13 Practical Issues in Vertical Scaling

Deborah J. Harris[1]

ACT, Inc.

13.1. Introduction

The capability to measure students along a continuum, such as measuring growth in mathematics from grade 3 to grade 6, has become more and more important, especially with the recent federal legislation No Child Left Behind Act of 2001 (NCLB) and the concept of adequate yearly progress, by which it is to be determined if students are making sufficient gains as they advance through the education system. An assessment with a vertical scale is the most common way of evaluating growth from one grade level to another.

Vertical scaling refers to the process of linking different levels of an assessment, which measure the same construct, onto a common score scale (see Holland, Chapter 2, for placement of vertical scaling into a linking framework). Many elementary and secondary test batteries report scores on a vertical scale, such as the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001) and ACT, Inc.'s Educational Planning and Assessment System (EPAS; ACT, 2000).

Why is there a need for a chapter addressing practical issues? Because when one constructs a vertical scale, decisions have to be made with respect to the definition of growth, scaling design, statistical methods, type of scales, and so forth (see Harris, Hendrickson, Tong, Shin, & Shyu, 2004; Kolen, 2003). Different decisions can lead to different vertical scales, which in turn can lead to different reported scores and different decisions. The literature shows that vertical scaling is design dependent

[1] The opinions expressed in this chapter are those of the author and not necessarily of ACT, Inc.

(Harris, 1991), group dependent (Harris & Hoover, 1987; Skaggs & Lissitz, 1988; Slinde & Linn, 1979), and method dependent (Kolen, 1981; Skaggs & Lissitz, 1986).

This chapter examines issues that a practitioner would encounter when developing a vertical scale for an operational testing program. Although there is no single right way to develop a score scale, there are many options available, and the practitioner who chooses a method with a careful eye to both the purpose of the scale (i.e., how the resulting scores are intended to be used) and to the literature is more likely to create a scale that will facilitate appropriate decision-making. The chapter considers five sets of issues: conceptual, technical, implementation, maintenance, and other. An example involving the vertical scaling of two math tests is given throughout the chapter to provide an illustration of some of the issues that are discussed. The example uses data from the PLAN and ACT mathematics tests. However, the reader interested in a more complete summary should consult the PLAN and ACT technical manuals (ACT, 1997a, 1999).

The literature cited in this chapter, as well as on Web sites and in technical manuals and in other documentation relating to operational verticals scales (although the latter are often scant on details), should be consulted for additional information. Specific papers are helpful to address specific issues; four sources are recommended for general treatments and overviews on vertical scaling: Kolen and Brennan (2004), a book on equating and scaling that covers many issues related to vertical scaling, Kolen (2003) a conference presentation that discusses several topics in vertical scaling that need to be addressed, Petersen, Kolen, and Hoover (1989), a chapter that covers basic scaling and linking information, and Harris et al. (2004), a conference presentation that discusses practical issues related to vertical scaling. Literature not specific to vertical scaling, such as equating literature, item parameter calibration literature, computer estimation program manuals, and score reporting literature, should also be consulted, as vertical scaling covers a wide range of issues. The companion chapters by Patz and Yao (Chapter 14) and Yen (Chapter 15) should also be consulted.

It is also recommended that the reader consult multiple sources, because inconsistencies abound: for example, the Rasch model was found to be both acceptable (e.g., Schulz, Perlman, Rice, & Wright, 1992) and unacceptable (e.g., Phillips, 1983) for vertical scaling applications. Similarly, grade-to-grade variability in ability was shown to increase (Andrews, 1995; Yen, 1986), decrease (Hoover, 1984a), or remain stable across grade levels (Bock, 1983). Harris et al. (2004) contained the beginning of a comprehensive review of the literature related to vertical scaling, which might be useful to those readers who either have difficulty gaining access to the original source material or who prefer to read a summarized version. In addition to providing information as to where various methods yield consistent, or inconsistent, results, a comprehensive summary of the literature also helps identify issues requiring further investigation.

