

Developing a Vertical Scale for the Florida Comprehensive Assessment Test (FCAT)

**James McBride and Laress Wise
Human Resources Research Organization
(HumRRO)
Alexandria, Virginia**

**Under subcontract to
Harcourt Educational Measurement
San Antonio, Texas**

FINAL DRAFT



San Antonio, Texas

May 2001

TABLE OF CONTENTS

Overview.....	1
Introduction.....	2
Scaling Methods and Practices	3
Data Collection Approaches	5
Scaling Methods	6
Grade-Equivalent Scaling	6
Thurstone’s Absolute Scaling Methods (Normal Curve Equivalent).....	7
IRT Scale Linking.....	8
Direct IRT Scale Transformation	8
IRT True-Score Equating.....	9
IRT Observed-Score Equating.....	9
Major Test Publishers’ Practices	11
Threats to Vertical Scaling	13
Score Precision Differences.....	13
Multidimensionality.....	13
Construct Differences	13
Exposure to Content.....	14
Practice and Fatigue Effects	14
A Plan for Vertical Scaling of FCAT 2001	14
Anticipated Constraints.....	14
Proposed Data Collection Design.....	16
Number of Anchor Items	16
Item Difficulty	17
Content and Format Coverage	17
Example	18
Approach to Scaling.....	18
Computation.....	18
Final Scale.....	19
Initial Use.....	19
Follow-On Studies	19
Protection Against Threats.....	20
Score Precision Differences	20
Multidimensionality	20
Construct Differences	20
Exposure to Content.....	20
Practice and Fatigue Effects.....	21
Annotated Bibliography.....	21
Appendix A: Figures.....	29
Appendix B: Glossary.....	33

OVERVIEW

To date, assessments in the Florida Comprehensive Assessment Test (FCAT) program have been grade-level specific, with an emphasis on comparing the performance of students at each grade with the performance of previous students. The Florida Department of Education is investigating the possibility of expressing all test scores on the same scale, regardless of grade level, to make comparisons between grades within a year. This will require the development of a single score scale, spanning the entire range of student performance for Grades 3 through 10. A scale of this kind is called a *developmental scale* or a *vertical scale*, and the process of developing such a scale and placing the assessment scores of all grades on it is called *vertical scaling*.

This purpose of this report is to

- describe vertical scaling, including how it is similar and different from the scaling currently used in FCAT;
- review historical and current scaling methods and practices;
- address the scaling practices of major test publishers;
- describe possible factors that can adversely affect the accuracy of vertical scaling;
- propose a plan for the vertical scaling of FCAT;
- describe the computational procedures for establishing a vertical scale for FCAT; and
- explain how the plan protects against threats to the accuracy of the vertical scale.

This report contains numerous technical terms used in the fields of psychometrics and educational measurement. Readers who are not familiar with these terms may wish to consult the glossary of terms in Appendix B.

INTRODUCTION

Harcourt Educational Measurement is assisting the state of Florida with its statewide educational assessment. To date, all of the assessments have been grade-level specific, with an emphasis on comparing the performance of students at each grade with the performance of previous students. This has been accomplished using a test score scale calibration procedure by which test scores within each grade can be expressed on a common, grade-specific scale that spans the duration of the assessment program. Although the scale calibration procedure has enabled comparisons of same-grade performance across years to be made, grade-to-grade comparisons have not been made to date. Each grade has had its own score scale, and it has not been possible to compare performance across grades. Consequently, no information has been available as to differences across grades in student proficiency or the degree of overlap between grades. Furthermore, the separate scales make it difficult to measure individual student growth from one grade to the next, as might be needed for a value-added model of school accountability.

Data from the 2001 assessment will be used to develop the capability to express all test scores on the same scale regardless of grade level and to make comparisons between grades in any given year. This will require creation of a single score scale for the state assessment, spanning the entire range of student performance for Grades 3 through 10. A scale of this kind is called a *developmental scale* or *vertical scale*, and the process of developing such a scale and placing all grades' assessment scores on it is called *vertical scaling*.

Vertical scaling is both similar to, and materially different from, the scale calibration procedures previously used to place scores at a specific grade level on a common metric from year to year. The similarities may be characterized as *procedural* and the principal differences as *substantive*. That is, many (but not all) of the statistical procedures that have been proposed for use in vertical scaling are formally similar to the more established procedures applied to horizontal scale calibration. Horizontal calibration puts scores from forms designed to have the same level of difficulty on a common scale; these procedures come from the closely related field of test equating. For example, an objective of both horizontal and vertical processes is development of a mathematical equation for transforming test scores from one scale to another; such transformation equations are often linear, and the procedures for deriving them are formally similar. Two notable, substantive differences include differences in the difficulty of test forms and the general ability levels of students at different grades. In the horizontal calibration problem, different test forms are similar in content and difficulty level, and form-to-form differences in the overall ability levels of students in a

given grade are small. In contrast, vertical scaling deals with student populations that differ substantially in ability from grade to grade and with tests that differ accordingly in difficulty. Another important substantive difference is that the tests designed for the assessment of students at different grades necessarily differ in content, reflecting grade-to-grade differences in curriculum. Curriculum and content differences between nonadjacent grades may be profound; this poses a problem for the enterprise of vertical scaling that generally does not occur in within-grade scale calibration.

Before turning to a specific plan for developing a vertical FCAT scale, we provide some background on previously used approaches to constructing vertical scales. This information is followed by a brief section listing key threats to the validity of vertical scales. We conclude with a discussion of specific plans for constructing the FCAT scale.

SCALING METHODS AND PRACTICES

Test score scale calibration procedures have their roots in the practice and the methods of test equating. Test equating procedures are concerned with a fairly narrow technical problem: how to account for small differences in the psychometric characteristics of different forms of a test, all of which were designed to be similar in content, difficulty, and length. Lord (1980) points out that, strictly speaking, test equating can be technically successful only if the tests to be equated are parallel—identical in all psychometric characteristics. The irony is that parallel tests do not need to be equated, as scores on one test are interchangeable with scores on the other. And nonparallel tests cannot be equated in the strict sense of the term. With that cautionary note, Lord proceeds to outline methods for approximating the equating of imperfectly parallel tests and to point out some of the statistical pitfalls that are necessarily encountered in the process. Much of the art of test equating lies in minimizing the consequences of attempting the impossible.

The objective of horizontal scale calibration differs from that of test equating in that the tests one attempts to place on a common scale are typically not parallel. By design, they may vary in both content and psychometric characteristics such as length, difficulty, and the interrelationships among constituent items. Consequently the task of horizontal-scale calibration is susceptible to all the pitfalls inherent in test equating and to additional sources of statistical error arising from the inevitable differences between the tests to be calibrated.

Vertical scaling is subject to all these sources of error and more, as the scaling problem extends to even greater differences in content, difficulty levels of the tests to be calibrated, and ability levels of the respondents to those differing tests.

Despite the extensive practice of equating and scaling in the test publishing industry, until recently relatively little research and methodology was available to guide practitioners. Angoff's *Scales, Norms, and Equivalent Scores* (Angoff, 1984) provides one of the most comprehensive overviews available of test equating and scaling issues and practices. Although Angoff's monograph alludes to the application of item response theory (IRT) to equating and scaling, it is almost entirely dedicated to approaches based on classical test theory. Originally written as a chapter in the 1971 edition of *Educational Measurement*, it was updated by Peterson, Kolen, and Hoover (1989); their 1989 chapter incorporates IRT methods as part of the state of the art. The recent volume by Kolen and Brennan (1995) provided for the first time a book-length compilation of equating and scaling theory and practice, addressing both classical methods and IRT applications.

