15 Vertical Scaling and No Child Left Behind

Wendy M. Yen[1]

Educational Testing Service

The chapters by Harris (Chapter 13) and by Patz and Yao (Chapter 14) are quite different examinations of vertical scaling issues. The Harris chapter surveys practical issues related to implementing vertical scales, and the Patz and Yao chapter primarily studies the complex technical issue of using multidimensional item response theory models with vertical scaling. Given the great differences between these chapters, it is difficult to provide an integrated discussion of them. Thus, although this chapter contains some brief comments on the Harris, and Patz and Yao chapters, most of this chapter contains general observations on vertical scaling, observations harvested from vertically scaling K-12 achievement tests for over 25 years. Over those years, interest in vertical scales has changed. In particular, the No Child Left Behind Act of 2001 (NCLB) has led to changes in both who is interested in developing vertical scales and why they want to develop them. These changes have produced differences in expectations, evaluations, and issues related to implementing vertical scales.

15.1. Comments on the Other Vertical Scaling Chapters

15.1.1. The Harris Chapter

The Harris (Chapter 13) chapter is an excellent survey of the conceptual, technical, implementation, and maintenance issues related to the development and use of vertical scales, and the chapter provides a particularly valuable reference list. The Harris chapter should be read by

[1] The opinions expressed in this chapter are those of the author and not necessarily of Educational Testing Service.

anyone interested in general issues of vertical scaling. One particularly useful aspect of the Harris chapter is that she raises questions that need to be answered by those creating vertical scales. Perhaps the most telling of these questions is, Do you really need a vertical scale? I will address that question in relation to NCLB requirements.

15.1.2. The Patz and Yao Chapter

The Patz and Yao (Chapter 14) chapter contrasts vertical scaling based on a "divide-and-conquer" approach with vertical scaling within a unified item response theory (IRT) model. In the divide-and-conquer approach, test levels are scaled independently. Then a procedure such as that of Stocking and Lord (1983) is employed to link the results of adjacent levels. In a unified approach, the test levels are scaled simultaneously. Patz and Yao discussed limitations of a common unified approach (concurrent, multigroup, unidimensional IRT calibration of test levels), and they examined a unified model alternative. A multidimensional multigroup model was employed, allowing scores to be weighted averages of underlying dimensions, with the weights varying by test level. Such a model permits the explanation, based on empirical results, of complex shifts in what tests are measuring grade by grade.

Multidimensional modeling holds promise for K-12 assessment, although, as the authors noted, more work is required on the models before they are ready for operational implementation. One caution that I would note is that K-12 test users are understandably very focused on NCLB accountability. For that reason, they have great interest in the scores and state standards against which they are being evaluated. They want to know how their students are doing relative to those standards and what they need to do to improve performance relative to those standards. They have minimal interest in any score or subscore that is empirically identified that they cannot directly relate to the state standards. Thus, to be useful to K-12 educators, any dimensions empirically determined from a complex scaling model need to be related to state standards.

15.2. Vertical Scales: An Historical Perspective

15.2.1. A Folding Ruler: An Aside

I have been interested in vertical scales for a bit more than 25 years. When I was about 5 years old, I used to follow my father around as he did home improvements. He had a folding ruler with which I would play. It was

yellow, with hinged 1-foot lengths that would unfold (making a nice thwacking sound) to 6 feet. If I held the extended ruler at one end, it would curve gracefully through space. To my disappointment, if I leaned it too much to the side, one of the looser hinges would suddenly bend sharply.

A vertical scale is akin to a folding ruler. Although educational achievement tests tend to have very strong first factors, they are multidimensional, paralleling changes in the curriculum. This dimensionality changes both within and across test levels. The direction of the scale (i.e., the relative importance of the different dimensions) changes as the test levels become more difficult. Thus, the scale bends or curves through space. Connections between some levels are stronger (i.e., have tighter hinges) than others, and sometimes the links between levels are too loose to maintain a sturdy connection between the test levels.