As mentioned earlier, valuable information regarding vertical scaling is also found by examining current practice. Vertical scales continue to be built and used by test publishers, despite the lack of a commonly accepted set of procedures. Although research done using simulated conditions can be very informative, what is actually done in practice might be of most interest to potential practitioners. Harris et al. (2004) provided an appendix with an initial attempt to document how vertical scales were operationally implemented by various publishers for their testing programs.

13.2. Conceptual Issues

The tendency is to jump into methodology immediately, but the conceptual issues really need to be considered first, both to ensure that there really is a need for a vertical scale and because the decisions made up front have tremendous impact on the resulting scales.

13.2.1. Do You Really Need a Vertical Scale?

The first issue to resolve is the actual need for a vertical scale. For example, if one is a grade-school administrator who wants to ensure that all graduating sixth graders know the capitols of all 50 states, there is no need for a vertical scale. All students can be given the same test, and raw scores can be used to monitor progress over time. However, for subjects where knowledge acquisition is gradual, or follows a sequence, moving students to where one wants them to end up is more of a process. For example, if one wants students to be able to multiply three-digit numbers, repeatedly testing on multiplying three-digit numbers is not really effective. Instead, one wants to monitor (know) if they know their basic multiplication facts, if they can multiply and carry, and so on. Administering the "final" test content at an earlier grade will not really enable one to target effective instruction. However, having a scale, or sequence, that follows the process from, say, numeral recognition through three-digit multiplication would allow one to monitor progress and provide intervention where needed. A vertical scale could be helpful for the later situation.

A vertical scale, therefore, is not the only option. A scale might not be needed (i.e., the raw score scale might be sufficient) or other options might be preferable to a vertical scale. For example, Lissitz and Huynh (2003) advocated vertically moderated standards as being more useful than vertical scales in assessing adequate yearly progress.

However, as mentioned earlier, a vertical scale is often helpful in guiding students along a continuum. As an example that will be followed throughout this chapter, consider the mathematics tests in the ACT and PLAN programs. They have a common philosophical basis in measuring students' knowledge and skills typically attained during a student's secondary school experience. The ACT is intended primarily for 11th and 12th graders; PLAN is primarily intended for 10th graders. It was determined that placing the two tests on the same scale would facilitate the goal of providing a longitudinal approach to educational planning, assessment, instructional support, and evaluation. See the PLAN and ACT technical manuals (ACT, 1997a, 1999) for details regarding the use of the PLAN/ACT scale.

13.2.2. Developing Test Specifications

Issues such as what grades to include in the assessment, what content to cover, what item types to use, what time limits, who is writing the items, and so on can have a large impact on the resulting scale. How content is defined across the grades (i.e., the amount of overlapping content in, say, the third and fourth grades) has a major impact on the resulting score scale.

Issues such as how to model grade-to-grade overlap depends, in part, on how the assessment structures content across grades. Kolen (2003) listed "Over what test content should grade-to-grade growth be defined?" (p. 6) as an issue in need of further study, illustrating the relationship between test content and the nature of growth.

Issues such as balancing completeness of coverage with motivation and frustration issues of administering too many items of inappropriate difficulty or interest to examinees in a given grade, deciding how many grade levels should receive particular items, the number of concepts that overlap, and so on are philosophical as well as practical or measurement issues. Construct dimensionality issues are also partially embedded in the nature of growth. The importance of content dimensionality in establishing vertical scales continues to be an issue.

The content specifications for the ACT and PLAN mathematics tests, taken from ACT (1999) are shown in Tables 13.1 and 13.2.

The test specifications make concrete some of the assumptions regarding growth concrete. For example, the inclusion of plane geometry in both PLAN and ACT specifications indicates this is a topic that one expects to be covered at both levels, whereas trigonometry is not. The more detailed specifications that are actually used for forms construction (sublevels of topics within the broader area of plane geometry), as well as statistical specifications, such as average or target *p*-values, would indicate how the progression of plane geometry across levels is thought to occur. For example, topics intended for the PLAN assessment might be more difficult for 10th graders than for 12th graders, and more advanced topics might be included on the ACT and not included on the PLAN.

