LEADERSHIP, ADVOCACY, SERVICE

CCSSO

THE COUNCIL OF CHIEF STATE
SCHOOL OFFICERS

Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems

January 2007

Richard J. Patz

The Council of Chief State School Officers

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

State Collaborative on Assessment and Student Standards

The State Collaborate on Assessment and Student Standards (SCASS) Project was created in 1991 to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. The Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects. SCASS projects accomplish a wide variety of tasks identified by each of the groups including examining the needs and issues surrounding the area(s) of focus, determining the products and goals of the project, developing assessment materials and professional development materials on assessment, summarizing current research, analyzing best practice, examining technical issues, and/or providing guidance on federal legislation.

Technical Issues in Large-Scale Assessment (TILSA)

TILSA is part of the State Collaborative on Assessment and Student Standards (SCASS) project whose mission is to provide leadership, advocacy, and service by focusing on critical research in the design, development, and implementation of standards-based assessment systems that measure the achievement of all students. TILSA addresses state needs for technical information about large-scale assessment by providing structured opportunities for members to share expertise, learn from each other, and network on technical issues related to valid, reliable, and fair assessment; designing and carrying out research that reflects common needs across states; arranging presentations by experts in the field of educational measurement on current issues affecting the implementation and development of state programs; and developing handbooks and guidelines for implementing various aspects of standards-based assessment systems.

CCSSO Staff and Consultants
Don Long, SCASS Director
Phoebe Winter, TILSA Coordinator
Arthur Halbrook, Editor

Council of Chief State School Officers
State Collaborative on Assessment and Student Standards
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000, Fax (202) 408-8072
www.ccsso.org

Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems

Prepared for the Technical Issues in Large Scale Assessment (TILSA)

State Collaborative on Assessment and Student Standards (SCASS)

of the Council of Chief State School Officers (CCSSO)

Richard J. Patz
CTB/McGraw-Hill

Acknowledgments

This report was written in collaboration with the Technical Issues in Large Scale Assessment (TILSA) Subcommittee on Vertical Scaling and Growth Models, under the leadership of Joseph Saunders. The Committee provided very valuable feedback that guided the development and revision of the paper. Howard Mitzel played a critical role in advancing this and supporting its goals. Reviewers Dale Carlson, Huynh Huynh, and Wendy Yen provided valuable feedback on an earlier draft of this paper. Richard Patz, who conducted this work as president of R. J. Patz, Inc., is now Vice President of Research at CTB/McGraw-Hill and may be reached at richard_patz@ctb.com.

Abstract

This paper describes vertical scaling as a method for linking a set of test forms of increasing difficulty, and explores the applicability of these methods for standards-based educational achievement tests under status-based and growth-based accountability priorities. The purpose of the paper is to inform state policy-makers and assessment and accountability specialists about vertical scaling methods, and the potential advantages and limitations of vertical scaling in standards-based educational accountability programs. The possible application of vertical scales to support growth models is discussed, as well as alternatives to vertical scaling that meet accountability system needs.

Contents

Introduction 1

Psychometric Tools for Vertical Scaling and the Measurement of Achievement and Growth 2

Measurement Models and Scaling Procedures 2

Linking and Equating 6

Standard Setting 8

Appropriate Contexts for Vertical Scales 11

Data Collection Designs 12

Advantages of Vertical Scales 15

Limitations of Vertical Scales 16

Characteristics of Well Constructed Vertical Scales 17

Some Recent Developments in Vertical Scaling Technology 18

Policy Context: NCLB and Growth Model Pilot Studies 19

Discussion 21

References 23

Introduction

Educational measurement practitioners commonly violate the basic principle: “When measuring change, don’t change the measure.” Unlike instruments for measuring physical characteristics (e.g., height or weight), our instruments for measuring status and growth in educational achievement must change in order to preserve the validity of the measurements themselves. We test our schools year after year and ask if this year’s students as a whole or in a specific grade are performing better than last year’s students or those students of two years ago. In order to support the relatively high stakes associated with the answers to these questions, states work diligently to ensure that this year’s tests are different (for security’s sake) but also comparable (for fairness’ sake) to the tests administered last year and the year before. Test forms for a given grade level are built to stringent specifications and statistically equated to ensure comparability so that our changed measures do not contaminate our measures of change.

Over the years, however, and heretofore largely apart from any accountability concerns, we have been interested in the change that occurs to individual students or groups of students as they progress across grades through our schools. How much more proficiently can students read now in grade 5 than they could three years ago in grade 2? To measure this type of change our measures must also change, here because unlike physical scales, our instruments for measuring achievement provide reliable information only if they are tailored to a somewhat homogeneous group of students (e.g., students in grade 5).

The creation of a set of tailored instruments that measure educational achievement for groups of systematically different abilities is the goal of vertical scaling. This paper describes vertical scaling methods and helps examine the question of whether the resulting changed measures can measure changes in individuals over years with enough accuracy to support the relatively high stakes that accountability testing necessarily entails. We explain what vertical scales are, how they are constructed, and what their advantages and limitations are.

We begin in the next section by describing vertical scales in their psychometric context so that their basic technical features can be understood. Then we look at the contexts in which vertical scaling may be fruitfully applied, as well as the advantages and limitations that vertical scaling may bring. We examine key accountability features of the *No Child Left Behind* (NCLB) *Act* and discuss the specific ways that vertical scaling may be applied in NCLB programs and their emerging “growth model” pilot studies. We conclude by making some recommendations on how states might evaluate vertical scaling in light of their priorities and the possible alternatives to vertical scaling.

Psychometric Tools for Vertical Scaling and the Measurement of Achievement and Growth

Vertical scales result from the application of a number of psychometric procedures, and understanding vertical scales, their usefulness and limitations, requires an understanding of several of these key building blocks. In particular, the fundamental components that we review for this vertical scaling discussion include measurement models, scaling, equating, linking, and standard setting.

Measurement Models and Scaling Procedures

Educational assessments are created to report on what students know and can do. Although any test consists of a specific set of test items, the information we seek about student proficiency is general and the specific test items are intended to support conclusions about more broadly defined areas of student proficiency. The proficiency we assess in a subject area may be thought of as not directly observable and measurable; instead, it may be imperfectly measured by observing student responses to test items.

When we say that we wish or require that a student or all students be proficient in reading, for example, defining what is meant by proficiency is of great importance. In modern educational measurement theory and practice, proficiency is considered to be a psychological construct, operationally defined as the unobserved ("latent") variable that explains individual differences in performance on the observable set of measures (e.g., reading test items). Establishing the relevance of the observable measures for supporting inferences regarding the psychological construct is a primary goal of test validation.

The latent proficiency variable is defined by a specific statistical model (e.g., item response theory model or classical test theory model) that describes how observable responses depend on those proficiencies. Because these statistical models enable us to estimate and assign numerical values for proficiency given an examinee's responses to test items, they are appropriately called measurement models. Scaling refers to the establishment of units for reporting measures of proficiency, and scaling occurs in conjunction with the identification of measurement models.

Measurement models and cross-grade scaling techniques have a long history that pre-dates the development of many methods currently employed in large scale assessment programs. For example, editions of the California Achievement Tests published before 1980 used Thurstone (1928) scaling, as described in Gulliksen (1950, pp. 284–286). Under this version of Thurstone scaling, raw scores (i.e., number correct scores) for equivalent groups of examinees are normalized and linearly equated. The Iowa Tests of Basic Skills have been vertically scaled using a "scaling test" design (discussed with other data collection designs later in this paper), wherein a special test composed of items selected from each test level is administered to special samples of students. Within- and between-grade variability and growth is determined on the basis of the scaling test results, and grade level specific forms are linearly equated to the scale (Hieronymous, Lindquist, & Hoover, 1982).

Currently, item response theory (IRT) serves as the basis for the most commonly employed measurement models used in state K-12 assessment programs. IRT models define a specific statistical relationship between examinees and test items (Lord, 1980, Hambleton & Swaminathan, 1985), and the IRT framework has proven to be a powerful technology for test construction, scaling, equating, and score reporting.