15.2.2. Pre-NCLB Interest in Vertical Scales

Before NCLB, the K-12 norm-referenced test (NRT) test publishers (such as CTB/McGraw-Hill, Harcourt Educational Measurement, and Riverside Publishing) conducted the vast majority of the vertical scaling. They produced these scales to satisfy users and to facilitate internal business systems. The primary uses of vertical scales were grade equivalents, functional level testing, scale scores for growth analyses, and computer-adaptive testing.

A large-scale K-12 test publisher cannot stay competitive without grade equivalents, which are demanded by customers. Grade equivalents are developed through the combination of a vertical scale and norms (Petersen, Kolen, & Hoover, 1989). The development of grade equivalents requires that normative scale score averages increase by grade. The vast majority of uses of grade equivalents are low stakes; they basically are a means of communicating the main idea of test results to those with minimal testing background, such as students, parents, and some teachers.

In functional level testing, a short locator test is used to identify the (vertically scaled) test level that is best matched to a student's current achievement. That level is then administered to the student. Scores obtained on different test levels that are linked via the vertical scale (e.g., scale scores, normative results) can be pooled for group reporting. Results that are not vertically linked—such as number-correct scores on the full test or on its subscores—cannot be pooled for group reporting. The promise of functional level testing is to obtain the most accurate measure

of a student's performance, given the multiple test levels that are available. The problem with functional level testing is that it is operationally cumbersome to administer different test levels to different students in the same classroom. The fact that raw scores cannot be pooled across test levels is also awkward. Despite its promise, users today rarely choose to use functional level testing in operational K-12 testing programs.

When K-12 test levels are vertically scaled, these scale scores can be used to longitudinally track academic growth of individual students or cohorts. Before NCLB, there were a few sophisticated school districts that chose to conduct such growth analyses; however, the vast majority of users depended on cross-sectional results, for example, comparing this year's fourth-grade students to last year's. Some large-scale research studies on hierarchical modeling conducted by university researchers used vertical scales, and at least one state (Tennessee) used vertical scales, in combination with national norms, to conduct value-added evaluations of teacher effects on growth in student achievement test scores (Braun, 2005; Sanders, Saxton, & Horn, 1997).

Vertical scales based on K-12 achievement tests or items can also be used in computer-adaptive testing (CAT). As with functional level testing, the goal is to get the most accurate measure of a student's achievement as efficiently as possible. Use of CAT algorithms for item selection and terminating testing virtually require use of IRT to calibrate the items on one scale. Whereas in the late 1970s and early 1980s CAT appeared to hold tremendous promise for K-12 testing, as well as for testing in many other settings (Weiss, 1983), to date only a small minority of school systems have used CAT in K-12 (e.g., Northwest Evaluation Association; Kingsbury & Hauser, 2004).

K-12 publishers rely on vertical scales to organize their internal psychometric analysis systems. Publishers have very large numbers of items whose psychometric qualities need to be stored, accessed, and used. There might be items for 13 grades for 10 or so content areas (e.g., word analysis, reading vocabulary, reading comprehension, language expression, language mechanics, mathematics computation, mathematics problem solving, science, social studies.). In those systems that employ IRT, the items' parameter(s) are stored in scale score units (i.e., based on the cross-grade vertical scaling). This greatly facilitates the selection of items to create test forms at a variety of appropriate difficulty levels for either shelf or custom assessments.

In addition, scoring systems are arranged using the (vertical) scale score system. For example, to score a particular test form/level, the item parameters or raw score-to-scale score conversion table is stored in scale score units. When a student's test form/level is identified, the appropriate

scoring table or algorithm translates the student's responses into a scale score. The normative tables contain scale score-to-norm conversions (e.g., scale score-to-grade equivalent, scale score-to-percentile). The norm tables are organized by grade/testing date (e.g., grade 3 fall, grade 3 spring) and are independent of the test form/level that the student took. Thus, the vertical scale provides an efficient backbone for the organization and access of items, test forms, and normative derived scores.

How successful were the pre-NCLB vertical scales in meeting the user and publisher needs?

