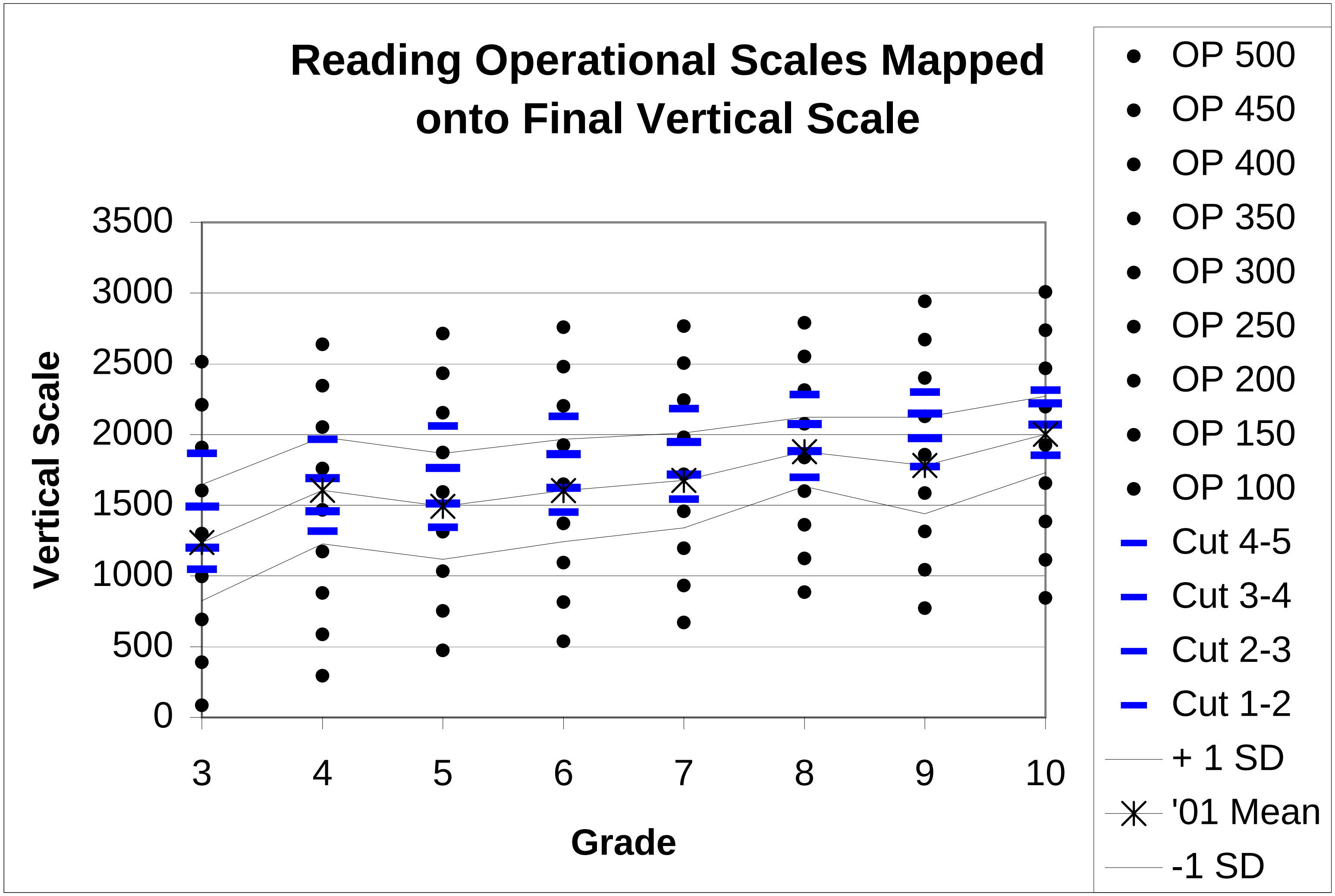

Figure 10. Proposed final vertical scale for Reading with operational scale benchmarks.