Whether a classical or an IRT approach is taken, the object of placing tests with different psychometric characteristics on the same scale is to determine the parameters of a linear or monotonic equation that may be used to transform test scores on any one test, X , to equivalent scores on the common scale, C . In classical approaches, the test scores may be expressed as raw scores (number correct), standard scores, or some other transformation, but the problem is the same—to determine the parameters A and B of the linear transformation

$$C = AX + B$$

where X is in the metric of scores on the test to be scaled, C is in the metric of the common scale, and A and B are the slope and intercept, respectively, of the linear transformation equation.

When multiple tests with different psychometric characteristics are administered to groups of examinees who differ in ability and are to be placed on the scale C , the general practice is to arrange the tests by levels, develop scale transformations for converting adjacent levels, and conduct a series of scale transformations—two levels at a time—until all of the test levels have been transformed to a common reference scale. Sometimes an interim scale is employed for these purposes; after all levels have been converted to the interim scale, it may then be transformed into a final scale, perhaps a scale that has more convenient properties than the interim scale and/or one that lends itself more readily to interpretation. Two such scales are (1) the familiar SAT score scale, which ranges from 200 to 800, with a mean of 500 and a standard deviation of 100 in the reference population, and (2) the IQ scale, which has a mean of 100 and a standard deviation of 15 (16 for some versions). Neither of these, however, is a vertical scale.

This kind of test scaling has two important methodological components: collecting data appropriate to the problem and applying the statistical methods involved in the scaling itself. These will be discussed in separate sections below.

Data Collection Approaches

Angoff (1984, pp. 93–120) lists six different "designs" for test equating. Each design encompasses a different data collection procedure as well as a different scaling procedure. However, many of the data collection designs used for test *equating* are not useful for vertical scaling of educational assessment tests because they require either equivalent ability groups, equivalent tests, or both. Two of Angoff's six designs—Design IV and Design VI—may be considered adaptable for vertical scaling of school tests. Design IV is described as "nonrandom groups, with a common equating test," and Design VI encompasses methods based on item data (rather than test scores), with some common items in the tests scaled together.

Angoff (p. 110) summarizes the data collection aspect of Design IV as follows: "Form X is administered to Group *a*; Form Y is administered to Group *b*. Form U, a test which is based on a set of items in addition to (or included among) those represented by Forms X and Y, is administered to both Groups *a* and *b* and is used to adjust for differences that may be found to exist between them." One critical element of Design IV is that the equating test, Form U, is scored separately. The relationships of scores on Forms X and Y to scores on Form U will be used to effect the scaling adjustments.

In contrast, the otherwise similar Design VI does not use test scores to derive the scaling adjustments. Instead, it requires some items in common between Forms X and Y and uses item scores rather than test scores to derive the scaling adjustments. Design VI is an appropriate data collection design for use with Thurstone's method of absolute scaling and is useful with IRT scaling methods as well.

There are two broad approaches to scaling using a common equating test: anchor test designs and scaling test designs. In the context of equating multilevel school assessment tests, anchor tests differ from scaling tests in two important ways: (1) anchor tests are designed to be appropriate in content and difficulty for a relatively narrow range of ability or achievement, while scaling tests span a much broader range of content and difficulty; and (2) a unique anchor test is typically administered at just two adjacent levels, while the same scaling test is administered over a broad span of ability levels. Although there can be only a single scaling test, there would be as many anchor tests as there are pairs of adjacent test levels to be scaled together. For example, if scores of third, fourth, and fifth grade students on grade-level forms R, S, and T are to be placed on a common scale using an anchor test design, there will be two anchor tests, one administered to students in Grades 3 and 4 and the other administered at Grades 4 and 5. In contrast, in a scaling test design the same scaling test would be administered to all three grades.

Scaling Methods

A wide variety of approaches may be used to link anchor test or common or equivalent group designs to different score scales. Kolen and Brennan (1995) provide a reasonably complete discussion of alternative approaches to equating two score scales, which we will not try to replicate here. Vertical scaling across many grades typically proceeds by designating one grade as the base and then linking each succeeding grade to the preceding one. The scores linked may be total raw scores or IRT score estimates derived from one model or another.

In many cases, the composite scale is derived from linear or equipercentile translations of grade-specific number-correct or theta scales. However, this is not necessarily the final step, since it is possible to apply monotonic transformations of an original composite scale to produce a new scale that has more specific meaning for grade-to-grade differences. Below we discuss two examples, commonly called grade-equivalent (GE) and normal-curve equivalent (NCE) scales.

Grade-Equivalent Scaling

The ITBS Developmental Score Scale (DSS) is a grade-equivalent scale, transformed to a different metric and used to report scores of all levels of the ITBS tests. As summarized by Peterson, Kolen, and Hoover (1989), vertical scaling of the multilevel ITBS tests involves the following steps:

1. A reference test called a *scaling test* is developed by sampling several items from every level of the ITBS tests. The number of items constituting the scaling test is specified to ensure that the test can be administered in a single sitting. This test includes content appropriate to Grades from 3 through 8 and is administered to representative samples of students in those grades. The distribution of raw scores at each grade is determined and transformed to an estimated distribution of true scores by linear rescaling to obtain a distribution with the same mean and shape as the raw score distribution, but with a smaller variance [$r(xx)$ times the raw score variance, where $r(xx)$ is the estimated within-grade reliability of the scaling test].
2. Median scores for the scaling test are determined for students at Grades 1 through 12; medians for grades 3 through 8 are taken from the actual data, and those for grades 1 and 2 and 9 through 12 are estimated through extrapolation.
3. A table of percentiles is constructed using the grade medians and the estimated raw score distributions at each grade. This table has one row for every grade in the sample and one column for every within-grade median score. The entries in each row are the within-grade percentiles corresponding to each median raw score. (The grade value assigned to each column is the grade and month at

which the scaling test was administered—for example, *gg.5*, where *gg* indicates the grade and *.5* indicates the fifth month of the school year. The numerical values so assigned define the “interim scale” employed in the vertical scaling of the ITBS.)

4. Raw-score to grade-equivalent conversion tables are developed for each level of the ITBS by administering that level to a representative sample of students at each appropriate grade. For each level of the test, percentile ranks are tabulated by grade from the sample data. For each test level, raw scores are converted to grade-equivalent scores by the following procedure: (1) the within-grade percentile rank of the raw score is determined; (2) the grade-equivalent score corresponding to that percentile rank is determined from the grade-specific row of the table developed at Step 3. (Interpolation is used in the case of scores that are intermediate between table values.)

In the ITBS scaling test approach, each level of the multilevel test is administered to a representative sample of students at the appropriate grade(s); the scaling test, which is used to define the scale, is administered to all grades of a separate sample. The anchor test approach is similar in that each test level is administered only to the grade(s) for which it was designed. It differs from other approaches in that an additional set of items—the anchor test—is also administered to the same students. The same set of anchor items is administered at adjacent grades, which provides the data for scale linking between adjacent grades. The items that comprise an anchor test are appropriate for the adjacent grades and may be either embedded in the test at each level or administered as a separate test.

Thurstone’s Absolute Scaling Methods (Normal Curve Equivalent)

Williams, Pommerich, and Thissen (1998) note that Thurstone’s methods of absolute scaling (Thurstone, 1925, 1938) were the most generally used approaches to vertical scaling before the introduction of IRT methods. They describe two different absolute scaling methods developed by Thurstone, one based on item data (Thurstone, 1925) and another based on test score data (Thurstone, 1938).

Both methods require data from two groups of examinees on a set of common items (in addition to the unique items of the two forms to be scaled). The common items are used to establish the scale transformation, which can then be applied to convert scores from one group’s scale to that of the other.