As described earlier, to develop grade equivalents, it is necessary that the vertical scales show increasing average performance over grades. In the development of test blueprints, the K-12 publishers carefully map content strands to provide overlap and connections between the measurements at different grades and test levels. To demonstrate between-grade growth, this design must be accurately connected to typical or modal curricula across the nation. In the vast majority of cases, K-12 test publishers have been successful in producing measures that showed grade-to-grade growth. This growth is not necessarily smooth, but smoothness is not expected when there can be variations over grades in the strength of the connection between tests and curricula. The last 2–3 years of high school typically show minimal growth between grades, perhaps due to a looser connection of norm-referenced tests to high-school curricula than elementary curricula. Lower motivation for older high-school students could also play an important role. Despite these difficulties, K-12 publishers produced measures with vertical scales that demonstrated normative growth over grades.

The vast majority of uses of NRT results are horizontal: to compare this year's results (a) to a national norm at the same grade level or (b) to last year's results for that grade for the same school/district/state. Other uses that rely on the vertical scale (e.g., grade equivalents) tend to be low stakes. Publishers do provide cautions about using results from different parts of a vertical scale (e.g., a student at a lower grade getting a high scale score is probably thinking about content differently than a higher grade student getting that same scale score). It is also generally acknowledged and accepted that cross-grade correlations of scores are lower than within-grade (between parallel form) correlations of scores.

It is worth mentioning that in the 1980s, a brouhaha arose about scale shrinkage that occurred with some IRT vertical scales (Camilli, 1988; Clemans, 1993; Yen, 1986; Yen & Burket, 1997; Yen, Burket, & Fitzpatrick, 1996). In scale shrinkage, scale score standard deviations and IRT item difficulty parameter standard deviations decrease over grades,

and IRT item discrimination parameter means increase over grades. Many hypotheses were generated to explain this phenomenon and there was much discussion about the implications of scale shrinkage. In actuality, because the vast majority of uses of NRT scores are horizontal, few test users were aware of the issue or cared about it, and scale shrinkage remained an issue primarily of academic interest. With the evolution of test design and IRT parameter estimation software, scale shrinkage disappeared.

Overall, the vertical scales developed by K-12 NRT publishers successfully addressed the needs of users and publishers.

15.3. The NCLB Era

Under NCLB, it is the responsibility of each state to develop its own challenging content standards and assessments to measure progress in achieving those standards. With the advent of NCLB, interest in NRTs has greatly declined, although some states do take an NRT core set of items (and vertical scale) and augment it to improve the coverage of unique state standards. There is interest in vertical scaling for Titles III and I of NCLB and for evaluation of growth.

15.3.1. Title III

Title III of NCLB states, "A State shall approve evaluation measures…that are designed to assess…the progress of children in attaining English proficiency, including a child's level of comprehension, speaking, listening, reading, and writing skills in English." Title III generates interest in vertical scales, both explicitly and implicitly. Nonnative English-speaking children enter our schools with a wide range of English skills, so in assessing these skills accurately, functional level testing (which assumes the existence of a vertical scale) can be particularly important. Behavioral scale anchoring (i.e., examples of what students know and can do at different scale scores) is of interest to those trying to attach meaning to the student scores. On these vertical scales of English proficiency, setting performance standards related to exiting English learner programs is of particular importance.

At Educational Testing Service (ETS), we have recently developed vertically scaled assessments of English acquisition skills for two different clients (Comprehensive English Language Learning Assessment

[Educational Testing Service, 2005] and the New York State English as a Second Language Achievement Test [Wang & Smith, 2003]). These assessments display properties different from traditional measures of achievement. For example, the lowest test level, measuring introductory skills, can include a wide range of content (letters of the alphabet, words, sentences, paragraphs) and show much greater units of growth that those seen at higher test levels. These differences reinforce the advice given by Braun (1988) that growth is most accurately evaluated by comparing students who start at the same place; when students start at different places on a scale, differences in scale units can greatly complicate interpretations. At the group level, cross-sectional results show much different growth patterns over the grades for listening and speaking than for reading and writing. listening/speaking skills rise rapidly in the early grades and then top out. reading/writing, which are academic skills, continue to rise throughout the grades. For traditional achievement measures, "grade" is the most relevant time measure; however, for English acquisition skills, both "number of years in the United States" and "grade" are relevant time measures. In interpreting cross-sectional growth over grades for English acquisition tests, immigration patterns also need to be considered. For example, whereas for traditional achievement measures, growth is expected across virtually all grades, for English acquisition tests, performance can dip at grades where a large influx of students new to the United States can occur (e.g., grade 9, where students are coming to the United States for high school). Thus, growth expectations for vertically scaled English acquisition tests can differ from the expectations for traditional educational achievement tests.