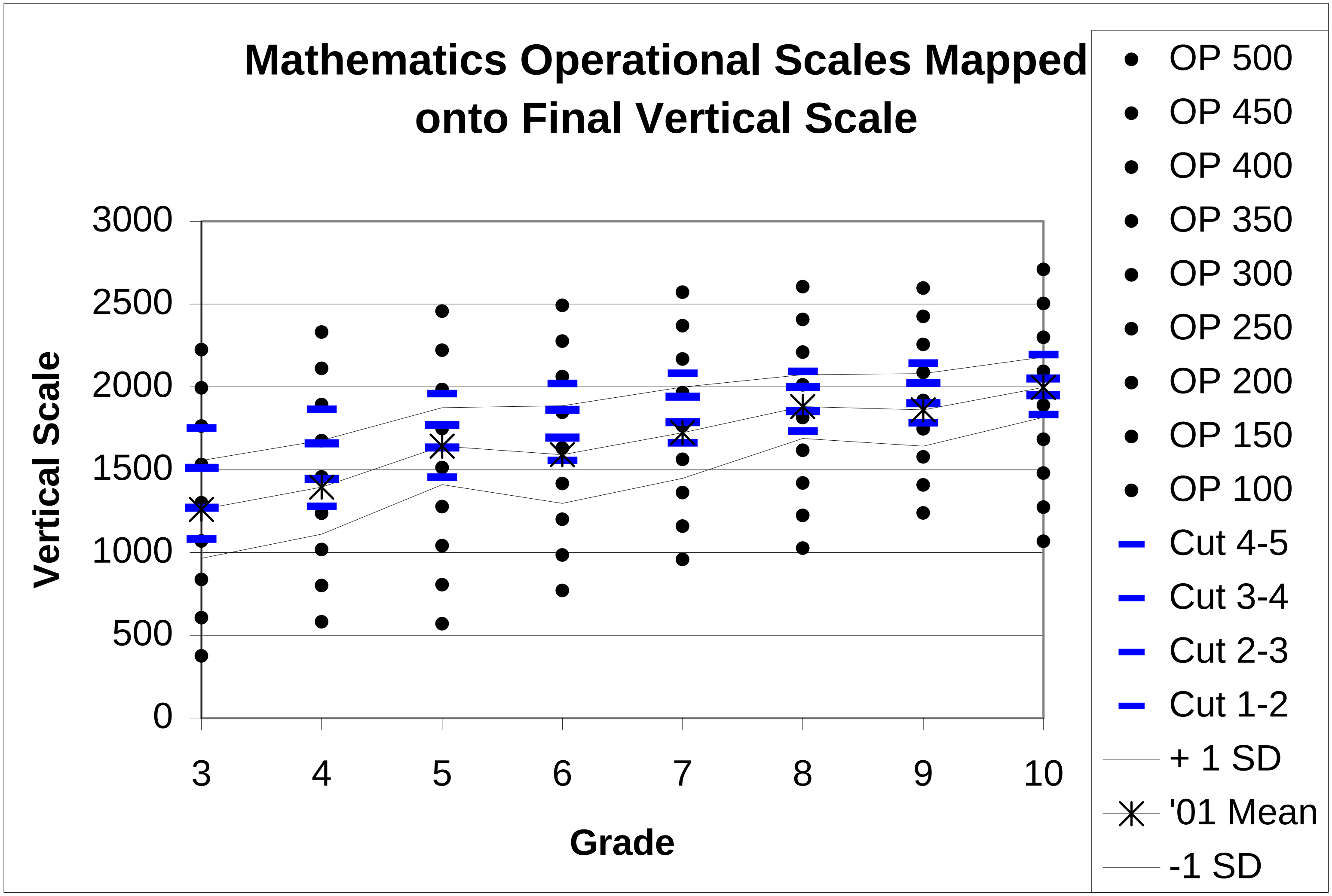

Figure 11. Proposed final vertical scale for Mathematics with operational scale benchmarks.

IMPLICATIONS AND CONCLUSIONS

In addition to illustrating the vertical scale concept, Figures 9 and 10 also stimulate several thoughts relevant to the application of the scale.

Impact of testing

It cannot go unnoticed that when actual FCAT score means were computed on all students and translated to the proposed Reading and Mathematics vertical scales, in all four cases where a new grade follows an initial grade, students' tested achievement was lower in the higher, new grade than in the lower, initial grade. Based on research on the consequential impacts of testing in another state (Thacker, Koger, Hoffman, & Koger; 1999, 2000), we are reasonably confident that these differences in means are a function of differences in testing history for these different grades. Pressure is created by a testing program to improve instruction that has just started for the new grades, but has been influencing the initial grades for several years. The new grades may be "playing catch-up" for the next couple of years. It is also interesting that the lower achievement in the new grades is more pronounced for the lower achieving (–1 standard deviation) students. Monitoring how achievement in new versus initial grades changes in the coming years will provide interesting data for further discussion on the consequential impact of the FCAT.