Table 13.1. Specifications for the ACT mathematics test

Content area	Proportion of test	No. of items
Pre-Algebra[a]	.23	14
Elementary Algebra[b]	.17	10
Intermediate Algebra[c]	.15	9
Coordinate Geometry[d]	.15	9
Plane Geometry[e]	.23	14
Trigonometry[f]	.07	4
Total	1.00	60

[a] Pre-Algebra. Items in this content area are based on operations using whole numbers, decimals, fractions, and integers; place value; square roots and approximations; the concept of exponents; scientific notation; factors; ratio, proportion, and percent; linear equations in one variable; absolute value and ordering numbers by value; elementary counting techniques and simple probability; data collection, representation, and interpretation; and understanding simple descriptive statistics.

[b] Elementary Algebra. Items in this content area are based on properties of exponents and square roots, evaluation of algebraic expressions through substitution, using variables to express functional relationships, understanding algebraic operations, and the solution of quadratic equations by factoring.

[c] Intermediate Algebra**.** Items in this content area are based on an understanding of the quadratic formula, rational and radical expressions, absolute value equations and inequalities, sequences and patterns, systems of equations, quadratic inequalities, functions, modeling, matrices, roots of polynomials, and complex numbers.

[d] Coordinate Geometry. Items in this content area are based on graphing and the relations between equations and graphs, including points, lines, polynomials, circles, and other curves; graphing inequalities; slope; parallel and perpendicular lines; distance; midpoints; and conics.

[e] Plane Geometry. Items in this content area are based on the properties and relations of plane figures, including angles and relations among perpendicular and parallel lines; properties of circles, triangles, rectangles, parallelograms, and trapezoids, transformations, the concept of proof and proof techniques volume; and applications of geometry to three dimensions.

[f] Trigonometry. Items in this content area are based on understanding trigonometric relations in right triangles; values and properties of trigonometric functions; graphing trigonometric functions; modeling using trigonometric functions; use of trigonometric identities; and solving trigonometric equations.

Table 13.2. Specifications for the PLAN mathematics test

Content area	Proportion of test	No. of items
Pre-Algebra[a]	.35	14
Elementary Algebra[b]	.20	8
Coordinate Geometry[c]	.18	7
Plane Geometry[d]	.27	11
Total	1.00	40

[a] Pre-Algebra. Items in this category are based on operations with whole numbers, integers, decimals, and fractions. The topics covered include prime factorization, comparison of fractions, conversions, scientific notation, square roots, percent, absolute probability, mean, median, and mode.

[b] Elementary Algebra. The items in this category are based on operations with algebraic expressions. The operations include evaluation of algebraic expressions by substitution; simplification of algebraic expressions, additions, subtraction and multiplication of polynomials; factorization of polynomials; and solution of quadratic equations by factoring.

[c] Coordinate Geometry. Items in this category cover topics on graphing in the standard coordinate plane. The topics include graphs of linear equations, measurement of lines, and determination of the slope of a line.

[d] Plane Geometry. Items in this category cover such topics as measurement of plane surfaces, properties of polygons, properties of triangles, the Pythagorean Theorem, and relationships involving circles.

Vertical scales are often created after test forms for different levels are created. It should be understood that the nature of the forms themselves—in particular, their content and statistical specifications in relation to each other—has a large impact on any resulting vertical scale, including ceiling and floor effects, and the amount of overlap between different levels.

13.2.3. How Is Growth Defined?

Perhaps the most publicized debate in the vertical scaling literature is in relation to using item response theory (IRT) as a scaling method. A key issue in the debate over scale shrinkage was the nature of growth and whether within-grade variance should increase, decrease, or remain constant as the grade increased. Camilli (1988) stated.

> The scale shrinkage controversy has opened up an important debate in educational measurement. It is not a debate between IRT methods and traditional scaling methods. In fact, it was shown in this paper that the two types of methods (IRT and percentage correct scores) could lead to similar conclusions about shrinkage within grades. The more interesting issue raised is how children learn, and this question goes far beyond measurement technology. (pp. 239–240).