For multiple-choice items, one popular IRT measurement model is the three-parameter logistic (3PL) model (Birnbaum, 1968). It describes how the probability of a correct response on a test item depends on examinee proficiency and three characteristics of the test item: difficulty of the item, ability of the item to discriminate between high and low proficiencies, and the probability of a correct response by a very low proficiency examinee. Figure 1 depicts this relationship (the item characteristic curve) for a multiple choice item under the 3PL model. For multiple-choice items, the third parameter (referred to as a pseudo-guessing parameter) may be expected to be close to the reciprocal of the number of answer options (e.g., ¼ for a multiple-choice item with four response alternatives).

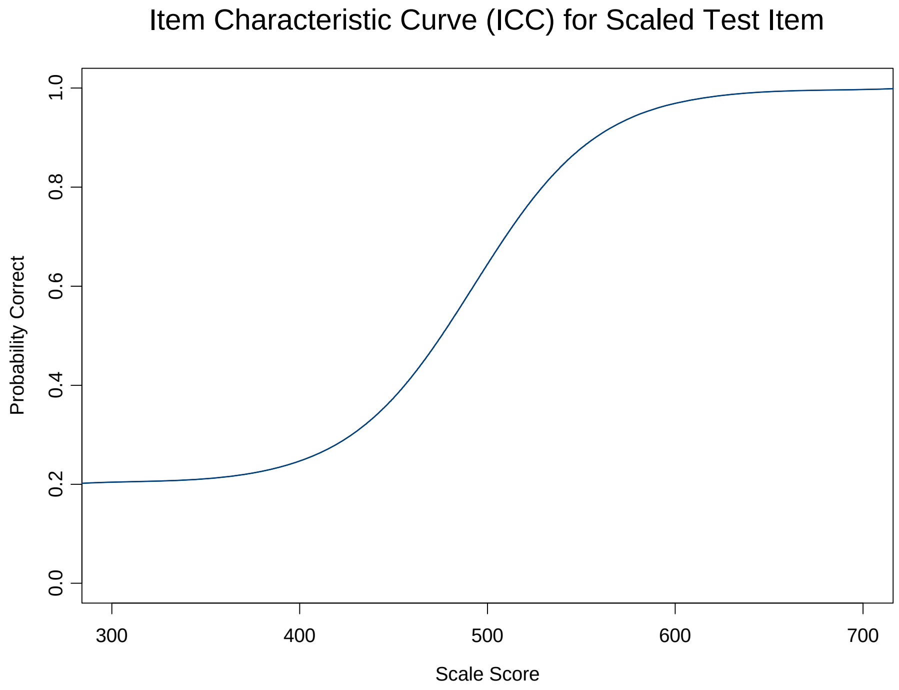

Figure 1. Item characteristic curve describes how the probability of success on a test item depends on examinee scale score.

Dichotomously scored constructed-response items may be modeled with a special case of the 3PL model (2PL model) that constrains the guessing parameter to be zero. If we also constrain the discrimination parameters to be equal, then we have the one-parameter logistic model developed by Rasch (1960). Both the 2PL and Rasch models have extensions that accommodate partial credit scoring of constructed-response items. Vertical scales have been created using each of these models. It is important to examine the appropriateness of any selected model(s), including statistical measures of model-data fit.

Estimating item parameters for each item on a test forms allows us to examine how examinees at any point along the proficiency scale might perform on each item. Figure 2 illustrates a set of item characteristic curves for a test form. For an examinee at any point along the proficiency range, we can determine items that the examinee is likely to answer correctly or incorrectly with high probability. Note that the small number of the items (10), the similarity of the ICC shapes for items, and the regular spacing between ICCs in Figure 2 are used for conceptual clarity here; sets of ICCs for typical achievement tests will typically be less regular. In particular, under the 3PL model, ICCs may have different slopes and different lower asymptotes, whereas we illustrate concepts with common or only slightly varying slopes and asymptotes.

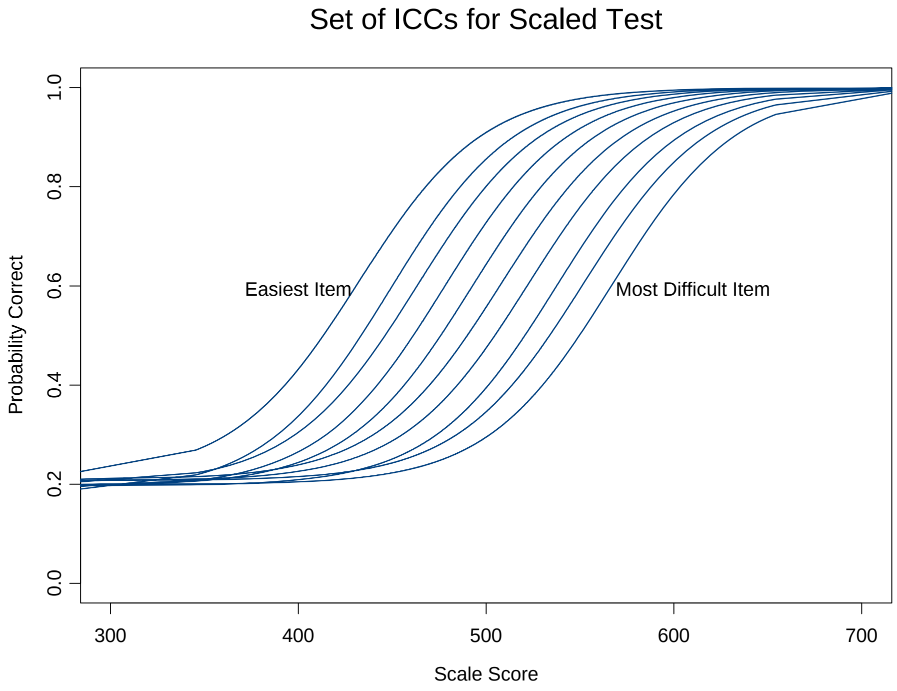

Figure 2. Item characteristic curves for a set of test items indicate which items will be answered correctly with high or low probability for examinees with a range of scale scores.

The set of item characteristic curves may be averaged to determine a test characteristic curve (TCC), and a TCC describes the relationship between proficiency and the expected number or proportion of correct responses on the entire test. Figure 3 illustrates a test characteristic curve.

It is important to note that all measurement models, but particularly item response theory measurement models, make strong assumptions, and the validity of the conclusions drawn from fitting IRT models depends on the accuracy of the assumptions. For example, although a student might be predicted to have a very high or low probability of success on an item according to the IRT model, the student may well perform differently than expected, perhaps because the student's ability to perform on the item depends in reality on important factors that are not captured in the IRT's simple, one-dimensional characterization of proficiency. Test developers explicitly examine test data for instances where external factors such as gender or ethnicity inappropriately influence student success on items (e.g., through differential item functioning [DIF] analyses), but additional factors (e.g., reading ability influencing success on a math test) are known to affect success and complicate our interpretations of achievement.

Measurement models provide a systematic way to associate numerical values (proficiency estimates) based on observable responses. However, the units to be used in reporting proficiency are not determined by the measurement model. This is established in the process of scaling. In a simple case, scaling may consist of identifying an arbitrary linear transformation of measurement model latent variable estimates in order to achieve desirable statistical properties (e.g., mean, standard deviation, or range) for the distribution of scaled proficiency estimates in a particular population of examinees. In the more complicated case of vertical scaling, measurement model latent variable estimates derived from a set of test forms of increasing difficulty must be placed in appropriate relation so comparisons may be made for examinees taking forms of different difficulty. Vertical scaling requires test forms to be linked, and we discuss methods for linking in the next section.