A common critique of testing programs like the FCAT is that the pressure on school systems to obtain acceptable test scores leads to a narrowing of the school curriculum. We have seen evidence for this argument (Thacker, Koger, Hoffman, & Koger; 1999, 2000). From that perspective, Florida's plan to add other subjects to the FCAT seems to provide an important balance to curriculum emphasis. Field research to assess the impact of testing on schools' curriculum is recommended.

Longitudinal versus cross-sectional data

In the introduction, we noted that the data on which the vertical scale is built are cross-sectional. Given that the data represent a comparison of different cohorts in different grades assessed in the same year, one can wonder how longitudinal data that tracks individual students across grades and years might look. For four grade/subject pairs, we were able to match students tested in an initial grade in 2000 with students tested in the next higher (new) grade in 2001. We converted their scores from both assessments to the proposed vertical scale and calculated their gain.[12] These are the same four initial grade/new grade pairs mentioned

[12] The matching was conducted by the Florida Department of Education using student ID and name.

above as showing a cross-sectional decrement or minimal gains in tested achievement. All students with FCAT scores were included in this analysis so that their longitudinal gains could be compared to the cross-sectional actual score gains for all students that are depicted in Figures 10 and 11.

Figures 10 and 11 and Table 10 contain data for students' Spring 2001 scores in Grades 5 and 9 Reading and Grades 6 and 9 Mathematics. For the longitudinal data in Table 10, only students who could be matched were included. In all four cases, means were 23 to 48 points higher and standard deviations were 10 to 21 points lower for the matched students than for all students. We can conclude that unmatched students, on average, scored lower on the FCAT and added to the variability in FCAT scores. While consistent, these differences are not large.

Table 10 presents score means and standard deviations on the vertical scale and means and standard deviations for students' gains across these two years. In contrast to the cross-sectional data, the longitudinal data indicated that the average achievement gains for these grades/subjects were positive. One explanation for the difference between the longitudinal and cross-sectional results is that the cross-sectional data compare this year's new grade students to **this year's** initial grade students, while the longitudinal data compare this year's new grade students to **last year's** initial grade students. Improvements in instruction in the initial grade could explain the different results. Similarly, improvements in instruction in lower grades in earlier years would have led to cohort differences that would affect the cross-sectional comparisons. Again, monitoring over time will provide insight into the consequences of FCAT on student achievement.

Table 10.
Longitudinal Analysis on Students' Matched FCAT Scores

Grade/ Subject	2000 Scores		2001 Scores		Gains		
	Mean	S.D.	Mean*	S.D.*	Mean	S.D.	% > 0
Reading							
4-5	1444	352	1516 (1493)	362 (374)	72.1 (-112)	215.9	64.0
8-9	1821	251	1829 (1781)	320 (341)	7.6 (-98)	192.0	51.9
Mathematics							
5-6	1591	254	1614 (1591)	285 (295)	22.6 (-51)	163.2	59.7
8-9	1856	218	1895 (1861)	207 (219)	38.9 (-20)	117.7	65.6

*Numbers in parentheses give data for all students.

**Numbers in parentheses show cross-sectional changes in cohort-to-cohort means per Figures 9 and 10.

Although the matched, longitudinal data might be not representative of all grades, the gain data are still instructive. There is considerable variability in gains and, while the majority of students gained in each case, there were also as many as 48 percent who failed to show tested improvement. Figures 10 and 11 also show where students one standard deviation above and below the mean fell on the vertical scale. In order to compare the longitudinal data to the cross-sectional trend data for low (–1 standard deviation) and high (+1 standard deviation) students, we divided our matched students into 9 equal-size groups based on their 2000 FCAT scores and then computed average gains for students in the second group, fifth group, and the eighth group. These groups include students whose scores were centered at –1 standard deviation, the mean, and +1 standard deviation, respectively. Results are shown in Table 11. There was no trend in terms of which segment of the student population gained the most, but it is clear that there was variability in students' tested growth in all segments. Part of that variability in gains was a function of the next topic, score reliability.