15.3.2. Title I

Title I of NCLB focuses on the adequate yearly progress in the percents of students reaching the Proficient performance standard established in each state. Thus, comparisons are made from year to year in the percents of Proficient students at a given grade and no statistical connection is required between the tests at different grades.

Typically, the NCLB assessments and their performance standards have been developed in a piecemeal fashion, because the legislation eased in the assessment requirements over the years. For example, NCLB legislation started with a requirement (in reading and mathematics) of one assessment in each of three grade ranges (grades 3 to 5, 6 to 9, and 10 to 12). Later, states were required to have assessments in each of grades 3 to 8. Also in the Title I legislation there is no requirement for longitudinal or growth

measures. For these reasons, few states have vertical scales for their NCLB assessments. Vertical scales that demonstrate cross-sectional growth over grades can be more difficult to develop if the content standards/curricula/test blueprints have not been designed from their inception to have hierarchical content strands with substantial between-grade overlap. Furthermore, performance standards that are set independently by grade might not "grow" on a vertical scale (e.g., Proficient for grade 7 might not be at a higher scale score than Proficient for grade 8). Thus, it might be more difficult to develop vertical scales that produce expected progressions over grades for NCLB state assessments than it was for NRTs.

Although the Title I legislation does not require it, there has been increasing interest in vertical scales among NCLB practitioners.[2] Why is that? I can speculate on several reasons. First, there might be a mistaken impression among some practitioners that a vertical scale is required. Second, there are those who want to use NCLB assessment results within evaluation and accountability systems. Within such systems, being able to distinguish input (i.e., performance before a particular instructional treatment) from output is particularly helpful. Some of those interested in accountability are specifically interested in value-added models, and some of these models require the use of vertical scales. Finally, I believe that most educators care dearly about student growth, and vertical scale is a catch-all phrase that, for many people, includes any type of growth measure.

15.3.3. Educators' Interest in Growth Measures

It became important to us to understand what educators wanted in terms of a growth measure in the NCLB era. Toward that end, we gathered in-depth information from educators in one state via phone interviews, large-group meetings, and a small working group (Smith & Yen, 2006). We discussed with them the pros and cons of three types of growth measure (vertical scales, state norms, and cross-grade regressions [expectations]) and listened to the issues that they were trying to address. Their interests seemed to center around answering the following questions:

[2] In November 2005, the U.S. Department of Education invited states to submit proposals for developing growth models for adequate yearly progress consistent with the principles of *No Child Left Behind*. In May 2006, the Department approved two programs as part of this pilot (U.S. Department of Education, 2006).

Parents:

- Did my child make a year's worth of progress in a year?
- Is my child growing appropriately toward meeting state standards?
- Is my child growing as much in Math as Reading?
- Did my child grow as much this year as last year?

Teachers:

- Did my students make a year's worth of progress in a year?
- Did my students grow appropriately toward meeting state standards?
- How close are my students to becoming Proficient?
- Are there students with unusually low growth who need special attention?

Administrators:

- Did the students in our district/school make a year's worth of progress in all content areas?
- Are our students growing appropriately toward meeting state standards?
- Does this school/program show as much growth as that one?
- Can I measure student growth even for students who do not change proficiency categories?
- Can I pool together results from different grades to draw summary conclusions?

Most of these questions are variations on one underlying question: Is the amount of growth observed reasonable or appropriate? There are two aspects inherent in answering such a question: the absolute and the normative. The absolute aspect compares a measurement to a fixed criterion, such as the score needed to be called Proficient. The normative aspect arises from interest in how the growth of this particular student (or group of students) compares with that of other students. A vertical scale by itself does not address either the absolute or normative aspect of growth questions.