Both the 1925 and the 1938 methods begin by transforming values of p on the common items to standard normal deviations, z . In the 1925 method, p is the proportion of these items correct, computed separately for each group; in the 1938

method, it is the percentile equivalent to an examinee's raw score on the common items in the group to which he or she belongs.

The second step is to plot the z -values obtained for one group against those of the other group and examine the scatterplot for linearity. If the correlation between the two sets of values is high and the scatterplot demonstrates a linear trend, the scaling proceeds.

The third step is to calculate the mean and standard deviation of the z -values for each group, then calculate the parameters of the linear equation that will transform the z -values of one group to transformed values that have the same mean and standard deviation as the z -values in the other group. Once derived, the transformation is applied to examinees' scores on the test that is to be rescaled.

IRT Scale Linking

Unlike classical test theory, item response theory (IRT) analysis of the item responses of examinees places both item difficulty and examinee ability on the same scale. Analysis of a different test administered to a different examinee group produces a similar but different scale. To equate the two tests using IRT methods simply requires transforming one of the IRT scales to the metric of the other. This is straightforward, provided that there are some common items or some examinees who took both tests. Linear scale transformations can be derived readily from the two different sets of values of the common items' difficulty parameters or of the common examinees' ability parameters. Kolen and Brennan (1995, pp. 163-208) provide a detailed summary of IRT scale transformation and equating procedures; here we will simply summarize some of the alternative approaches to scaling based on IRT. There are three broad alternative approaches: (1) direct-scale transformation; (2) true-score equating; and (3) observed-score equating. We discuss each approach below.

Direct IRT Scale Transformation

Kolen and Brennan list three data collection designs that lend themselves to IRT scaling methods: (1) administration of two (or more) different tests to randomly equivalent groups; (2) administration of the different tests to a single group with counterbalancing; and (3) administration of common items to nonequivalent groups. Under the randomly-equivalent-groups design, two or more equivalent samples each take a single test. Under the single-group design, the entire sample takes two (or more) tests, with order of administration counterbalanced. Under the common-items/nonequivalent-groups design, each sample takes a different target test, with some items taken by both samples; the common items (called anchor items) may be embedded in each form of the target tests or may consist of a separate anchor test form.

Under all three designs, IRT parameter estimation and ability estimation are performed, using the item response data that has been collected. Under the randomly-equivalent groups design, these analyses may be performed separately for each test, and the equivalence of the groups produces item and ability parameters that are on the same scale for all tests. Under the single-group design, the parameters for all of the tests may be estimated in a single analysis of all of the data; this produces parameter estimates that are all on the same scale. Under the nonequivalent-groups design, the parameters for each test (including the anchor items) are estimated separately; the relationships between the two resulting sets of parameter estimates for the anchor items may be used to estimate the scale transformation from one test's parameters to those of the other.

Variations of the procedures outlined above also may be used. For example, the IRT analyses may be performed separately for each test under the randomly-equivalent-groups and single-group designs, and scale transformations may be calculated based on the ability parameters. In addition, response data from all tests of the common items/nonequivalent-groups design may be analyzed in a single run of the IRT parameter estimation program. The presence of the common items will, in principle, put all item parameter and ability estimates on the same scale.

IRT True-Score Equating

IRT true-score equating is based on test characteristic curves, which are mathematical formulae for the expected value of the test score—that is, the true score—conditional on examinee ability level. If two or more tests are calibrated on the same ability scale, true scores on the different tests are equivalent if they correspond to the same ability level. Thus, if the IRT item parameters of the different tests are known, the true scores of each test may be calculated as a function of ability level, and equivalent scores may be determined via tabular, graphic, or numerical estimation methods. In practice, parameter estimates are used in place of the unknown parameter values. Data collected under the randomly-equivalent-groups and single-group designs lend themselves readily to true-score equating, as the IRT parameters are estimated on a common scale.

IRT Observed-Score Equating

Despite the similarity of their names, IRT observed-score equating is very different from true-score equating. In observed-score equating, the IRT model parameters are used to estimate the observed-score distributions of the different tests for a given population, and two tests are equated by equipercentile equating based on those distributions. As with true-score equating, data collected under the randomly-equivalent-groups and single-group designs lend themselves readily to observed-score equating, as the underlying distribution of ability is the same for both tests.

Major Test Publishers' Practices

Among the most significant practitioners of vertical scaling are the three major U.S. publishers of educational achievement tests: CTB/McGraw Hill, Harcourt Educational Measurement, and Riverside. Their test batteries include the California Achievement Tests (CAT), the Comprehensive Tests of Basic Skills (CTBS), the Stanford and the Metropolitan Achievement Tests (SAT and MAT), and the Iowa Tests of Educational Development (ITED) and Iowa Tests of Basic Skills (ITBS). All six of those batteries employ vertical scales to express the proficiency of students from kindergarten through twelfth grade. However, each publisher has its own approach to scale development. CTB and Harcourt both employ anchor-test scaling methods based on IRT; CTB uses the three-parameter logistic (3-PL) response model, while Harcourt prefers the Rasch one-parameter model. Riverside uses IRT for scaling some tests, but not for its educational achievement tests. In one technical document, Riverside notes that "numerous studies performed at the University of Iowa during the past two decades have shown IRT scaling methods to be problematic when applied to the vertical scaling of achievement test batteries. Educational achievement growth across time does not fit well with the mathematical IRT model. For this reason, scaling tests were used to produce a single common score scale across all levels of ITBS/TAP/ITED, Forms K and L. This scale is called the Developmental Standard Score (DSS) scale" (Riverside, 1994, p. 68).

Riverside's scaling test approach is summarized in some detail by Peterson, Kolen, and Hoover (1989). At its core is the practice of developing a single test, the scaling test, with content that spans a wide developmental range. For example, in the process of scale development the same scaling ITBS test is administered to representative samples of students from Grades 3 through 9; the test includes content appropriate to the entire grade span. The scaling test performance of students at each grade is used to convert raw scores on the scaling test to an interim grade-equivalent scale. New ITBS forms are also administered at grade-appropriate levels and their raw scores are equated to the scaling test scores. The end result is that raw scores of every level of the ITBS tests are transformed into the DSS scale based on the scaling test.

For equating levels of CTBS/4, CTB collected scale-linking data using a hybrid of random groups and anchor test design. At each grade, they randomly assigned students to two tests—an on-level test and a specially-constructed linkage test. The linkage tests designed for adjacent grades overlapped, providing common items administered to nonequivalent groups. IRT scaling of CTB's tests uses the three-parameter logistic model. Adjacent test levels' item parameters were placed on a common scale after being estimated in a single IRT analysis of the two on-level tests and the two overlapping linkage tests (CTB, 1991, p. 112). This process was repeated for pairs of grades, then characteristic curve equating methods (Stocking and Lord, 1983) were used to place all levels on a single interim IRT scale. The final scale was a linear transformation of the interim scale

designed to make all scale scores positive integers spanning the range from 1 to 999.

Yen and Burket (1997) conducted a simulation study of vertical scaling of the CAT/5 Reading Comprehension and Mathematics Computation tests. They provided an example that includes some data from different tests administered to a single group—for example, first graders were administered test levels 11 and 12, and second graders were given levels 12 and 13. Simultaneous estimation of the 11/12 item parameters was accomplished in one data analysis run, resulting in placement of all items from both levels on one difficulty scale. A second scale was produced by a similar process for the response data from levels 12 and 13. These two scales were different, but linearly related. Item characteristic curve methods (Stocking-Lord) were used to resolve the scale differences, with the result that item parameters of all three levels were placed on a single vertical scale of item difficulty/examinee ability.

In contrast to CTB's use of the 3-PL model and its practice of equating through the IRT item parameters, Harcourt tests use the Rasch model and are equated through the test scores. Harcourt uses a single-group data collection design. As summarized in the norms booklet for Stanford 8 (The Psychological Corporation, 1992), each student in the "equating of levels program" was administered tests at two adjacent levels—on-grade and one level below grade—in counterbalanced order. This provides two test scores for each student; equating is done using test scores.