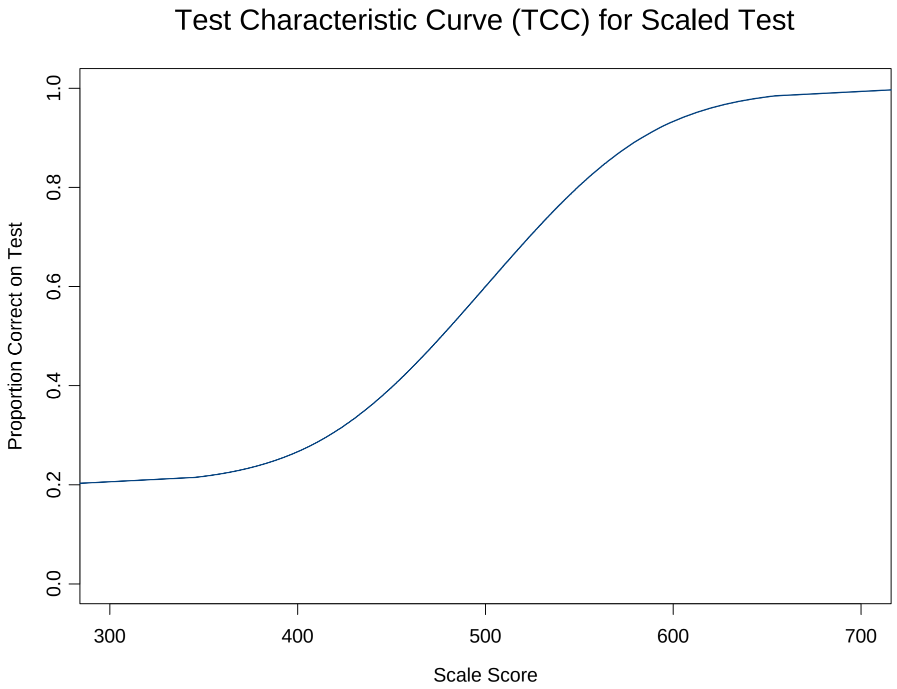

Figure 3. The test characteristic curve establishes the relationship between the percent (or number) correct on a test and the scale score. This information may be the basis for number correct to scale score conversion tables.

Linking and Equating

Creating vertical scales involves linking test forms to a common scale. Formally equating forms, which results in interchangeable test scores, requires that forms be parallel in their content and technical characteristics. Forms that are not parallel in structure but measure a common proficiency may be linked (e.g., see Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Overlap in content standards at adjacent grades may support the proposition that forms for adjacent grades measure a common construct, but differences in the standards and psychometric properties of the test forms (e.g., test difficulty) mean that these forms are not parallel and so they may be linked but not equated. Vertical scaling of achievement test batteries is most commonly established through a series of linking procedures. Since the links used in this case relate forms of intentionally different difficulties, they are referred to as vertical links, and the resulting score scale is called a vertical scale. This is in contrast to the linking of test forms of equivalent difficulty, which can be called horizontal linking or equating.

There are a number of technically sound approaches to establishing equating and linking relationships. Some require that a common set of examinees or statistically equivalent groups of examinees take the two forms to be linked or equated. Alternative approaches require that the forms to be linked or equated contain a set of test items in common (see Kolen & Brennan [2004] for a thorough introduction to and review of equating and linking procedures).

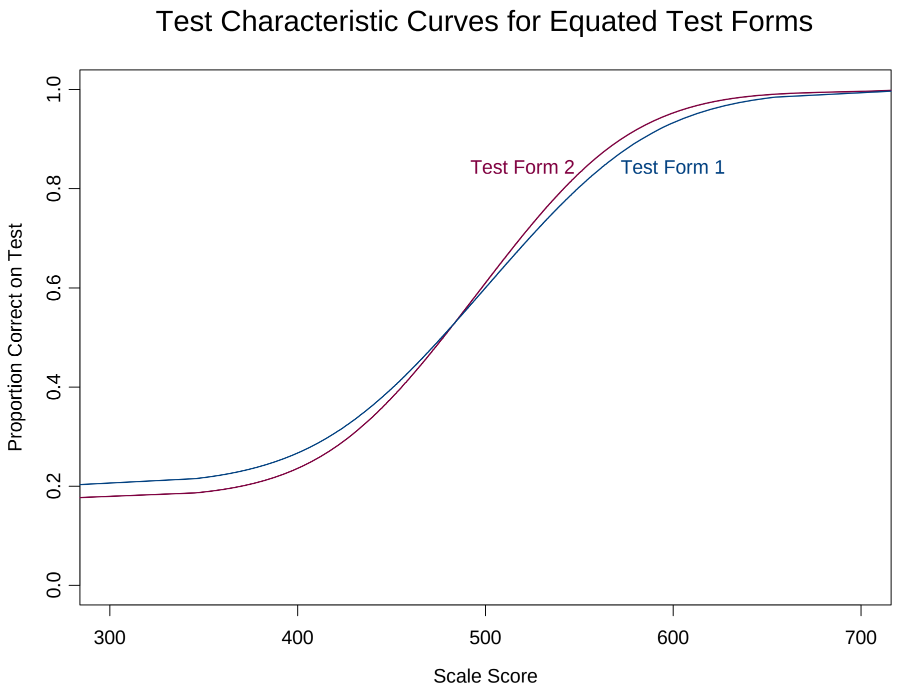

Figure 4. Parallel test forms may be equated, resulting in similar test characteristic curves. Differences in TCCs represent the equating adjustments required by unintentional differences in test difficulty and/or discrimination. Perfectly parallel test forms would have identical TCCs and scoring tables.

When two test forms are equated or linked to a common scale, their test characteristic curves may be depicted in relation to the scale, and differences in the technical characteristics between the forms may be examined. In the case of equating, the differences between TCCs, as illustrated in Figure 4, reflect the fact that these forms are not perfectly parallel. In test construction, efforts are made to minimize TCC differences, but such differences occur commonly and are adjusted for in the equating process.

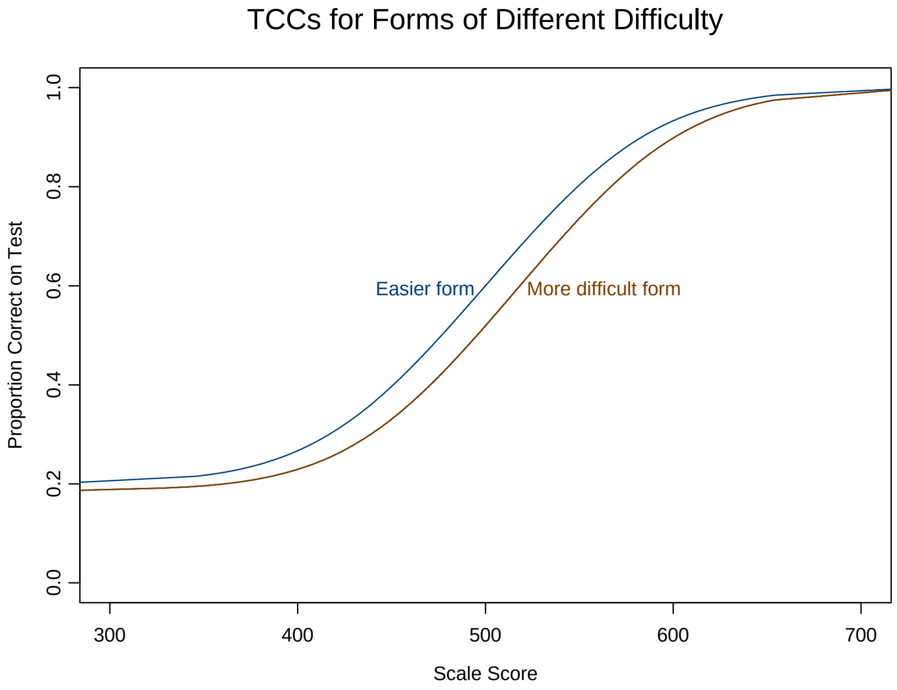

Figure 5. Vertically-linked test forms have systematically different test characteristic curves by design.

By contrast, when linking vertical forms, the TCC differences are expected, as they are created intentionally by making one test form more difficult than the next. Figure 5 depicts the TCC relationships between vertically linked forms. We can see, for example, that any given scale score is associated with a higher proportion correct on the easier form and a lower proportion correct on the more difficult form.