The primary reason for creating vertical scales is to measure learning across time. Without an understanding of the nature of growth, it is not possible to clearly evaluate whether a vertical scale is functioning as it should. For example, if the true nature of growth shows increasing variability over time, then a vertical scale that shows constant variability over time would not be judged as adequate. These issues are philosophical and deal with child development, psychology, and how the educational curriculum is implemented. The pattern of growth might vary across grades (i.e., increase from, say, first to fourth grade, then remain constant) and across academic subjects (the nature of mathematics might yield a different growth pattern than, say, English; punctuation might be different from comprehension). Different ways of constructing and implementing a curriculum might also impact growth across time. A spiral curriculum (where a concept is covered at multiple points in time, at increasing depth) might yield gradual growth, whereas a different implementation might yield a more stair-step pattern of growth. Additionally, how one chooses to assess growth will have an impact, as growth is generally operationally defined by some assessment tool.

In addition, there is the interaction with test construction/design. Should specifications be developed to meet a preexisting growth model or should the model of growth be developed based on empirical information obtained from an assessment built to a philosophy of curriculum? Given that results will differ depending on choice of particular practices, scaling methods, assessment forms, and so forth, how does one decide what to do? For example, Harris and Hoover (1987) found that examinees received higher ability estimates if the test level they were administered was calibrated on less able examinees. How should this information be used in selecting procedures? Could findings like this be manipulated for advantage? Or, are aspects of these issues somewhat irrelevant to most practitioners, as Yen and Burket (1997) suggested, as long as most comparisons tend to be within a grade, using the same instrument (e.g., fourth graders administered the ITBS are compared to other fourth graders administered the ITBS)?

One problem in trying to address the issue of defining growth is that test publishers rarely make the information explicit. It seems likely that most definitions are determined operationally, based on a combination of empirical data, the test development process, and preconceptions regarding the nature of growth. For example, a practitioner who believes within-grade variance should remain constant over grades might not develop test specifications or a data collection design with this in mind, but might reject scaling methods that resulted in large changes in within-grade variance over grades.

In our example, the nature of growth for the PLAN and ACT mathematics scale was determined using two main sources of information: curriculum surveys, content experts, and educators, and empirical data. The former were used to develop the test specifications, which included the content covered on both assessments and the targeted difficulty and complexity of the content. Empirical data were then used to operationally define, for example, within-grade variability.

13.3. Technical Issues

The separation of technical and implementation issues followed here is admittedly arbitrary. The intent is to separate the decision to use, say, the three-parameter logistic model in scaling from the particular choice of estimation program used to estimate item parameters.

When initially developing a score scale, decisions need to be made as to the number of score points, how the scale will be anchored, how vertically scaled levels are mapped into the score scale, how equated raw scores or thetas are mapped onto the scale (linearly? normalized? arcsine transformation?), and how, where, or if gaps or clumping (multiple raw scores mapping into a single reported score) occur. Is the scale to have a target mean and standard deviation for a particular population or sample? Are the scale values integers? Two digits? Are the values chosen likely to be confused with some other scale? Does the scale aid in score interpretation or detract from it? Most scales are not equal interval, despite some claims to the contrary. Is this clear to users? What is the best scale to measure growth?

Yen and Burket (1997) discussed the need for criteria regarding what makes for a desirable scale. Even if we could define the gold standard in terms of what characteristics a good scale should have, we are still left with the problem of how to obtain these properties. How do we manipulate the results obtained from some objective set of procedures and software? Do we smooth? If so, how much and with what method? What are the ideal characteristics that a scale should possess? Tomkowicz and Schaeffer (2002) provided a case study into manipulating results to obtain a final scale with what they viewed as desirable scale characteristics. How much subjective manipulation is acceptable? And as there really is little that is objective about the choice of software to use, the methods to use, the data to use, and so forth, does it really matter?

13.3.1. Data Collection Design

One of the most obvious choices in collecting data to scale a test is the data collection design (see Kolen, Chapter 3). According to Kolen and Brennan (2004), “It needs to be made explicit that the differences between the grade-equivalent scales of test publishers lie mainly in the method of data collection (e.g., scaling test versus anchor test), not in the statistical method used to link the test levels” (p. 235).