It is difficult, if not impossible, to perform a comparative evaluation of practical implementations of different vertical scaling procedures because of the limitations of real test data and the unavailability of any absolute criterion. One matter of interest in evaluating vertical scales has been the pattern of grade-to-grade standard deviations of the resulting scale scores. Analyses of their published norms (Riverside, 1994) indicates that in reading and mathematics tests at least, the Riverside developmental scale scores show increasing score variation from grade to grade. Similar analyses of CTB reading and mathematics test norms (CTB Macmillan/McGraw-Hill, 1991, 1993) showed the opposite: scale-score standard deviations decreased as grade levels increased (although Yen and Burket note that this trend is not consistent across all content areas). Figure 1 in Appendix A gives an example based on the norms of CTBS/4 total mathematics scores for Grades 1 through 3. Analyses of the Harcourt reading and mathematics test norms (The Psychological Corporation, 1992, 1993) indicates an intermediate result: scale-score standard deviations for the Harcourt tests tended to be nearly constant across grade levels. Figure 2 in Appendix A illustrates this for MAT 7 total mathematics scale scores at the same grades. The points of divergence mentioned above tend to occur at higher grade levels.

It is tempting to attribute this seemingly profound difference among the publishers to peculiarities of their respective scaling methods. Williams,

Pommerich, and Thissen (1998) state that Thurstone's scaling methods have historically found standard deviations to increase with age and grade, while the recently introduced IRT scaling methods have been inconsistent in this regard but have sometimes shown an opposite trend. Yen and Burket (1997) point out that correlations between grade and scale-score variation can vary by content area, ability distribution, and other factors; differences among the scaling methods themselves are not definitively responsible.

THREATS TO VERTICAL SCALING

A number of factors can adversely affect the accuracy of vertical scaling. Some of these are discussed here.

Score Precision Differences

When tests are equated, an inequitable situation can arise if the tests are unequal in precision (standard error of measurement). Score equating or calibration cannot compensate for these differences. In the case of vertical scaling, differences in the precision of the tests before rescaling can adversely affect score comparability under most equating methods (for example, equipercentile). This is unlikely to be a serious problem, however, unless one (or more) of the forms is too unreliable to be used by itself in reporting results for its target grade.

Multidimensionality

Multidimensionality of the item content domain can make scaling inaccurate. While this is particularly an issue in the case of IRT scaling—because IRT models such as the Rasch and the three-parameter logistic explicitly assume that item sets are unidimensional—it is no less an issue with some other equating and/or scaling approaches, such as Thurstone's 1925 method of absolute scaling. Low or moderate values of Coefficient Alpha provide an indication of possible dimensionality problems. DIMTEST, developed by William Stout, or factor analyses using TESTFACT can be used for further investigation of dimensionality.

Construct Differences

A fundamental assumption in vertical scaling is that the tests being scaled measure the same attributes or constructs. If they do not, the interpretability of the resulting developmental scale scores may be degraded such that identical scale scores obtained from tests at different levels may not carry the same meaning. This is particularly of concern when the content domain is subject to

differences in curriculum such as those between school grades, especially nonadjacent grades. Note that while construct differences may introduce multidimensionality, these two threats are not identical. For example, the content domains of adjacent grades in mathematics achievement may be multidimensional but may not differ as to the constructs measured. Conversely, the tests designed for the two grades may differ at the construct level, yet each grade's content domain may be unidimensional.

Exposure to Content

Data collection to support vertical scaling may involve administering out-of-level tests to some students (for example, the common design in which two levels are administered at each grade). If lower-grade students have not been exposed to all of the material covered in one or both test levels, but all the upper-grade students have been exposed to it, the scaling will be affected.

Practice and Fatigue Effects

Some data collection designs require each student to be administered two test forms. In such cases, practice effects may inflate scores on the second test taken, and fatigue effects may deflate scores to some extent. These effects, if they occur and are not controlled (by counterbalancing test order, for instance), will have an impact on the scaling results.

A PLAN FOR VERTICAL SCALING OF FCAT 2001

This section outlines a proposed plan for the vertical scaling of FCAT 2001 tests for Grades 3 through 10. The features of this plan are designed to take into account some constraints that we anticipate and, at the same time, reflect sound scaling practices.

Anticipated Constraints

We anticipate that FDOE will want to adhere to the items listed below. These will serve as constraints on the vertical scaling design and methods employed.

1. At each grade level, all multiple-choice test items are to be scaled using the three-parameter logistic (3-PL) IRT model. This has been FDOE's practice in the past, and we understand that the use of the 3-PL model for both item analysis and test scoring is to continue. We note that the use of the 3-PL IRT model for item analysis, test scoring, and year-to-year scale linking does not necessitate vertical scaling based on IRT methods, but we see no reason to

propose a different model here. The 3-PL model most accurately reflects observed relationships between ability level and the probability of passing each item.

2. All FCAT test administration for a grade/subject combination is to be conducted during a single test administration period in either February or March 2001. We assume there will not be an opportunity to administer tests at times other than the period designated for the FCAT 2001 assessment.
3. All students participating in FCAT 2001 are to be administered on-grade-level tests. We assume it will be unacceptable for some students' assessments to be based solely on out-of-level tests.
4. All students participating in the FCAT 2001 assessment are to be administered a single test form in each content area. We assume it will not be practical to administer both an assessment form and another form to any students. If this assumption is correct, the use of linking forms is precluded.
5. To facilitate creation of a vertically linked scale, it will be necessary to include a number of common items in test booklets for two or more successive grade levels. Not all of these anchor items need be included in the FCAT assessment test scores. Some of the anchor items will be included in the core section of the test booklet and used for operational scoring. Other anchor items will be inserted in special sections used for vertical scaling only.
6. There will be a separate core test level for every grade. No items will be used in operational scoring at more than one grade level.

If the above assumptions about practical constraints are correct, then the number of available approaches to data collection for purposes of FCAT vertical scaling is limited. In particular, certain approaches are eliminated because they are inconsistent with the constraints. Specifically, these constraints eliminate data collection designs of the following kinds:

- A. Those that require administration of a separate additional test to some students, such as a scaling test (as is administered to students of all grades for vertical scaling of the ITBS);
- B. Use of a full-length linking test (as is administered to some students at adjacent grades for vertical scaling of the CTBS and CAT);
- C. Use of a separate anchor test form (similar to but shorter than a full-length linking test) administered to adjacent grades.;

- D. Equivalent-groups designs of all kinds. Such designs entail administering two or more test levels per grade and would therefore not be feasible if our assumptions are correct.

Proposed Data Collection Design

We propose to use an embedded anchor test data collection design for vertical scaling of FCAT 2001 tests. For each subject area, there will be one set of anchor test items for each pair of adjacent grades (see Figure 3 in Appendix A.) The appropriate anchor test items will be embedded within FCAT 2001 assessment test booklets either as operational or as unscored vertical scaling items. Unscored vertical scaling item positions will be used for items that may be too easy or too difficult to contribute appropriately to the operational scores for a given grade.

Following are some of the issues to be addressed in defining the anchor for each pair of grades:

1. How many items should be in each anchor?
2. What range of item difficulties should be covered by the anchor?
3. How should the anchor items be distributed by content strand and item format?

Subject to further discussion, we propose the procedures described below.