When we create a series of vertically-scaled tests, as may be done for NCLB reading or math tests for grades 3-8, the result of vertical linking is a series of TCCs that progress systematically across the scale score range. Figure 6 illustrates a series of test characteristic curves for a set of vertically-scaled test forms. Depending on a number of factors, including the choice of IRT model, such a set of TCCs for a real test program might appear different than those displayed in Figure 6. For example, the range of lower asymptotes might be greater and the slopes of the curves more variable under a three-parameter logistic measurement model, or the lower asymptotes might be zero and all slopes equal under the Rasch model. We note in passing that the creation of forms of intentionally different difficulty also occurs in computer adaptive testing (CAT), where the form is constructed by selecting items on the fly to match the proficiency level of the examinee. Comparability of CAT administrations rests on some of the same IRT methodology used for vertical scaling.

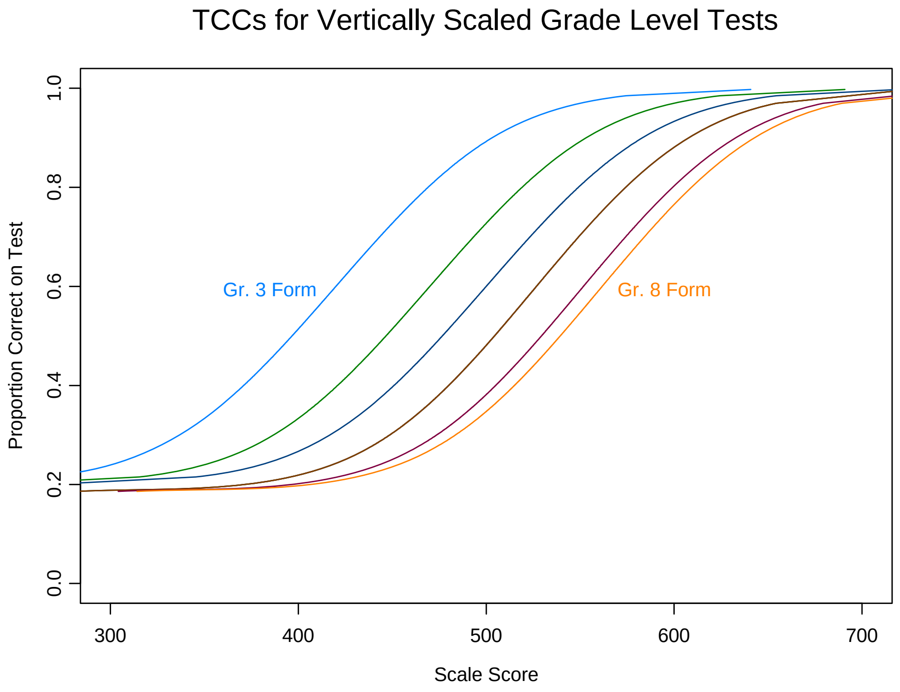

Figure 6. Test characteristic curves for a series of vertically-scaled test forms progress systematically across the scale score range.

Standard Setting

The last critical psychometric procedure that we review in relation to vertical scaling is standard setting. Standards are the centerpiece of NCLB assessment systems, and the methods available for setting performance standards for any set of tests are greatly influenced by choices regarding scaling procedures in general and vertical scaling decisions in particular.

Setting a performance standard on educational assessments is a process designed to determine a test scale cut-score that distinguishes performance below a given standard from performance that meets or exceeds the standard. Bookmark standard setting (Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001) is one of a number of methods developed to facilitate this task (see Cizek, 2001, for a review of standard-setting methods). Cut-scores in the bookmark method are determined when panelists identify a set of tasks (i.e., test items) that examinees are required to answer correctly with a given probability of success. When a standard setting panelist says that a proficient examinee should have at least a 67% probability of answering a particular item correctly, that panelist is making a judgment that helps identify what it means for an examinee to be proficient. The bookmark standard setting procedure defines a systematic way to collect from a group of panelists many such judgments regarding many test items, in order to determine a definition of proficiency (or multiple levels of proficiency) that best represents the consensus of the panelists.

The central judgment in a bookmark standard setting task is illustrated in Figure 7. For a specified probability of success, a judgment is made regarding which items we expect a student to answer correctly with at least this probability in order for the student to be designated proficient (or any other performance level of interest). The success probability criterion, illustrated as 0.80 in Figure 7, is defined in the instructions for the standard setting task (see Huynh, 1998, 2006, for a technical discussion of this criterion). If (as illustrated) a judgment is made that the first 3 items must be answered correctly with probability at least 0.80, then a bookmark is placed between the third and fourth items in an ordered item book. A corresponding scale score is identified based on the bookmark placement, and this scale score defines the performance standard that students must reach in order to be classified as proficient.

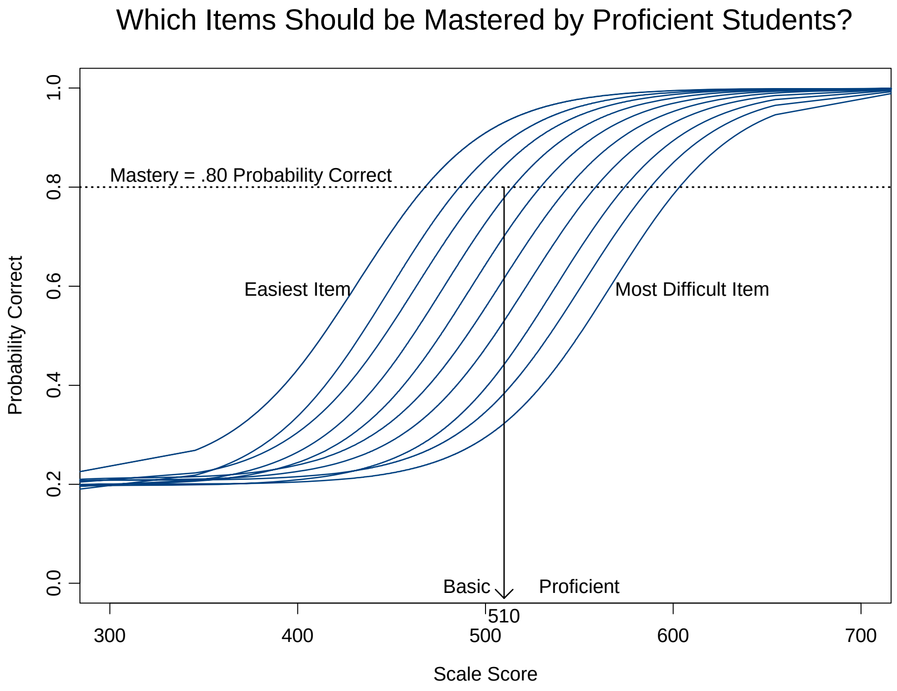

Figure 7. When test items are scaled, standard setting may proceed by identifying the set of items that must be answered correctly with a specified probability.

When tests are vertically scaled it becomes possible to set standards for each test level in a manner that utilizes information from the other test levels. Under bookmark standard setting, this may take a particular form illustrated in Figure 8. Items from both the grade 5 and grade 6 forms are considered simultaneously, and performance levels for both grades may be determined together. Specifically, under this approach it becomes possible to identify which *additional* items must be mastered to reach proficiency in grade 6 above and beyond what was required in grade 5. Perhaps more fundamentally, it becomes possible to make certain that the grade 6 performance standard is in fact higher than the grade 5 performance standard. When grades 5 and 6 are measured by tests that are not linked, or when standard setting occurs independently by grade regardless of any link, then performance standards can fail to progress in developmentally appropriate ways. The contribution of vertical scaling to the creation of developmentally appropriate performance standards over grade levels is perhaps the most compelling reason for states to consider incurring the additional costs and complexities associated with vertical scale creation.