Table 11.
Score Gains for Initially Low Scoring, Mid-Level Scoring, and High Scoring Students

Grade/ Subject	Second 1/9th in 2000 (Centered on –1 SD)		Middle 1/9th in 2000 (Centered on Mean)		Eight 1/9th in 2000 (Centered on +1 SD)	
	Mean	SD	Mean	SD	Mean	SD
Reading						
4-5	84	221	61	185	42	193
8-9	-24	201	-1	158	40	207
Mathematics						
5-6	4	190	25	136	30	127
8-9	57	125	25	83	17	81

Reliability

When working with gain scores, one should always be concerned about reliability (Cronbach & Furby, 1970). As indicated earlier, the reliability of gain scores is typically troublesome, particularly so when the two component scores are highly correlated with each other. Using standard deviation data from Table 10, reliability estimates for the Spring 2000 test and the Spring 2001 test (FCAT Technical Reports; 2000, 2001), and the correlation between 2000 and 2001 scores computed on the matched student sample, we computed reliability estimates for the gain scores. As expected, because the correlations between years were high, individual student gain score reliabilities tended to be low, in spite of acceptable reliabilities for the scores from 2000 and from 2001. These data are shown in Table 12.

Table 12.
Reliability Estimates for Students' Vertical Scale Scores

Grade/ Subject	2000-2001 Correlation	Reliability Estimates		
		2000	2001	Gain
Reading				
4-5	0.82	0.90	0.87	0.37
8-9	0.80	0.89	0.90	0.53
Mathematics				
5-6	0.82	0.92	0.87	0.41
8-9	0.85	0.92	0.92	0.48

In addition to the reliability estimates, we also computed correlations among FCAT scores and score gains using the matched student sample for Grades 8 and 9. This analysis adds another way of understanding the instability of student-level gain scores. These correlations are presented in Table 13. As we might expect, correlations between reading scores and mathematics scores (both within and across years) were in the 0.70's. In sharp contrast, the correlation between measured gains in Reading and gains in Mathematics was only 0.16. We corrected this observed correlation for the unreliability in the two gain measures, estimating the correlation between true gains in Reading and Mathematics to be approximately 0.32.

Table 13.
Correlations among Scores and Score Gains for Longitudinal Sample Grade 8 to 9

	2000 Reading Score	2001 Reading Score	Reading Gain	2000 Math Score	2001 Math Score	Math Gain
2000 Reading Score		0.80	0.03	***0.79***	***0.73***	-0.18
2001 Reading Score	135729		0.62	***0.72***	***0.73***	-0.05
Reading Gain	135729	135729		0.17	0.26	**0.16**
2000 Math Score	134523	134518	134518		0.85	-0.36
2001 Math Score	133136	133136	133136	132002		0.19
Math Gain	132002	132002	132002	132002	132002	

Note: Samples sizes are below the diagonal.

Vertical scaling at the school level of analysis

Vertical scale score gains will also be used at the school level of analysis to answer questions about the average achievement gains for students within schools. Table 14 presents descriptive and reliability data. The descriptive data are based on the matched student samples. For schools with at least 15 students, school-level gains were calculated as the average gains of the students within the school. Table 14 shows the means and standard deviations of the resulting school-level gains. The mean school-level gains mirrored the mean student-level gains in Table 10, and as expected, the standard deviations for schools were smaller than the standard deviations for students.

Table 14.
Longitudinal School Level Gains

Grade/ Subject	N of Schools*	Mean N of Students within Schools	Gains		Reliability by Class Size***		
			Mean	SD**	15	100	200
Reading							
4-5	1606	99	69.7	43.7 (22.7)	0.898	0.983	0.992
8-9	413	326	6.9	40.6 (10.6)	0.945	0.991	0.996
Mathematics							
5-6	672	230	27.0	45.62 (10.8)	0.912	0.986	0.993
8-9	409	329	38.2	23.35 (6.5)	0.932	0.989	0.995

*Schools with at least 15 matched students.

**Numbers in parentheses are the standard deviations expected from student-level standard deviation and mean school size if schools were all randomly equivalent.

**Reliability estimates are not necessarily accurate to three decimals. Three are reported, however, to avoid rounding to 1.00.

Reliability estimates showed a marked contrast to the student-level data. We used the Spearman-Brown prophecy formula to project the reliability of school scores based on representative school sizes and the student-level reliability estimates. With as few as fifteen students, school-level gains can be estimated with reliabilities that are traditionally acceptable. Higher reliability, of course, would be obtained for larger schools. This is not an unexpected finding, but is one of the natural consequences of increasing the amount of data used to calculate scores.