Number of Anchor Items

Each anchor should have about twenty-four items to provide adequate stability and allow for appropriate coverage of test content and item formats. As discussed below, the items should span the combined content of the assessment at the two grades being linked but will necessarily exclude items that are too difficult for students in the lower grade. We propose to take sixteen of the operational items from the lower grade and eight of the operational items from the higher grade for a total of twenty-four items. The sixteen operational items from the lower grade will be administered as unscored test items in the higher grade, spread across three or more vertical scaling test forms. Similarly, the eight operational items from the higher grade will be administered as unscored test items at the lower grade, spread across two or more vertical scaling test forms. The use of special vertical scaling test booklets that include unscored scaling items in addition to core items is necessary because the items from one grade probably have not been tried out with students from the other grade and thus cannot be used operationally at that other grade level.

We recommend using twenty-four items for linking each pair of grades for two reasons. First, we believe that it will be possible to cover the content tested in

two adjacent grades with about this many items. Second, we believe it will take this many items to ensure adequate precision. Horizontal equating is sometimes performed with fewer items. (For FCAT 2000, fifteen to sixteen items were used at each grade level.) Because of differences in content and difficulty levels across grades, vertical scaling will require a greater number of items to achieve similar precision. While we recommend twenty-four, it would be reasonable to consider a range of anchor sizes, perhaps from twenty-one to twenty-eight, that reflect different trade-offs between cost/effort and precision. Flexibility may be particularly important where varying numbers of items per passage make it difficult to meet an exact target.

Item Difficulty

The range of difficulty covered by each anchor should run from the easiest item administered to the lower grade through at least the midpoint of the item difficulties for the higher grade (for example, sixteen lower-grade items that span the full range of difficulty for that grade and eight higher-grade items that span the lower half of the difficulty range for that grade). We believe that items from the upper half of the difficulty range for the higher grade could be too difficult for students in the lower grade to respond to meaningfully. Specifically, if nearly all students at the lower grade are performing at chance levels ($p = .25$ or less), the item will not provide much useful information. It might also be the case that the least difficult items from the lower grade do not provide meaningful information for the upper grade, but we have no specific reason to believe that this would be the case. We propose using the IRT difficulty (b) parameter as the primary indicator of item difficulty.

Content and Format Coverage

All of the content included in the standards for the lower grade should be covered by the anchor items. Care should be taken, however, in selecting items from the assessment for the upper grade to avoid areas of content that are clearly not covered in the lower grade. While it would be desirable to include all item formats in the anchor set, we understand the decision to exclude constructed-response (CR) items from the anchor item sets. The anchor items are not scored for one of the grades to which they are administered, and the costs of hand-scoring items would not be justified when the items are not used operationally. For mathematics, gridded-response (GR) items may be considered above Grade 6 to the extent that this format is used in both of the grades being linked.

The exclusion of CR items from the vertical scaling anchor item sets should not pose a serious problem so long as the correlation between scores from CR and multiple-choice (MC) items remains high. In the FCAT 2000 equating analyses, difficulty parameters from the CR items generally fell right on the regression line

defined by the b -parameters for the MC items so that the results would have been very similar if CR items had been excluded from that equating.

Example

The following example, using the Grade 5 mathematics assessment, summarizes the form design plan proposed above (see Figure 3 in Appendix A). Two different anchors must be included in the operational and unscored vertical scaling test items administered to Grade 5 students. The Grades 5-6 anchor will include sixteen Grade 5 items and eight Grade 6 items. Two types of special vertical scaling test booklets will be constructed, each having four of the Grade 6 items in addition to other items to fill in the positions for the unscored vertical scaling items. The remaining eighteen Grade 5 items in the grades 5-6 anchor will be included among the fifty operational (scored) items in the Grade 5 assessment. Three additional vertical scaling test booklets will be designed for Grade 5 to include items in the Grades 4-5 anchor. These booklets will each contain five or six Grade 4 items that are in the unscored item positions, for a total of sixteen items. The Grades 4-5 anchor will also include eight additional items that are administered as Grade 5 scored items.

In summary, twenty-four of the fifty scored Grade 5 mathematics items will be included in the grades 4-5 or grades 5-6 anchor. Of these, sixteen will also be Grade 4 scored items, and eight will be Grade 6 scored items. There will be five special cross-grade anchor booklets, three with additional Grades 4-5 anchor items, and two with additional Grades 5-6 anchor items. Note that the fifty operational Grade 5 items will also include fifteen to twenty items that were used operationally in prior years and form the anchor for year-to-year equating. Of course there can be overlap between the operational items used for grade-to-grade linking and the 1520 operational items used for year-to-year linking. A chart illustrating this design is attached (see Figure 3 in Appendix A).

Approach to Scaling

Computation

We propose to use an item-centered method that involves first performing a separate calibration of the items administered to each grade, including the operational items and linking items in the unscored item positions. A Stocking-Lord approach will then be used to identify a linear adjustment to the scale for the higher grade so that the transformed anchor item parameters for the higher grade closely match the parameters estimated from the lower grade. Finally, we will transform the combined scale to have the desired properties, so the direction of linking should not matter. (We could just as easily start with the scale for the

highest grade and link each lower grade through the next higher grade up to this scale. After transformation to a final target scale, the results would be the same.)

Final Scale

Once the scale for each grade is linked to the scale for the next lower grade, the resulting composite scale can then be mapped onto whatever scale FDOE would like to use. This mapping may be simply a linear transformation of the composite scale, but it could also be a nonlinear adjustment. A decision on the final reporting scale is not needed at this time, since the data collection and initial scaling can proceed independently from the final scale adjustments.

Given the desire to use the final scale in a value-added model, we believe that the final scale should be derived so that expected gains from each grade to the next are equal. For example, if the “expected” scores (either current average or the level required for proficiency) were 30, 40, and 50 for Grades 3, 4, and 5, respectively, the expected one-year gain would be 10 points for students at each grade level. Figure 4 in Appendix A provides an illustration of this scale. Note that the degree of overlap in scores for different grade levels must be determined empirically and should not be inferred from the illustration. The proportion of students at one grade who score above the mean for the next higher grade will be a key statistic in determining the degree of overlap.

Initial Use

Barring unforeseen results, the creation of the vertical scale should not require extensive analysis time. It remains to be seen, however, whether an additional two weeks or so could be inserted in the FCAT 2001 schedule so the results from that assessment could be reported directly on the new scale. It may be more prudent to use grade-specific scales for the initial reporting of FCAT 2001 results (retaining the existing scales for the grades tested previously) and introduce conversions to the new scale in fall 2001.

Follow-On Studies

It would also be prudent to plan follow-on studies to check the stability of the vertical scale over time. Changes in the curriculum at specific grades and other modifications might lead to different results in later replications. The pace of such change should be relatively slow, so replication every two or three years until stability over time is demonstrated should be sufficient.

Another type of follow-on study would involve the use of expert panels to review item data and develop anchor descriptions of what students at different points on

the scale know and can do. Such anchors could help align curriculum to content standards and have diagnostic uses.

Protection against Threats

Earlier in this report we listed several threats to the accuracy of vertical scales. The discussion that follows summarizes the steps taken in the proposed plan to minimize these threats.

Score Precision Differences

Each of the forms for different grades being linked will be of sufficient precision to support operational reporting. Precision differences should not be a threat to the proposed design.

Multidimensionality

The proposed number of linking items for each grade should be sufficient to protect the linkages from minor departures from unidimensionality (for instance, local item dependence for reading items linked to a common passage). Scaling results for the current operational grades do not indicate any problems in fitting IRT models to the content of those grades, even when different item formats are included.

Construct Differences

Content differences across different grades may continue to pose a problem for the creation of the vertical scale. The proposed approach links each pair of successive grades separately, reducing problems associated with differences across wider grade ranges. Within each anchor item set, it will be possible to examine the correlation of the sixteen items from one grade with the eight items from the next higher grade. If that correlation is low relative to the reliability of scores from these item sets, a more extensive analysis of construct differences will be undertaken.