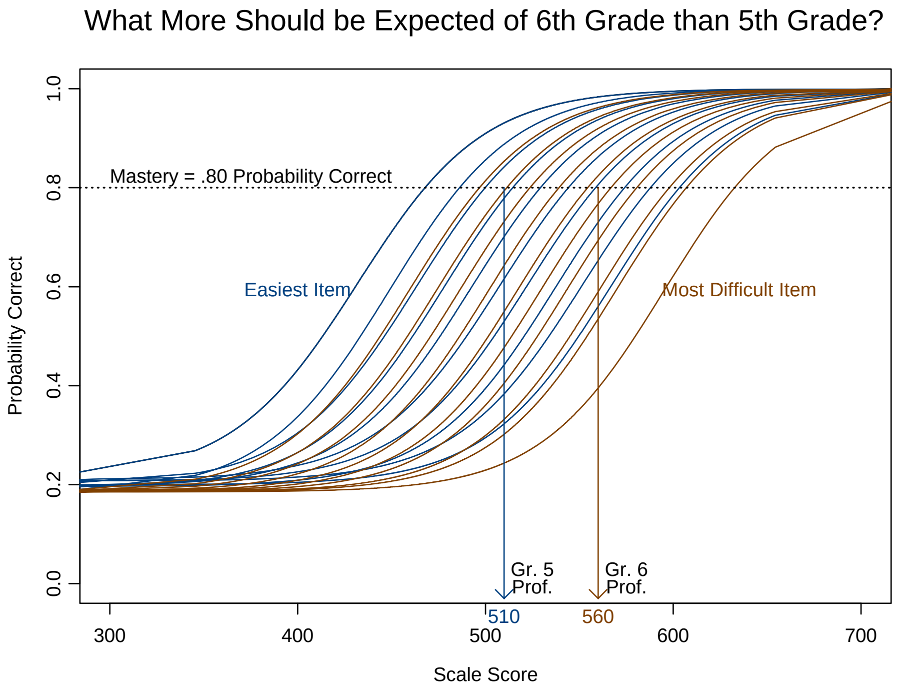

Figure 8. Vertically-scaled forms support examination of developmental appropriateness of performance standards progression over grade levels. Additional items that must be mastered at the next grade may be examined and specified during standard setting workshops.

Although the benefits of vertical scaling for standard setting have been described above for the bookmark approach, these benefits may be realized when alternative approaches to standard setting are employed. For example, under approaches that ask panelists to estimate the probability of success on each item for an examinee at the cut-score (e.g., “modified Angoff”), participants can be asked to estimate the probability of success for examinees at the grade 3 cut-score and for examinees at the grade 4 cut score. Eliciting judgments of this ordered variety, and mapping them to cut-scores on a vertical scale would also be expected to result in grade-to-grade cut-scores that progress in difficulty.

A developmentally appropriate progression of performance standards, informed by examination of content standards and empirical data on the relative difficulty of items across vertically-linked forms, might be called “vertically meaningful” standards. The meaning of one grade's performance standard may be characterized based on its relationship to the content and performance standards of adjacent grades. We note that this characterization is distinct from a progression of performance standards that result in regular patterns of population performance level classifications. When cross-grade patterns of performance level classifications are considered in the definition of performance standards, the resulting performance standards are said to be vertically moderated. We note that these characterizations although distinct are not mutually exclusive, and we believe that methods for achieving both moderated and meaningful standards by attending to developmental and population concerns simultaneously are deserving of additional research. In a similar distinction, Crane and Winter (2006) describe the result of such moderation procedures as *consistency* in performance standards and distinguish it from *coherence* in performance standards, the latter requiring a more holistic view of the standards and their articulation across grades. For a discussion of vertical moderation in the setting of performance standards, see Cizek (2005) and related articles in this special edition of *Applied Measurement in Education*.

To be concrete, developmentally appropriate performance standards set on vertically scaled tests would appear as scale score cut-points that increase monotonically over grades. For example the Proficient cut scores on Colorado's CSAP Reading Test are 526, 572, 588, 600, 620, and 632, in grades 3-8, respectively. By contrast, when the grade level tests are not vertically scaled, then the performance level cut-scores will not be directly comparable and will not progress along a continuous scale. For example, California's STAR tests set the scale in each grade level so that the Proficient cut-score is 350 in every grade, and the Basic cut-score is 300 in every grade (for links to information about performance standards in the states, visit the "State Education Agencies" link at http://www.ccsso.org).

Appropriate Contexts for Vertical Scales

Although the statistical and psychometric scaling and linking procedures required to produce vertical scales might be employed in any of a wide variety of contexts, meaningful interpretations about growth and progress will only be supported when the test forms and the student populations involved have certain characteristics. Patz and Hanson (2002) described the general requirements in this way: "When differences in population proficiency at adjacent levels are modest in comparison to differences between examinees within levels, and when the expectations or standards against which examinees are to be measured overlap extensively, then linking the adjacent test levels to a common scale will make sense and help to provide meaningful information." They note that these conditions are generally well-satisfied in achievement test batteries measuring proficiency over a range of adjacent grade levels in broad domains such as reading and mathematics.

A significantly more refined examination of the appropriate contexts for conducting vertical scaling is the "vertical alignment" work sponsored by CCSSO's Technical Issues in Large-Scale Assessment (TILSA) State Collaborative on Assessment and Student Standards (SCASS). In particular, Wise and Alt (2006), and Wise, Zhang, Winter, Taylor, and Becker (2006) describe a systematic approach to assessing and creating the necessary conditions for meaningful vertical scales.

We simply note here that some sets of content standards are clearly more amenable to vertical scaling than others. In North Carolina, for example, the K-5 Language Arts content standards are based on common goals "in order to provide continuity of language study and increasing language skill development" (North Carolina Standard Course of Study, available at http://www.ncpublicschools.org/curriculum/languagearts). By contrast, North Carolina's Social Studies content standards indicate that the state's history and geography will be the focus of grade 4, whereas United States history is the focus of grade 5. Given these characterizations, one would expect a vertical scale to provide meaningful information for language arts but not social studies in North Carolina. As states add testing at a high school grade to their grades 3-8 testing programs, it is particularly appropriate to examine the degree to which the high school content standards align with those of grade 8 before making a commitment to place the high school test on a 3-8 vertical scale.

Data Collection Designs

Constructing sound vertical scales requires the collection of data according to an appropriate psychometric assessment design. Patz and Yao (2006) provide a more complete discussion of the statistical issues in the creation of vertical scales. We characterize three general psychometric data collection approaches here, illustrated in Figure 9.

If adjacent forms share common items, then the vertical scales may be created by administering each form to students from grade for which it is targeted. Under this "common item design" the sets of common items between adjacent levels, which may be called "vertical anchor items" should be sufficiently numerous (15 or more is desirable) and representative of the domain to provide a solid link. The selection of vertical anchor items defines what growth on the resulting scale means, and the fact that this is a (perhaps small) subset of items covering a subset of content standards may be viewed as a drawback of the common item design. Under this approach, where the vertical anchor items are embedded in tests on which students will be scored, it is best that the vertical anchors be determined by the content standards shared between grade levels. If vertical anchor items are identified according to this criterion, then it matters not if the item was originally intended or is operationally being used at grades above or below any given grade. Including on an operational test form vertical anchor items aligned to content not yet covered in the curriculum at the lower grade may be deemed unfair since the approach involves testing students on content they may not have had an appropriate opportunity to learn. It is important that vertical anchor items appear in similar contexts (e.g., in similar positions in the test book) in each level in which the items appear.