To further understand the potential utility for school-level data, we also calculated correlations among school-level data (see Table 15). These data showed generally stronger intercorrelations among all of the variables, with some exceptions. The bold italicized data indicate very strong within year and between year relationships between school means for FCAT Reading and Mathematics. An interesting difference, however, occurs for reading gains and mathematics gains. Schools with initially higher Reading scores tended to show more student growth than schools with initially lower Reading scores (*r* = 0.55). This mirrors the pattern at the individual level, on cross-sectional data, depicted in Figure 4. Mathematics gains, in contrast, are unrelated to the Spring 2000 initial values (*r* = 0.00) which, again, is consistent with the individual level, cross-sectional data (see Figure 5). One consequence of this is that school means for students' Reading and Mathematics gains were very modestly correlated (*r* = 0.24). Since school-level gains were reasonably reliable, this correlation is a reasonable estimate of true gains. Applying the correction for unreliability, as we did for the student-level data, increased the correlation only slightly (*r* =0.26). This suggests that the effectiveness of reading and math instruction are relatively independent.

Table 15.
School-Level Correlations among Scores and Score Gains for Longitudinal Sample Grade 8 to 9

	2001 Reading Score	Reading Gain	2000 Math Score	2001 Math Score	Math Gain
2000 Reading Score	0.97	0.55	***0.95***	***0.96***	-0.17
2001 Reading Score		0.75	***0.96***	***0.94***	-0.06
Reading Gain			0.65	0.57	*0.24*
2000 Math Score				0.97	0.00
2001 Math Score					-0.26

N = 413 Schools.

Considering expected growth

We indicated in the introduction that vertical scaling, per se, could not establish an expectation for how much students should learn from one year to the next. On the other hand, Florida's Achievement Levels establish expectations for schools at each grade level. Top performing schools are benchmarked by the proportion of students above the cut point between Achievement Levels 2 and 3. Failing schools are benchmarked by the proportion of students below the cut point between Achievement Levels 1 and 2. Therefore, differences across grades between corresponding cut points on the vertical scale can supply one perspective on expected student growth. For example, expected growth could be defined by the differences across grades in either the 2-3 cut points or the 1-2 cut points.

Table 16 presents the cut points in vertical scale units between Achievement Levels 1 and 2 and between Levels 2 and 3. The means of students' reported scores are also given for comparison. The table also shows differences between each adjacent grade and the average of the differences. These are the same data used in the construction of Figures 10 and 11, but this display more clearly shows the variability of between-grade differences in cut points.

Table 16.
Achievement Level Cut Points, “Expected” Gains, Current Grade-Level Means, and Current Gains

Grade	1-2 Cut point		2-3 Cut point		Mean	
	Vertical Scale Score	Gain from Lower Grade	Vertical Scale Score	Gain from Lower Grade	Vertical Scale Score	Gain from Lower Grade
Reading						
3	1048		1200		1236	
4	*1317*	*269*	*1458*	*258*	*1605*	*369*
5	1344	27	1512	55	1493	-112
6	1451	107	1623	111	1604	112
7	1543	92	1716	93	1676	71
8	*1698*	*154*	*1884*	*167*	*1878*	*203*
9	1773	75	1974	91	1781	-98
10	*1853*	*80*	*2070*	*96*	*2001*	*221*
Average Gain		115		124		109
Mathematics						
3	1080		1270		1260	
4	1279	198	1445	175	1394	134
5	*1454*	*175*	*1633*	*188*	*1642*	*248*
6	1556	102	1693	60	1591	-51
7	1662	106	1787	93	1723	132
8	*1734*	*72*	*1852*	*66*	*1881*	*158*
9	1783	49	1902	49	1861	-20
10	*1833*	*50*	*1948*	*47*	*1998*	*137*
Average Gain		108		97		105

Note: italics indicate initial Grades (Grades 4, 8 and 10 Reading; Grades 5, 8, and 10 Mathematics).