Different data collection designs can be used to create vertical scales, such as scaling test, common items, or single group to scaling test (a separate test containing, for example, both third and fourth-grade items, which is administered to both third- and fourth-grade students in addition to the regular third- or fourth-grade test, respectively), common items (a set of anchor items appearing, e.g., in both the third- and the fourth-grade tests), or single group (where one group of students is administered, e.g., both a third-grade test and a fourth-grade test). Common items can be internal or external, they can span the entire range of content and difficulty, or any subset of the range. The number of items required to provide adequate linking using common items has not been determined. What characteristics the sample of examinees need to display has not been determined, nor has the number of examinees required for vertical scaling been agreed upon. No general rules exist in terms of how to edit items or data; there is no consensus on how to use goodness-of-fit indexes in determining whether to retain items or examinees in establishing vertical scales. No single combination of methodology, data collection design, and sample has been found to be superior to others to a generalizable extent, and most designs seem to work well in at least some of the settings studied.

It should be noted that the way a design is implemented also can vary. For example, a scaling test can cover the full range of, say, grade 3 to grade 8, or there can be two scaling tests that cover, say, grades 3 to 6 and 5 to 8, and so on. In addition, some common-item designs are implemented with overlap to both a higher and lower level (e.g., grade 5 overlaps with both grade 6 and grade 4) or to only a lower level (e.g., grade 5 only overlaps with grade 4). At times, two distinct designs (e.g., scaling test and common item) on a particular battery might have more in common than the same method (e.g., common item, across two batteries). For example, the common-item design used in Boughton, Lorie, and Yao (2005), where the common items are scattered throughout a test form and the linking is one-directional in that a grade 5 test also includes grade 4 items, but a grade 4 test does not include grade 5 items, is very different from the common-item design used in Tong (2005) and Hendrickson, Wei, Kolen, and Tong (2005) where the common items are concentrated at the ends of the test forms, and tests overlap with both the next higher and the next lower grade (i.e., a grade 5 test has both grade 4 and grade 6 items). (It should be noted that the determination of the common item pattern is not just a data collection issue, it is also impacted by the test specifications and the nature of growth.)

Practical issues such as testing time and the nature of the items (it is difficult to have common-item designs with some types of passage-based item or constructed response item) as well as the nature and number of examinees available for scaling also impact how data are collected. Crocker and Algina (1986) stated that the these sorts of practical issue are often the "prime criterion" in selecting a data collection design for equating, but that the main criteria should be the tenability of the design assumptions, practicality, and accuracy. This is likely to also hold for vertical scaling.

Hendrickson, Kolen, and Tong (2004) found an interaction between scaling design (common item vs. scaling test) and calibration procedure; Loyd and Plake (1987) also found that design can have a substantial influence on the results. Andrews (1995) found that score scales developed with different methods and different designs differed enough to consider scaling design as an "important factor" when creating a vertical scale.

Raju, Edwards, and Osberg (1983) examined the effect of anchor-test size in vertical scaling with Rasch and 3PL and found that shorter anchors (as few as six items) could be as effective as longer ones. Barron and Hoover (2001) found context effects to be problematic in using common items to create a vertical scale. Harris (1991) found that although both designs appeared adequate, Angoff's Design 2 (counterbalanced, single group design) exhibited more stability than Angoff's Design 1 (random groups design). Kolen (Chapter 3) provided a current updated description of Angoff's designs. Holmes (1982) compared a single group method and two external anchor common-item methods and found that the single group method consistently produced the most accurate results, although the advantage was small.

Various operational vertical scales have been established using different data collection designs. The Stanford Achievement Test Series (Harcourt Educational Measurement; 1985) and Metropolitan Achievement Tests (The Psychological Corporation, 1988) used a single group design variant: Each student was administered two adjacent levels. The Mississippi Curriculum Test (Tomkowicz & Schaeffer, 2002) used internal anchor items to link to the TerraNova K-12 assessment system. The Iowa Tests of Basic Skills (Iowa Tests of Basic Skills, 2003) used a scaling test design.