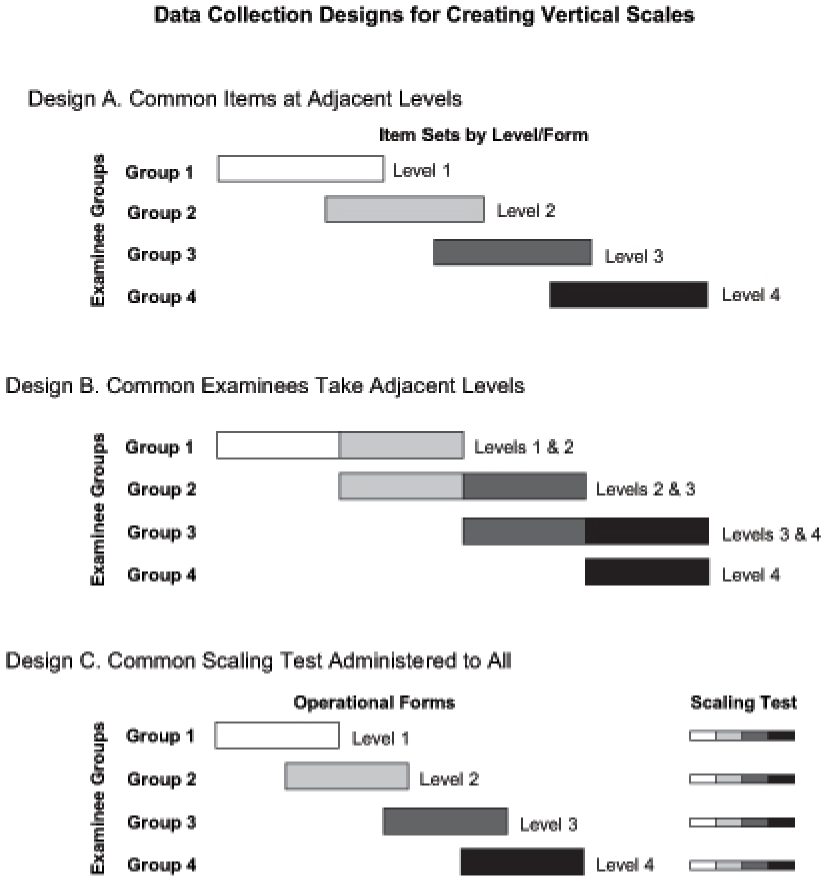

Figure 9. Creation of vertical scales requires data collection according to one of several linking designs. Forms may have items in common, or students may take multiple test forms, or an external scaling test may be employed.

An alternative approach may be adopted if it is desirable that vertically-scaled forms have no items in common. In a separate research study to be used in creating the vertical scaled forms, students may take two forms of the test—one targeted for their grade level and another targeted for an adjacent level. This "common examinee" approach has been used in the creation of nationally standardized, vertically-scaled achievement test batteries (e.g., CTB/McGraw-Hill's *TerraNova*, *CTBS*, and *CAT* batteries). In those settings, students might take a test level above their grade if the vertical scaling study is conducted in the spring, or a test level below their grade level if the study occurs in the fall. A somewhat less desirable variation of this design would have randomly equivalent groups of examinees take each form instead of having the same examinees take each form.

The third design illustrated in Figure 9 involves the use of a separate scaling test, which is constructed by sampling a small number of items from every level of the test. Examinees in the vertical scaling study take both the scaling test and a test targeted for their grade level. Patterns of within- and between-grade growth in student achievement are determined by performance on the scaling test, and grade specific forms are linked to a common scale using this information. This approach has been used to scale the *Iowa Test of Basic Skills*, for example.

Under the common item design, where the vertical anchor items are embedded in the operational test forms, no separate research study is required to create the vertical scale. Data from the operational test administration will support the construction of a vertical scale. Under the common examinee and scaling test designs, some students need to take a separate test form (an operational form from an adjacent level or a special scaling test common to all levels) in a special study. For vertical scale score interpretations to be most valid, the motivation of students on all tests involved should be the same. For example, if some students will take adjacent test levels, it is important that students have the same motivation on each. This may be achieved, for example, if the students do not know which test form will "count" for them and which one is for research purposes, and if the order of administration is counter-balanced so each form is the first form for half of the students.

Although special studies require additional testing time and possibly additional costs, establishing a vertical scale in a separate study may provide additional flexibility. For example, a form that is not intended to support the scoring of an individual student for accountability purposes may include items aligned to standards that the student has not been taught. The items may function in this case as a "pre-test" that explicitly measures content that has not yet been taught, and growth on scales thus constructed would be interpreted accordingly. This configuration would likely identify quite a different growth pattern than would be identified if all content that students encounter is aligned to standards taught prior to testing.

Reasonably robust vertical scales may be constructed using any of these approaches, and the approaches may be tailored to support specific intended uses of the tests. If comparisons across vertical links become very important, then larger numbers of vertical anchors would be appropriate. If comparisons across two or more grade levels (or four grade levels as NAEP has considered) are of particular interest, then the data collection designs may be modified to better support those interpretations.

It would be difficult to overstate the importance of very careful planning of an integrated set of test form specifications and an appropriate psychometric data collection design to the successful creation of sound vertical scales. The work of test developers is significantly complicated when the forms they create need to support vertical scales, and this is true under any of these data collection designs.

One way some state assessment programs have created vertical scales involves linking their state tests to vertically-scaled, nationally standardized achievement tests. This approach will generally involve significantly lower costs and less complexity in assessment design and data collection, since no vertical anchor items are required. The disadvantage of this approach is that comparisons of performance at adjacent grade levels rests on multiple links, each of which is prone to some degree of noise or error. The state test at a given grade level is linked (horizontally) to the nationally standardized test at this grade level, the nationally standardized test is linked to its adjacent grade form based on the publisher's vertical scaling analyses (typically conducted years earlier using a national population), and the adjacent grade level of the standardized test is linked to the state's adjacent grade level test through another horizontal linking analysis. Comparability of scores across grade levels is therefore significantly weaker than that which may be derived from a carefully conducted vertical scaling analysis of the state test.

It is important to recognize that different data collection designs—and different data analysis and modeling approaches for a given designs—will result in vertical scales with different properties. Vertical scaling has a bit more "art" involved, when compared to other types of linking or equating, and the interpretations supported by the vertical links should be understood accordingly.

Advantages of Vertical Scales

Vertical scaling may bring several compelling features to achievement tests. Vertical scales facilitate the estimation and tracking of growth over time, as repeated measures (i.e., comparable scale scores) on individual students using different, age- and grade-appropriate test forms becomes possible. This should help determine how much growth has occurred over time and in different regions of the proficiency range. Second, it would appear that vertically-scaled achievement tests allow comparisons of one grade level to another and one cohort of students to another at any point in time. Figure 10 depicts the results of a vertical scaling analysis of one state's mathematics test. This type of graphical and/or statistical depiction of the relationship of achievement at multiple grade levels is only possible when the grade level tests are vertically scaled.

Vertical scaling of test forms also enables important comparisons regarding test items. Vertical scaling can lead to more efficient field testing of new content, as items targeted for one grade might be found to be of more appropriate difficulty for an adjacent grade, assuming that the targeted content standard is present in both grades. Final form selection for a target grade can then identify appropriate items from a larger pool, when each item in the pool has parameters on a common scale. In addition, as noted above, standard setting judgments can be made more developmentally appropriate. In addition, the standards may be made more precise in theory, since a richer set of items (from adjacent levels of the test) may be ordered and the scale may thus be more finely segmented as the density of items increases.

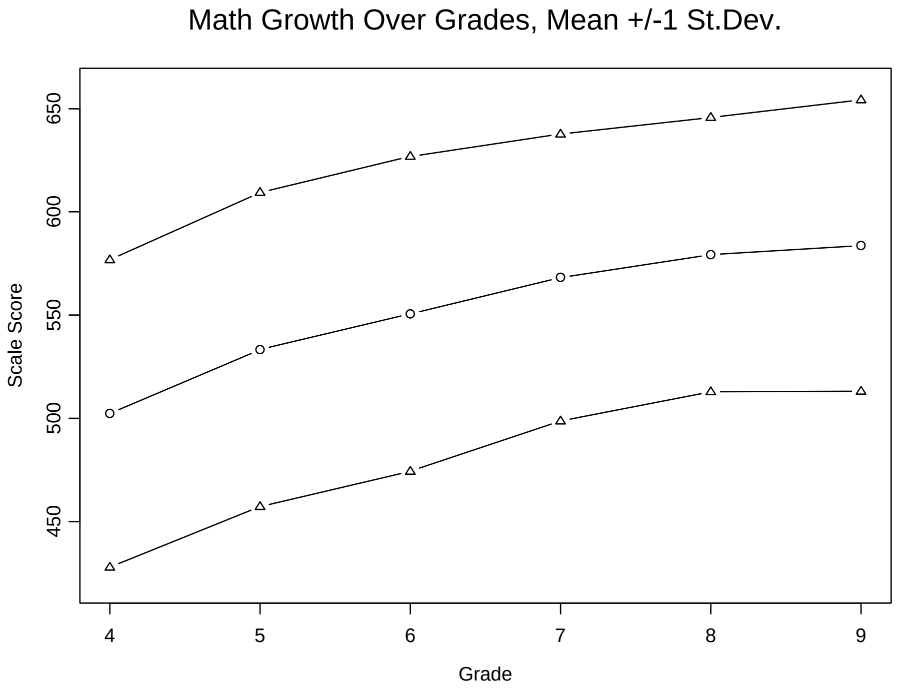

Figure 10. Population results for a vertically-scaled state achievement test. Variability within grade levels typically exceeds variability between grade level means.