It is beyond the scope of this report to provide conclusions from the data in Table 16, other than to make an obvious observation. The cut points do increase from grade-to-grade while the cross-sectional means do not always do so. If growth standards were to be based on simple differences in cut point levels, however, growth might be expected to be much greater for some grades than others. Note also that standards for the initial grades are high relative to adjacent grades in that the difference between each initial grade standard and the corresponding standard for the next lower grade is always greater than the difference between the initial grade standard and the standard for the next higher grade. Setting standards for student growth is a matter of policy and cannot be derived unambiguously from these data. We do, however, suggest that the data be considered in policy decisions regarding standards for student growth. Consideration should be given not only to the grade-to-grade differences themselves but also to overall trends depicted by the data, including the relationship between Achievement Level cut points and actual scores.

CONCLUDING REMARKS

The research has been something of an adventure, including the twists and turns of form design and the difficulties of achieving a satisfactory number scale. Five points seem clear, however.

1. A vertical scale has been constructed that expressed the numeric relationship of performance on each of the FCAT grade-level Reading and Mathematics tests to a common measurement scale. Scores expressed on this scale have the same validity and reliability as operational FCAT scores because only the numeric aspect of the scale has been altered.

2. Using the vertical scale to make assessments of achievement growth for individual students seems risky because of the unreliability introduced by computing differences between otherwise reliable FCAT scores.

3. On the other hand, using the vertical scale to make assessments of average achievement growth for the students within schools is far less problematic. By aggregating individual student gains to compute school mean gains, reliability is significantly increased.

4. The vertical scale does not specify expected student growth. However, data on the vertical scale, including trends for Achievement Level cut points and current levels of student growth, can provide valuable input to creating policy related to expected gains.

5. Our final point stems from the dynamic context of school improvement. That is, both longitudinal and cross-sectional analyses of achievement gains for differential levels of students must be an on-going activity in order to gain further insight about changes in educational achievements.

REFERENCES

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

Cronbach, L. J. & Furby, L. (1970). How we should measure “change” – Or should we? *Psychological Bulletin, 74,* 68-80.

Diaz, T. E. & Wise, L. L., (2000). *Effects of Item Order for the Florida Comprehensive Assessment Test (FCAT).* San Antonio, TX: Harcourt Educational Measurement.

Human Resources Research Organization and Harcourt Educational Measurement (2001). *Florida Comprehensive Assessment Tests: Technical Report For Operational Test Administrations of FCAT 2000.* San Antonio, TX: Harcourt Educational Measurement.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McBride, J. R. & Wise, L. L., (2000). *Developing a Vertical Scale for the Florida Comprehensive Assessment Test (FCAT)*. Unpublished. Alexandria, VA: Human Resources Research Organization.

Stocking, M. L. & Lord, F. M., (1983). Developing a common metric in item response theory. Applied Measurement, 7, 201-210.

Thacker, A.A., Koger, L.E., Hoffman, R.G., & Koger, M.E. (2000). *The transition from KIRIS to CATS, Year 2: Instruction, communication, and perceptions at 31 Kentucky schools* (FR-00-23). Alexandria, VA: Human Resources Research Organization.

Thacker, A. A., Koger, L. E., Hoffman, R. G., & Koger. M. E. (1999). *The Transition from KIRIS to CATS: Instruction, Communication, and Perceptions at 20 Kentucky Schools.* (HumRRO Draft Report DFR-WATSD-99-23). Alexandria, VA: Human Resources Research Organization.

Thissen, D. (1991). *Multilog™ User’s Guide.* Lincolnwood, IL: Scientific Software.

APPENDIX A: FURTHER OUTLIER ANALYSIS FOR FCAT VERTICAL LINKING

When common items are used to link Item Response Theory (IRT) scales created from different forms, each common item has separate IRT parameter estimates from each form. For each form, the item parameters define a function that gives the expected probability of a correct response for different achievement levels. (This function is called the item characteristic curve.) The idea of the Stocking/Lord linking between forms is to identify a linear adjustment to the item parameters such that a student with a given ability would have the same probability of answering correctly regardless of which form he/she took. To forecast potential problems in linking one form to another, it is common to review scatterplots that compare item parameters for each form. Thus, we routinely create three separate scatterplots plots: one that plots “a” parameter values from each form, one for “b” parameters from each form, and one for “c” parameters from each form. We examine these plots for items that are outlier items that could have an undue effect on the Stocking/Lord linear transformation solution.

For the vertical grade-to-grade linking, the “a,” “b,” and “c” plots tended to be more scattered than typically seen in FCAT’s year to year linking for a single grade. We subsequently traced this to differences between items operational in the higher grade versus items operational in the lower grade. We then explored separately linking the grades using only the items operational in the higher grade and linking the grades using only the items operational in the lower grade of each link as a means of understanding how the linking items were behaving. This led to our concerns about item position, our further exploration of p-values, and the decision to weight items in final linking to create a balance between items operational in the higher grade and items operational in the lower grade.