Exposure to Content

Care will be taken in selecting anchor items to avoid items from the higher grade that are not covered in the curriculum of the lower grade.

Practice and Fatigue Effects

The proposed design does not include overlength forms. Each student completes the same number of items as usual, so there is no reason to expect unusual practice or fatigue effects.

ANNOTATED BIBLIOGRAPHY

Angoff, W. H. (1984). *Scales, Norms, and Equivalent Scores*. Princeton, NJ: Educational Testing Service.

Standard reference from first publication in 1971 through the present.

Bolt, D. M. (1999). “Evaluating the Effects of Multidimensionality in IRT True-score Equating.” *Applied Measurement in Education* 12(4), 383-408.

This is a simulation study that compares IRT true-score equating against conventional linear and equipercentile equating in a set of data in which degrees of multidimensionality are introduced. The data are based on LSAT data, and the degree of multidimensionality is expressed in terms of correlations between dimensions: “Results from 2 simulation studies suggest the IRT method performs at least as well as the conventional methods when the correlation between dimensions is high (≥ 0.7) and may be only slightly inferior to the equipercentile method when the correlation is moderate to low (≤ 0.5).” Note that this study does not involve vertical scaling, just the effect of multidimensionality on equating.

Camilli, G. (1999). “Measurement Error, Multidimensionality, and Scale Shrinkage: a Reply to Yen and Burket.” *Journal of Educational Measurement* 36(1), 73-78.

CTB (1991). *CTBS/4 Technical Report*. Monterey, CA: Author.

CTB Macmillan/McGraw-Hill (1992). *CAT 5 Technical Bulletin 1*. Monterey, CA: Author.

CTB Macmillan/McGraw-Hill (1993). *CAT 5 Technical Bulletin 2*. Monterey, CA: Author.

DeChamplain, A. F. (1996). “The Effect of Multidimensionality on IRT True-Score Equating for Subgroups of Examinees.” *Journal of Educational Measurement*, 33(2) 181-201.

This article reports a comparison of IRT true-score equating results based on data from separate ethnic groups (Caucasian, African-American, and Latin-American) in which the degree of multidimensionality had been found to vary. These are LSAT data, in which two dimensions were found for Caucasian and African-Americans, but three dimensions for Latin-Americans. “The differences between

equating functions for the 3 ethnic groups and the total test-taker population were small and tended to be located at the low end of the score scale.”

Dorans, N. (1990). “Equating Methods and Sampling Designs.” *Applied Measurement in Education* 3, 3-17.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Kolen, M. J. and Brennan, R. L. (1995). *Test Equating Methods and Practices*. New York: Springer.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1982). “Item Response Theory and Equating--A Technical Summary.” In *Test Equating*, edited by P. W. Holland and D. B. Rubin. New York: Academic Press.

Peterson, N. S., Kolen, M. J., and Hoover, H. D. (1989). “Scaling, Norming, and Equating.” In *Educational Measurement*, edited by Linn, R. L.. New York: American Council on Education and Macmillan.

Extends and updates Angoff’s classic chapter from the first edition.

Riverside Publishing Company (1994). *Riverside 2000 Technical Summary I*. Chicago: Author.

Schulz, E. M. and Nicewander, A. (1997). “Grade-equivalent and IRT Representations of Growth.” *Journal of Educational Measurement* 34(4), 315-332.

The authors note that developmental scales based on age or grade historically showed increasing variance of test scores with increasing age, but that recently developmental scales based on IRT have shown scale-score variance to be constant or decrease with increasing age. They report a study, using IRT scaling methods, in which they find the variance of grade-equivalent scores increased markedly with grade. The study was based on assumptions of 1) negatively accelerated growth in the mean of a latent educational achievement variable; and (2) constant within-grade variance.

The Psychological Corporation (1992). *Stanford Achievement Test Series, Eighth Edition, Multilevel 1991 National Norms Booklet*. San Antonio, TX: Harcourt Brace Jovanovich.

The Psychological Corporation (1993). *MAT 7 Multilevel Norms Book (Spring)*. San Antonio, TX: Harcourt Brace.

Williams, V. S. L., Pommerich, M., and Thissen, D. (1998). “A Comparison of Developmental Scales Based on Thurstone Methods and Item Response Theory.” *Journal of Educational Measurement* 35(2), 93-107.

This study reports on the creation of a developmental scale for the North Carolina End-of-Grade Mathematics Tests. Identical test forms were administered to adjacent grades. Three variations of Thurstone’s scaling were tried; IRT scaling was based on BiMain (BILOG-MG predecessor) and Multilog. Mean scale scores increased in Grades 3 through 8 for all methods. Results of both IRT methods were highly similar to those of the Thurstone 1938 procedure with respect to means. The pattern of standard deviations across grades was inconsistent for the IRT methods, but generally decreased under the Thurstone 1925 procedure based on transformed p -values, increased slightly with one version of his 1938 procedure which is based on transformed percentile ranks; and was inconsistent with the third Thurstone procedure, a variant of the 1938 method. Some leading practitioners have found a concise but informative summary of differences between Thurstone and IRT scaling : “While applications of Thurstone scaling generally display scale expansion across age groups or grades, IRT studies have shown variances to be decreasing, homogeneous, or increasing . . . apparent growth trends vary across scaling procedures . . .”

Wise, L. L., Hoffman, R. G., and Thacker, A. A. (1999). *Evaluation of Calibration and Equating Procedures for the Florida State Assessment*. HumRRO Report FR-WATSD-99-41. Prepared for Harcourt Educational Measurement. Alexandria, VA: HumRRO.

Yen, W. (1983). “Use of the Three-Parameter Model in the Development of a Standardized Achievement Test.” chapter 8 in *Applications of Item Response Theory*, edited by Hambleton, R. K. Vancouver, BC: Educational Research Institute of British Columbia.

This chapter summarizes CTB’s use of the 3-PL model in test development, including vertical scaling.

Yen, W. M., and Burket, G. R. (1997). “Comparison of Item Response Theory and Thurstone Methods of Vertical Scaling.” *Journal of Educational Measurement* 34(4), 293-314.

This simulation study compares Thurstone’s absolute scaling with the 3-PL IRT approach recently introduced into CTB practice. The data are constructed to be highly similar to CTB’s own data from several tests. The study is especially attentive to scale shrinkage, reported by others in simulation studies of IRT equating where the data were less realistic. This study found no shrinkage under

the 3-PL model. For one data set, they found some scale expansion for both the 3-PL and Thurstone's methods.