Cross-grade growth expectations, which are connected to proficiency levels, answer these questions without the assumptions or development costs of a vertical scale. Such cross-grade growth expectations are obtained from longitudinal data, say from grade 3 to grade 4, that are analyzed using regression techniques; scores at a subsequent grade level are regressed onto scores at a previous grade level. Figure 15.1 provides one example of a report that could display the growth results for one district relative to the regression and the absolute performance criterion (Proficiency) established by the state. In this example, grade 3 and grade 4 have independent scales, with no vertical scale connecting them. The state regression line shows the relationship of the scores for the two grades when students are tracked

from grade 3 to grade 4. Results for one district can be compared to the state results. In this particular example, the district showed above-average growth (relative to the state) for low-scoring students and below-average growth for high-scoring students. It is also possible, using graphs such as this, to separate out results for different programs within a district and compare their relative amounts of growth. Examples of individual student score reports based on longitudinal regressions are presented in Smith and Yen (2006).

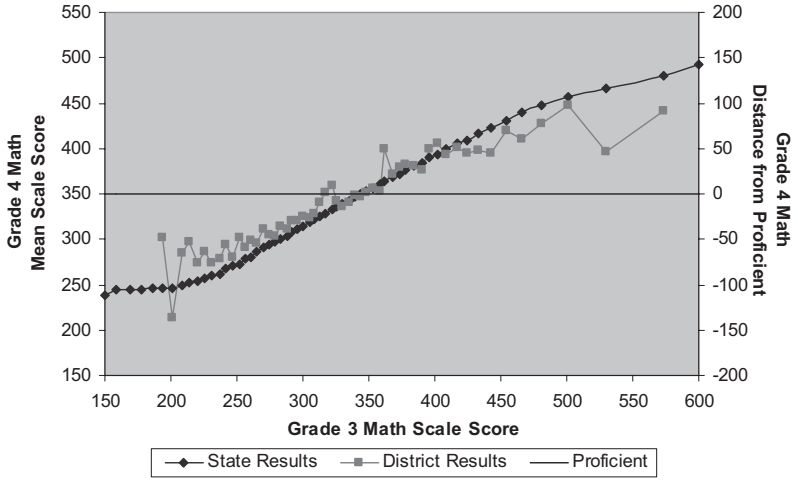

Figure 15.1. Sample longitudinal regressions of grade 4 Math on grade 3 Math at a state level and a district level.

15.4. Summary

Pre-NCLB, vertical scales were ubiquitous in K-12 assessment. The vertical scales developed by K-12 publishers satisfied general criteria for a usable scale; that is, their average scores increased by grade. The most common uses of vertical scales were embedded within grade equivalents, which were used in low-stakes settings. High-stakes usages that relied heavily on the vertical scale properties were fairly rare. Publishers did provide cautions about use of the vertical scale results.

Under NCLB, Title III requires "evaluation measures…that are designed to assess…the progress of children in attaining English proficiency…." Vertical scales are an obvious means of satisfying this requirement. In evaluating the properties of vertical scales for English

language attainment, such as expectations of increasing scores by grade, special care is needed to consider the properties of the different scales (such as academic vs. nonacademic skills) and the special characteristics of this student population. Particular care is needed in comparing amounts of growth in different parts of these scales.

In satisfying Title I of NCLB, vertical scales are not required. Vertical scales might not demonstrate grade-to-grade growth as clearly for state assessments developed under NCLB if the content of those tests, and the related curricula, have not been developed to be hierarchical. Under NCLB, users of the test scores are interested in evaluating academic growth in aspects that are both absolute (e.g., compared to a proficient cut-score) and relative (e.g., relative to how much other students grow). A vertical scale by itself does not address either of these aspects, and alternative analysis procedures can be used. For example, cross-grade longitudinal growth expectations (regressions) based on nonvertically scaled tests can address most of the growth questions being asked without the assumptions or expense involved in the development of vertical scales.