In the spirit of leaving no stone unturned, we returned to these “higher-only” and “lower-only” linking procedures, examined the “a,” “b,” and “c” plots for items that might be behaving differently in the higher and lower grade forms relative to the other items in each of the respective linkings. Differences in individual item parameters sometimes interact. For example, the effects of a lower “c” parameter might be largely offset by an increase in the “a” parameter. In identifying outliers, we focused on the predicted probabilities of correct responses, rather than the individual item parameters. For each item, we computed the probability of passing for low (one standard deviation below the lower grade mean), moderate (at the mean) and high (1 standard deviation above the mean) performing students. We compared the predicted probability using the item’s parameters from the lower-grade form with the probabilities computed using the parameters from the higher-grade form after the Stocking/Lord adjustment was applied. We computed the mean and standard deviation of all of the differences at each Achievement Level and flagged items where the differences were particularly large (more than 2.58 standard deviations, corresponding to a 0.01 significance level).

Tables A-1 and A-2 list items for which the overall Stocking/Lord adjustment did not fully eliminate differences in expected performance for low, medium, or high achieving students. These items are not necessarily "bad" items. For example, there may be valid curricular reasons that a particular item was harder (or easier) or more or less discriminating in one grade or the other. In addition, there might be subtle, but real, content/construct shifts from one grade to the next, such that an item does not relate to the content/construct of one grade the same way that it relates to the content/construct of the other grade.[13] On the other hand, outlier statistics could also signal a testing context difference that would not be a valid reflection of differences between grades. Such differences include:

(1) differences in cueing by adjacent items,

(2) changes in item layout in the test books, or even

(3) misprinting in one of the books.

The items in the tables below met our "potential outlier" criteria ($t > 2.58$, where the probability of chance differences this large was less than a 0.01). We would like to have these items reviewed for context and/or printing differences in their respective test forms that may have created differences in student responses unrelated to the standard being assessed. If such is the case, these items should be excluded from linking. If no test books' problems are detected, all but the most extreme items will remain in the linking on the assumption that grade-to-grade differences in their IRT data signal valid curricular differences.

On the basis of the statistics alone, we have identified 11 items that meet more stringent criteria. These include 3 Reading items and 8 Mathematics items. The items were identified as items with highly significant deviations in the predicted probability of a correct response (i.e., $t > 4$) where the probabilities differed by more than 0.10 (i.e., where the differences were also practically significant). These are shaded in the tables below. All 3 identified items for Reading were operational in the lower grade and therefore field test position items in the higher grade. Since there were more lower-grade items to start with, eliminating these items will help in restoring the balance between higher- and lower-grade items. Removing them will affect 2 linkages. All 8 items identified for Mathematics, 6 were operational in the lower-grade. Removing the 8 mathematics items will affect 5 linkages. We have recomputed Stocking/Lord linking constants after removing these items. The impact on overall scaling is negligible for both subjects.

[13] That is, the construct captured by IRT scaling for one grade may be different from the construct captured by IRT scaling for the other grade.

Table A-1.
Items Flagged as Possible Outliers for Reading[14]

Operational Grade		Off-Grade			Outlier Predicted Difficulties			
Grade	Item Number	Grade	Form	Item Number	Ability Level	Lower Grade	Higher Grade (Adj.)	t
3	12	4	12	45	High	0.631	0.726	3.812
3	34	4	13	43	High	0.825	0.887	2.585
4	18	3	12	46	Med	0.586	0.503	-3.517
					High	0.839	0.771	-3.916
4	3*	5	11	47	Med	0.407	0.559	4.076
					High	0.599	0.729	5.327
4	29*	5	13	46	Med	0.250	0.376	3.431
					High	0.554	0.650	4.100
5	31	6	11	49	Low	0.742	0.533	-3.033
5	17	6	12	51	High	0.874	0.773	-2.965
5	37	6	13	47	High	0.605	0.726	3.834
5	43	6	13	51	High	0.670	0.768	3.117
6	13	7	11	50	Low	0.685	0.546	-3.117
7	4	6	14	48	Med	0.599	0.745	3.398
7	22*	8	13	46	High	0.558	0.696	4.679
8	13	9	13	51	Low	0.310	0.487	2.964
9	4	8	14	48	Med	0.663	0.758	2.773
9	24	8	15	47	Med	0.801	0.680	-3.650
9	13	10	12	48	High	0.340	0.430	2.837
9	28	10	13	46	Med	0.596	0.719	2.860
10	21	9	14	47	Low	0.484	0.398	-2.922