APPENDIX A: FIGURES

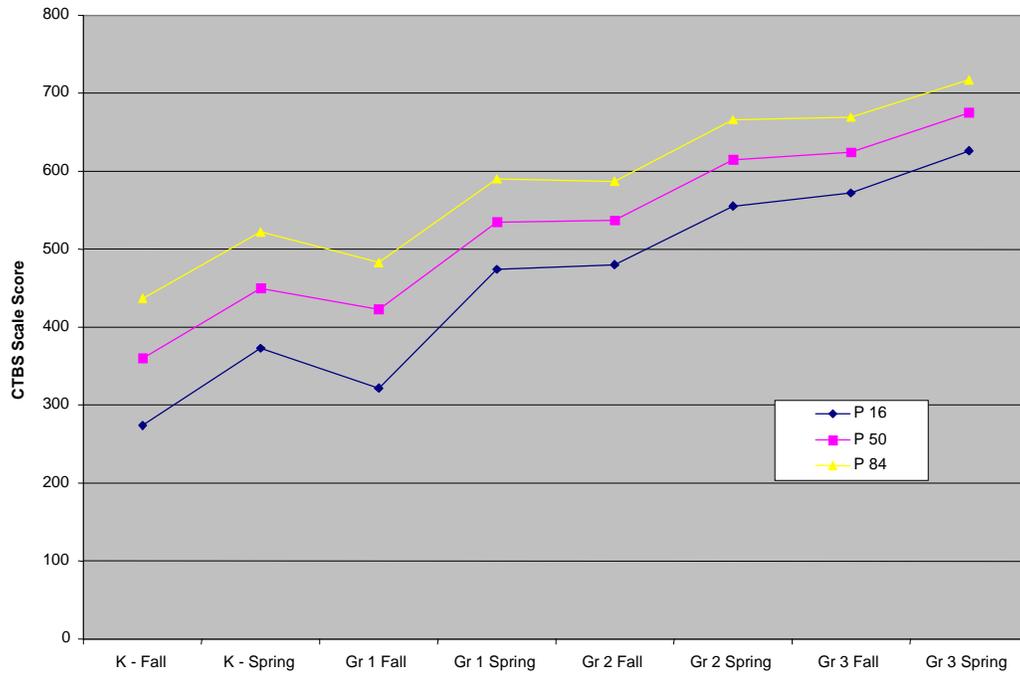


Figure 1. CTB 4 Total Mathematics Norms Data, Showing Scale Scores for P16, P50, and P84 for Grades 1 through 3

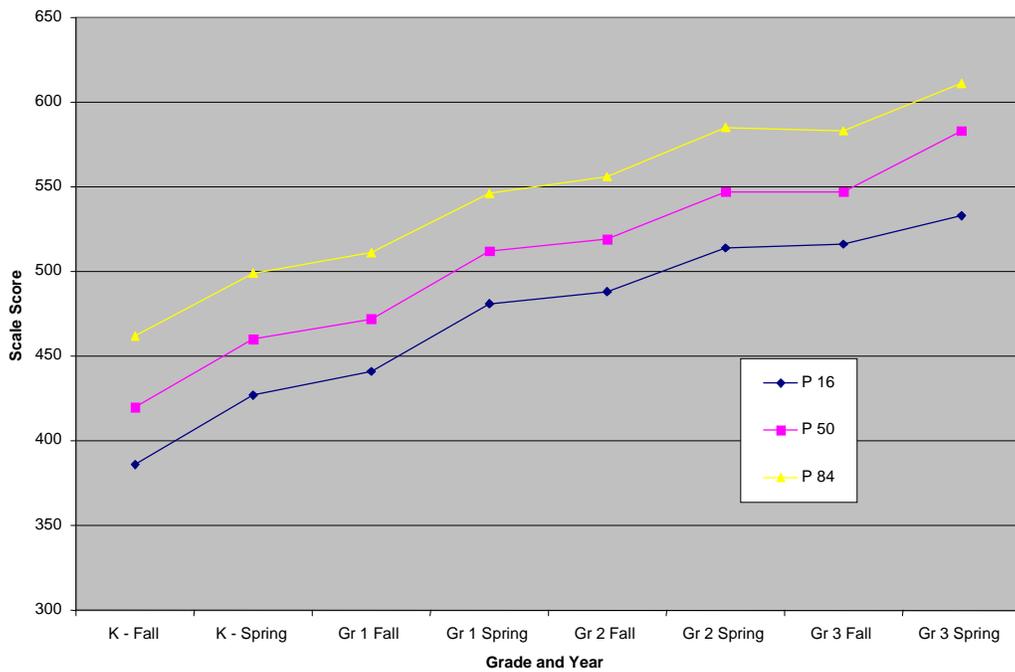


Figure 2. MAT 7 Total Math Norms Data, Showing Scale Scores for P16, P50 and P84 for Grades 1 through 3