Limitations of Vertical Scales

There are significant limitations to the inferences that vertical scaling of test forms can support. Because adjacent test levels are not parallel forms, vertical scaling does not constitute an equating of forms in the strict sense but instead represents a form of test linking, as described above. In particular, since vertical links provide for weaker comparability than equating, the strength of the validity of interpretations that rest on the vertical links between test forms is weaker. For example, in typical state assessment programs, grade-level forms are built to be parallel and carefully equated across years. Comparing students or groups of students who take parallel forms will generally be strongly supported, and this is appropriate since most decisions of consequence for students and schools rest on these types of comparisons. By contrast, comparisons of student performance from one test level to another (e.g., students in different grades, or one student over multiple years) are based on a lesser degree of comparability and should accordingly have lower stakes.

The validity of inferences based on vertical linking will generally diminish with the distance between levels under most vertical scaling research designs. That is, although an achievement test battery may measure achievement in mathematics on a scale that spans grades 3-8, comparisons between scores based on the grade 3 and grade 8 forms will typically not be well supported, although comparisons based on adjacent forms may be very well supported. Specifically, if a third-grade student earns a scale score above the grade 6 proficiency cut-score, it would not be appropriate to conclude that the third grader had mastered the grade 6 standards. Vertically linking forms intended to measure achievement at non-adjacent grade levels has proven difficult. NAEP attempted to vertically link forms for grades 4, 8, and 12, but abandoned the approach as the comparisons of students separated by 4 or 8 years were found to be "largely meaningless" (Haertel, 1991). It would seem likely that in these applications the most general requirements noted above—overlap in content standards and proficiency distributions for levels to be linked—may not have been satisfied adequately.

A measurement scale should have the property that the units of measurement possess the same meaning across the range of the scale. Although true of familiar measures of height and weight, for example, this property is only at best approximated in scales built for measuring latent variables such as proficiency or achievement. A ten point difference in scale score units may mean something different at the low end of the score scale than it does at the middle or high end, for example. This challenge to interpreting changes in scale scores is made more difficult when vertical scales are involved. Growth from 300 to 350 in scale score units may have a different meaning for third graders than for fifth graders, for example. The failure of achievement scales to achieve true "equal-interval" properties suggests that caution and additional validation efforts are appropriate when changes in scale scores become a focus of interest or accountability.

Additional statistical and observational research is needed to quantify more precisely these limitations of vertical scales. Such research might shed light on the manner and degree to which the validity of comparisons degrades across multiple links of vertically scaled grade level tests. With our current understanding of the limitations of vertical scales, it is prudent to keep the stakes associated with vertical comparisons relatively low.

Characteristics of Well Constructed Vertical Scales

In order to construct vertical scales that will support meaningful interpretations, several elements are required. A state must have a set of vertically-aligned content standards with considerable grade-to-grade overlap and a systematic, intentional increase in difficulty. A robust vertical scaling design specifying the psychometric procedures and data collection approaches to be used, including sufficient numbers of common items across levels or sufficient numbers of students taking multiple forms, is needed. It is highly desirable that the data collected to create the vertical scales be gathered during an operational administration or under conditions closely approximating the operational conditions, and that statewide data or large, statistically representative samples of students be involved in the vertical scale data collection.

When vertical scales have been well constructed for use in large-scale educational testing programs, we would expect to see a number of technical characteristics. Since the forms are intended to progress in difficulty, we should see evidence that this has been achieved. Test characteristic curves, for example, should show evidence of increasing difficulty across grades. For sufficiently large and diverse samples of students, scale score means would be expected to increase with grade level, and the pattern of increase would be expected to be somewhat regular and not erratic. Erratic or non-increasing patterns of mean scale score growth or large differences in scale score standard deviations from one grade to the next would warrant special scrutiny. When scrutinizing such results, it is important to consider as fully as possible the context of the test (e.g., Are the stakes lower in some grades than other? Are there significant changes in the curriculum at certain grades?), as well as all of the psychometric and statistical issues.

States should expect that all information pertaining to their vertical scales would be well documented in an appropriate technical report. Statistics reported should also include correlation of adjacent-level test scores (under the common examinee design) and/or correlation of item difficulties across test levels in the common item design. High degrees of correlation suggest that the examinees and/or items would be ordered the same way on adjacent test levels, which may be taken as a degree of validation that vertical scaling is appropriate. The specific methods employed, the data collection design, appropriate descriptive statistics regarding test items and groups of examinees, and the resulting scale score patterns should be thoroughly documented.

Some Recent Developments in Vertical Scaling Technology

Vertical scaling has seen a resurgence of interest in recent years, perhaps much of it attributable to the change in federal education policy in the United States that resulted in many states adopting new testing programs for grades 3-8.

For example, Patz and Yao (2006) examined approaches to modeling cross-grade growth trajectories in a hierarchical modeling framework. The approach may support the identification of more regular underlying growth patterns than those that may be inferred by fitting simpler IRT models and examining the resulting grade-by-grade trends.

Patz and Yao (in press) also examined the application of multidimensional IRT models to vertical scaling problems, as have Reckase and Martineau (2004). It is widely recognized that the assumption of unidimensionality underlying standard IRT measurement models is an over-simplification of reality, and these recent investigations into multidimensionality in vertical scaling data support this observation. Nonetheless, it is not yet clear whether or how more complex multidimensional models might bring greater validity to the essentially unidimensional classification decisions required by standards-based accountability testing. This remains an area rich with research possibilities.

Kolen and Brennan (2004) added a chapter on vertical scaling to their measurement text, and they include a discussion of several ways to evaluate the appropriateness and validity of vertical scaling results. Karkee, Lewis, Hoskens, and Yao (2003) applied the Kolen and Brennan criteria when examining vertical scaling results under a variety of model estimation and forms linking strategies.

Policy Context: NCLB and Growth Model Pilot Studies

The landscape of educational assessment in the United States changed in fairly dramatic ways following the passage and subsequent implementation of the *No Child Left Behind (NCLB) Act* of 2001. NCLB embraced and advanced in policy standards-based education reform, and coupled it with a very strong emphasis on the quantification, measurement, and monitoring of school performance. Attention in states has been focused on defining educational achievement standards and measuring the progress of all students and certain subgroups of students in meeting the standards. Rewards and sanctions flow to schools and districts based on their ability to meet clearly articulated performance goals, measured in large part by new student testing programs aligned to the standards. NCLB is a complex law with many facets. For our purposes here, we simply highlight several features related to growth and progress, since the measurement of growth and progress might more generally be a motivation for the creation of vertical scales.

NCLB's definition of what constitutes "progress" for accountability purposes is very simple. Schools are said to make "adequate yearly progress" (AYP) if they meet a set of prescribed performance standards. School achievement in relation to performance standards is quantified as the percent of students at or above the proficient performance level. The level of performance that must be achieved is designated by the state each year, increasing to a goal of 100% by 2014. The critical characteristics that distinguish NCLB's definition of AYP from other standards-based accountability programs are that i) the designated level of performance must be reached by all identified subgroups of students; ii) the required level of performance is common to all schools in a state without regard to school characteristics, inputs, or demographics; and iii) no attention is paid to trends in performance, be they increases or decreases in the percent at or above proficient relative to previous years. For example, highly-resourced schools trending sharply downward can be designated as having made adequate progress, whereas under-resourced schools making dramatic year-to-year gains may continue to be designated as failing. Resources and trends do not affect basic AYP designations. Growth is required and rewarded only indirectly as, in the later years of the program, state targets will rise to 100%, forcing in theory all schools to follow an upward trend. NCLB does have a "safe harbor" provision that opens the door to consideration of growth. It currently specifies that a 10% improvement in percent at or above proficient (relative to the number of students not proficient in the base year) will protect a school from sanctions that would otherwise be associated with school performance below the designated level.