[14] The items highlighted in yellow, which are also indicated with one asterisk (*), were identified as items with highly significant deviations in the predicted probability of a correct response (i.e., *t* > 4) where the probabilities differed by more than 0.10 (i.e., where the differences were also practically significant). This means that these items were significantly more difficult for the higher grade than the lower grade. The items highlighted in yellow, which are also indicated with two asterisks (**), were identified as items with highly significant deviations in the predicted probability of a correct response (i.e. *t*7-4 > -4). This means that these items were significantly more difficult for the lower grade than the higher grade.

Table A-2.
Items Flagged as Possible Outliers for Mathematics[15]

Operational Grade		Off-Grade			Outlier Predicted Difficulties			
Grade	Item Number	Grade	Form	Item Number	Ability Level	Lower Grade	Higher Grade (Adj.)	t
3	2	4	11	44	Low	0.649	0.538	-2.703
3	31*	4	11	45	Med	0.337	0.495	3.775
					High	0.589	0.830	4.382
3	21	4	13	43	Low	0.398	0.292	-2.595
4	20	3	15	44	High	0.505	0.631	3.368
4	25**	3	15	45	High	0.870	0.693	-4.873
4	29*	5	11	47	Med	0.433	0.612	2.977
					High	0.671	0.850	7.759
4	3	5	12	45	Low	0.745	0.662	-3.133
4	7	5	13	47	Low	0.728	0.493	-5.916
4	15	5	13	48	High	0.814	0.697	-3.983
4	33*	5	13	50	Low	0.273	0.395	2.980
					Med	0.412	0.589	2.944
					High	0.630	0.773	6.313
4	30	5	13	46	High	0.543	0.611	3.339
5	39	4	14	46	High	0.691	0.846	2.947
5	7**	6	13	46	Med	0.735	0.584	-4.118
					High	0.887	0.785	-3.533
6	3	5	14	47	Med	0.630	0.769	3.430
6	38	7	13	45	Low	0.182	0.319	3.109
6	10	7	13	45	Med	0.389	0.567	2.618
7	15	6	14	46	Low	0.420	0.271	-2.583
7	8	6	15	45	Low	0.171	0.364	3.223
7	22	8	12	49	High	0.475	0.600	3.736

[15] The items highlighted in yellow, which are also indicated with one asterisk (*), were identified as items with highly significant deviations in the predicted probability of a correct response (i.e., *t* > 4) where the probabilities differed by more than 0.10 (i.e., where the differences were also practically significant). This means that these items were significantly more difficult for the higher grade than the lower grade. The items highlighted in yellow, which are also indicated with two asterisks (**), were identified as items with highly significant deviations in the predicted probability of a correct response (i.e. *t*7-4 > -4). This means that these items were significantly more difficult for the lower grade than the higher grade.

Table A-2.
Items Flagged as Possible Outliers for Mathematics[15]

Operational Grade		Off-Grade			Outlier Predicted Difficulties			
Grade	Item Number	Grade	Form	Item Number	Ability Level	Lower Grade	Higher Grade (Adj.)	t
7	35*	8	13	47	Low	0.216	0.305	4.042
					Med	0.264	0.353	3.879
7	40*	8	12	45	Low	0.193	0.307	5.022
					Med	0.574	0.673	3.922
7	14	8	11	45	Med	0.261	0.350	3.885
8	22	7	15	47	Med	0.720	0.430	-3.928
8	40*	9	12	46	Low	0.307	0.383	4.212
8	20	9	12	50	Low	0.246	0.181	-3.122
8	18*	9	13	48	Low	0.259	0.332	4.048
8	29	9	12	47	High	0.491	0.575	2.647
8	38*	9	11	46	Low	0.180	0.251	3.964
					Med	0.463	0.581	4.712
9	21**	8	14	50	Med	0.742	0.611	-4.809
9	28	8	14	45	Med	0.707	0.628	-3.112
9	11	10	12	47	Med	0.167	0.302	2.860