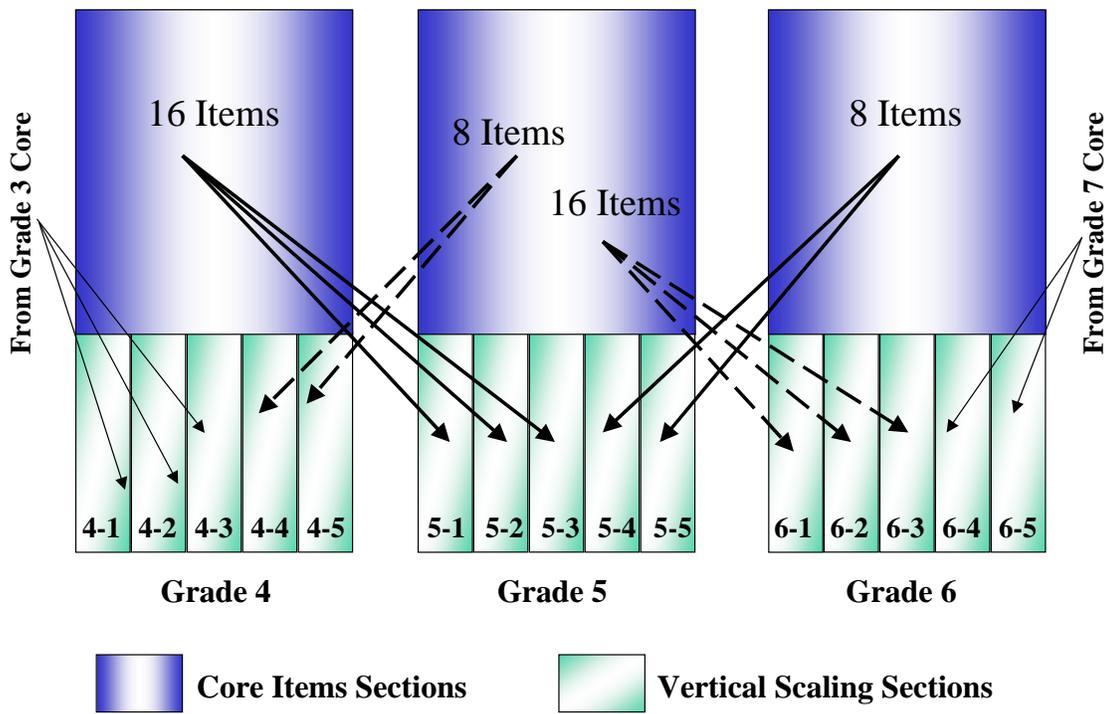


Figure 3. Illustration of Vertical Scaling Forms Design

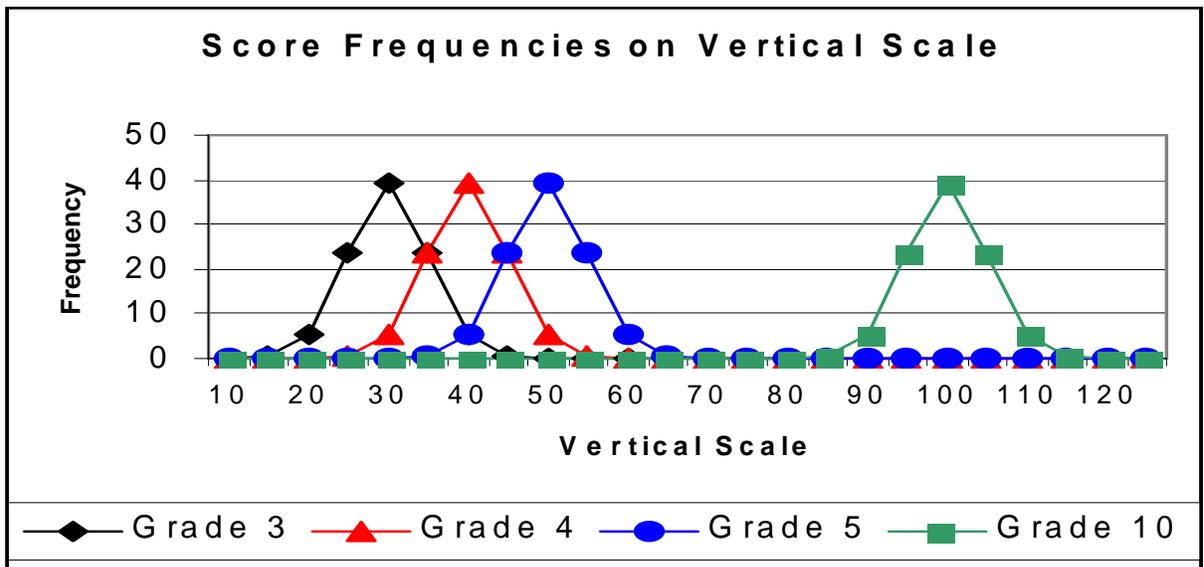


Figure 4. Illustration of Score Distributions on the Vertical Scale

APPENDIX B: GLOSSARY

absolute scaling – a method of nonequivalent-groups test equating based on the classical item statistics of anchor items embedded in the tests to be equated. In absolute scaling, each item's scale values in one group are transformed to equivalent scale values in the other group. Two tests' anchor item p-values are first transformed to normal deviates (z) within each test; the linear relationship between the two sets of anchor item z statistics is used as the basis for equating.

anchor items – items included in two or more forms for purposes of linking scores on those forms.

a-parameter – the slope parameter in the mathematical function (ICC) giving the probability of the correct answer for different levels of ability.

b-parameter – the difficulty parameter in the mathematical function (ICC) giving the probability of the correct answer for different levels of ability. This parameter is equal to the ability level at which the probability of knowing the correct answer is .5.

c-parameter – the guessing parameter in the mathematical function (ICC) giving the probability of a correct answer for different levels of ability. This parameter is equal to the probability that an examinee who does not know the answer will, nonetheless, answer the item correctly.

calibration – in applications of *item response theory*, the process of estimating item parameters. This process “calibrates” the items to a particular ability scale by specifying the location, slope, and lower asymptote of a mathematical function that relates the probability of a correct answer to examinees' abilities.

classical test theory – the traditional approach to item analysis that emphasizes item difficulty and item-total correlation (See Lord and Novick, 1968). This term is normally used to distinguish it from item response theory (IRT).

coefficient alpha – a formula for estimating test score reliability from the correlations among the items in a test form (internal consistency).

common items – a set of items common to more than one (usually all) forms used in a given administration. Typically these items are used as anchor items to put scores from each form onto a common scale.

construct – the trait or ability being measured by a test.

constructed-response (CR) – a response format in which the examinee must write in an answer as opposed to checking or bubbling one of a list of possible answers. Responses may range from a single word to an entire essay.

correlation – a statistic developed by Karl Pearson summarizing the relationship between two variables. Absolute values range from 1.0, when one variable can be perfectly predicted by the other, to 0.0, when there is no relationship at all. A value of -1.0 indicates a perfect negative relationship.

counterbalancing – altering the order in which two or more test forms are administered. Counterbalancing is used when the test forms to be equated will be administered to a single examinee group. Its purpose is to evaluate and counteract "order effects" such as learning, practice, and fatigue that might otherwise bias equating results.

developmental scale – a scale indicating stages of progress in acquiring understanding or skill.

dimensionality – the number of different things (dimensions) measured by a test. For example, arithmetic word problems could measure both reading ability and numerical computations (two dimensions).

equating – the process of putting scores for different forms onto a common scale.

equipercentile – an approach to equating used when different forms are administered to the same or equivalent groups. The number correct score corresponding to a given percentile for the first form is equated to the number correct score in the second form that corresponds to the same percentile score.

equivalent groups – two or more groups that are equivalent should have the same distribution of the ability being measured. Usually this refers to a process in which examinees are randomly assigned to groups so there is no reason for the groups to differ by more than random sampling error.

examinee ability (ability) – the construct measured by ability tests.

extrapolation – using data to estimate values outside the range of the data.

factor analysis – a statistical method for inferring the number of dimensions measured by a set of test items and for relating scores on different items (or other variables) to abilities on different underlying dimensions.

grade-equivalent – a scale on which the mean or median for each grade is the grade itself.

horizontal scale – a scale used to compare different students at the same level (for example, grade).

interim scale – a scale used to estimate ability or item difficulty that has not yet been linked to the standard reporting scale.

item difficulty – the proportion of examinees in a given population who answer an item correctly. For multiple-choice items, guessing inaccurately enlarges the proportion who earn credit on the item beyond the proportion who actually know the correct answer.

item-response theory (IRT) – an approach to item analysis based on using specific forms of the function relating knowing or answering correctly to the construct being measured (see Lord and Novick, 1968).

linear – an approach to equating that involves multiplying all original values by one constant (coefficient) and adding a second constant (intercept).

linear adjustment – an adjustment to score estimates that is linear.

mean – the average value in a range of numbers.

median – the middle value in a range of numbers. Half of the values are greater than the median and half are below.

monotone – a functional relationship that preserves the order (greater than) function, but is not necessarily linear. Thus all values larger than a given value on the original scale will be mapped onto new scale values that are greater than the new scale value corresponding to the given value.

multiple-choice (MC) – an item format in which a set of possible answers are presented and the examinee selects one of them.

nonequivalent groups – groups in which it is not reasonable to expect that abilities will be the same. For example, since this year's students may have learned more than last year's students, it is not reasonable to suppose that the achievement of the two groups would necessarily be the same.

nonlinear adjustment – a scale adjustment that is not linear (see *linear*). Usually it is at least monotonic (see *monotone*).

normal curve equivalent – a way of scaling scores from an arbitrary distribution by transforming the original scores to percentiles (percent of examinees scoring at the same level or lower) and then finding the value of the

cumulative normal distribution (the normal curve, or bell curve) that has the same percentile.

observed score – a score that is directly observable, usually the number correct. Used in contrast to "true score," an unobservable construct.

p (difficulty; item difficulty) – a classical test theory parameter defined as the proportion of examinees in a specified population who answer an item correctly.

percentile - the percentage of examinees who score below a given score level.

randomly equivalent groups – a common way of making groups equivalent through the random assignment of examinees to groups. Most commonly the groups are defined by specific test forms.

raw score – an untransformed, number correct score; see also *observed score*.

regression (regression line) – regression is the process of estimating a regression line that gives a prediction of one variable through a linear (straight-line) function of the other variable or variables.

reliability – an estimate of score consistency across forms or testing occasions. In classical test theory the variation in scores across testing occasions or equated forms is called error. Reliability is the proportion of the total score variation that is not error.

scale – a set of numbers assigned to different test outcomes.

scale transformation (transformation) – the process of translating numbers on one scale into a different set of numbers.

scaling – the process of putting items or examinees on a given scale. See also *item calibration*.

scatter plot – a display in which each entity is represented by a single point in a plane. The coordinates of the point are defined by the entities' values on two different measures.

standard deviation – the square root of the variance.

standard error of measurement – the standard deviation of measurement errors.

standard normal deviate – see *z-value*.

Stocking-Lord approach - one approach to putting results from two forms on a common scale by relating separate estimates of IRT parameters for a common set of items.

test characteristic curve – a function giving the expected total score as a function of ability.

test equating (see *equating*).

test form – a set of items administered together to a group of examinees.

true score – the expected or average number correct score over an infinite number of repeated administrations of the same or parallel forms. True scores are usually not observed.

variance – the average of the squared differences between the values of a variable and its mean.

vertical scale – a scale used to compare students from different grades.

z-value – a standard transformation of individual values, typically raw scores, usually generated by subtracting the mean and dividing by the standard deviation. The resulting scores always have a mean of 0 and a standard deviation of 1.