States can meet their NCLB Title I accountability requirements without creating vertical scales. Although vertical scales offer many advantages to states, as discussed earlier in this paper, compliance with NCLB requirements is not among those advantages.

In 2006 the United States Department of Education (ED) invited states to propose "growth models" as a part of their accountability systems. Under the pilot programs, states could modify their criteria for determining adequate yearly progress to include growth measures. The guidance in the invitation was fairly general and not prescriptive, presumably allowing states a good degree of flexibility in their approaches to incorporating growth. Nonetheless, of 20 states that submitted applications under the pilot program, eight states' proposals were approved for a peer review, and only two states—North Carolina and Tennessee—had their proposals approved in the initial round of submissions. In a second round of submissions in fall 2006, Delaware had its proposal approved, and Arkansas and Florida received approval of their growth proposal, subject to conditions regarding other parts of their NCLB programs.

An explication of the ED principles and procedures for evaluating growth model proposals is beyond the scope of this paper. Nonetheless, we are able to make some observations regarding vertical scales. ED's answers to direct questions from CCSSO about vertical scaling (available at http://www.ccsso.org/content/PDFs/GrowthModelsNotesDec05.doc) suggested that vertical scaling would not necessarily be required and that "simpler approaches" could be approved.

The first-round approval of North Carolina and Tennessee applications is noteworthy for several reasons, one of which is the clear implication that vertical scaling is not required for states to add a growth component to their accountability systems. Although both North Carolina and Tennessee have used vertical scales in their assessment systems, they do not rely on vertical scale change scores—and the assumption that the magnitude of such change scores has the same meaning regardless of the location on the scale or the grade level of the student. Rather than rely on such strong assumptions about the technical characteristics of vertical scales, these states characterize the magnitude of score changes in relation to historical norms (see Smith & Yen, 2006, for a related example). Such an approach would appear to be a reasonable complement to the explicitly non-normative basis for NCLB status designations (i.e., performance in relation to a standard as measured by test forms that are equated separately by grade level from one year to the next).

Note that a decision to define growth targets normatively does not preclude the use of vertical scales. Statistical projections of expected (or exceptional) growth may be made using historical data on performance characterized by scores on vertically scaled tests. Examining such historical data from vertically-scaled tests could enrich the interpretation of performance, and changes in performance, across ranges of the scale and for students in different grade levels.

For example, if all scales for all grade levels are not vertically linked but instead set so that the proficient cut score is 300 and advanced is 500, then differences in average scale scores from one grade to the next will not be meaningful except in relation to norms for such differences. If the historical average score is 320 for grade 4 and 315 for grade 5, then a student earning 320 in both grades 4 and 5 will have grown more than the average student. If, by contrast, the scales were vertically linked then the average changes from one grade to the next would be expected to be positive, and the magnitude of changes along the scale could be interpreted (with some caution for reasons discussed above) in relation to the increasing difficulty of the content from one grade to the next. Normatively defined expectations for growth along vertical scales are possible and prudent, and interpreting growth from a developmental perspective as well as a normative perspective becomes possible.

States may choose to focus not on changes in scale scores but rather on changes in performance levels (defined as intervals on the scale) as students move from one grade to the next. This is the approach proposed and approved for Delaware. When changes in performance levels at different grades are to be treated (or valued in a “value table”) interchangeably, then this approach rests on assumptions that may be best evaluated by a vertical scaling analysis. Here again, vertical scaling can benefit the state testing program without the scale scores themselves becoming the focus of either status or growth interpretations.

Florida’s growth model proposal does include the measurement of growth across years on the state’s vertically scaled tests. Thus, a variety of approaches to measuring growth have been approved, and vertical scaling appears to be neither required nor precluded under current federal policy.

Discussion

Vertical scaling offers many compelling features to an assessment system. Item response theory is a powerful framework for building tests and understanding their measurement properties, and IRT-based vertical scales promise greater comparability of scores, greater efficiency in test construction and field testing, and better standard setting judgments.

This power of IRT is derived in large part from some relatively strong assumptions, and the validity of interpretations based on the vertical scales depends upon evidence available to support assumptions. For these reasons, it makes sense to examine these assumptions critically in light of the available data.

Most fundamentally, creation of valid vertical scales requires a set of content standards that reflect a degree of continuity over grade levels. Vertical links between test forms of different difficulty are less reliable than horizontal equating relationships, so the stakes associated with inferences based on vertical links should be lower, too. Assessment programs that do not employ vertical scales eliminate opportunities for misinterpretation or over-reliance on the cross-grade comparisons that vertical scales invite.

Building a vertical scale is not merely a matter of psychometric procedures but instead requires careful design work at all stages of test development, from creating test blueprints to setting performance standards. These additional complexities add some cost and complexity to test development, although some of this cost may be offset by more efficient field testing and item utilization that vertical scaling analyses enable.

We note that neither vertical scales nor item response theory are required for measuring growth using tests. Relatively straightforward pre-test/post-test designs using alternate test forms can and have been employed to measure growth in simple and interpretable ways. Such approaches may have limited application under NCLB, however, where one summative test is of interest and where comparability must be maintained as items on the test are changed each year.

NCLB Title I does not require vertical scales, and vertical scales are not required to bring a defensible growth component into an NCLB state accountability system. Nonetheless, the absence of a vertical scale in NCLB Title I assessment systems implies some significant limitations. Most importantly, perhaps, the developmental appropriateness of a progression of grade-level performance standards may not be directly observed and assessed.

When their use is appropriate and their construction is sound, vertical scales can significantly enrich the interpretations of test scores and growth trajectories. They provide a systematic way to examine the developmental characteristics and appropriateness of systems of state performance standards across grade spans. This is in contrast to “vertical moderation” approaches that achieve grade-to-grade consistency by normative analyses. One fruitful area for additional research concerns the most valid approaches to achieving both developmentally appropriate and normatively consistent performance standards across grade levels.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

Cizek, G. J. (Ed.). (2005). Adapting testing technology to serve accountability aims: The Case of vertically-moderated standard setting. A Special Issue of *Applied Measurement in Education*.

Crane, E. W., & Winter, P. C. (2006). *Setting coherent performance standards*. Washington, DC: Council of Chief State School Officers.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C., (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley.

Haertel, E. (1991). *Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress* [ERIC Clearinghouse Document Reproduction Service No ED404367]. Washington, DC: National Center for Education Statistics.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hieronymus, A. N., Linquist, E. F., & Hoover, H. D. (1982). *Iowa Test of Basic Skills: Manual for school administrators.* Chicago: Riverside.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23,* 38-58.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard setting based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Karkee, T., Lewis, D. M., Hoskens, M., & Yao, L. (2003). *Separate versus concurrent calibration methods in vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed). New York: Springer-Verlag.

Lewis, D. M., Mitzel, H., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Patz, R. J., & Hanson, B. (2002). *Psychometric issues in vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Patz, R. J., & Yao, L. (2006). Vertical scaling: Statistical models for measuring growth and achievement. In S. Sinharay & C. Rao (Eds.), *Handbook of statistics, 26: Psychometrics*. Amsterdam: North Holland.

Patz, R. J., & Yao, L. (in press). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales.* New York: Springer-Verlag.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D., & Martineau, J. (2004). *The vertical scaling of science achievement tests*. Paper commissioned by the Committee on Test Design for K-12 Science Achievement. Washington, DC: National Research Council.

Smith, R. L, & Yen, W. M. (2006). Models for evaluating grade-to-grade growth. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance* (pp. 82-94). Maple Grove, MN: JAM Press.

Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychological Review, 35,* 175–197.

Wise, L., & Alt, M. (2006). *Assessing vertical alignment*. Washington, DC: Council of Chief State School Officers.

Wise, L. L., Zhang, L., Winter, P., Taylor, L., & Becker, D. E. (2006). *Vertical alignment of grade-level expectations for student achievement: Report of a pilot study*. Washington, DC: Council of Chief State School Officers.