In our example, the goal of the scaling was to place PLAN scores on the existing ACT score scale. Data from a random groups design were used as the primary data (12th graders were randomly administered the ACT or the

PLAN), with data from a random groups design (10th graders were randomly administered the ACT or the PLAN) and two single group designs (11th graders and 12th graders administered both ACT and PLAN, in a counterbalanced order) used to evaluate potential scales and to check assumptions of the scaling.

13.3.2. Scaling Methods

Different methods of developing scores include normatively, Guttman scaling (which might be unrealistic in practice; see Kolen & Brennan, 2004, who suggested that the probabilistic approach is more likely to be appropriate in practice than the deterministic approach), Thurstone scaling, Hieronymus scaling, and IRT scaling. Kolen and Brennan, and Petersen et al. (1989) discussed these methods, and scale construction in general, including linear and nonlinear transformations, creating scales that incorporate content meaning, or normative meaning, or score precision information, as well as developmental score scales such as grade equivalents. The PLAN and ACT mathematics tests were scaled using an equal-standard-error-of-measurement property (see ACT, 1989).

13.3.3. Reported Scale

Reported scores are generally integer scores or decimal scores rounded to a preset number of decimal places. When using IRT methodology, it would be possible to report ability estimates such as thetas or logits, rather than scores, although it generally is not done. It is assumed that examinees and general users of test results would have difficulty interpreting estimated theta or logit values. Commonly, some underlying scale is developed as a result of the scaling method, which is then transformed in some way to a reported scale. This can involve linear or nonlinear transformations, truncation, extrapolation, and rounding.

Numerous examples of different types of scales exist. For example, Angoff (1971) listed percent mastery, standard scores, percentile ranks, normalized standard scores, age-equivalent scores, grade-equivalent scores, and IQ scores. Petersen et al. (1989) discussed having primary and secondary scales. They advocated creating reported scales that facilitate score meaning and minimize likely score misinterpretations, such as being confused with another score scale that already exists.

Kolen and Brennan (2004) provided additional examples of scales based on psychometric models, including Thurstone and Rasch; domain scores are also discussed. Additional issues include how to compute raw scores on a test (e.g., number correct, pattern scoring, corrected for guessing) and how to scale tests within a test battery. (Should all tests be scaled the same way and/or have the same range of score values, even if the range is not optimal for all tests within the battery?) Should estimated true scores be used? For multiple-choice tests, should scores below the “chance” level be truncated? If you use normative information in creating a scale, on what group should the norms be based? If you use an equal standard error of measurement (SEM) method, what reliability values should be used? For example, if the number of achievable scores differs for math and language arts, do we want the same number of reported scale score points? (Kolen & Brennan, p. 345, suggested some ways to determine a reasonable number of score points.)

One important issue with some constructed response items is that raw scores for a prompt generally have meaning based on the scoring rubric. Depending on how those scores are combined with other items and then transformed into a reported score, this direct meaning might be lost.

In our example, the reported scale for PLAN is 1-32 on the 1–36 ACT scale. Because the PLAN assessment does not contain the more difficult items that the ACT assessment does, it was determined that the maximum scale score achievable for PLAN should be less than the maximum score achievable for ACT. A top of 32 was arrived at empirically, from examining data, test specifications, and scale characteristics.

13.3.4. Criteria

What are meaningful ways to compare different vertical scales resulting from different methodologies? What criteria do we use? Effect sizes? Heuristics/common sense? Is there some objective measure that could be applied, such as the reliability of gain scores on the scale, or empirical studies involving multiple test forms and multiple occasions? How do we determine if one scale is better than another or if a particular scale is acceptable? One very important and neglected area is how to evaluate if a scaling is acceptable or best.

Harris and Crouse (1993) summarized the various criteria that have been applied in equating studies and gave an example of how different criteria change the resulting decision on what is best equating; something similar should be done for vertical scaling.

Arce-Ferrer, Frisbie, and Kolen (2002) used the standard error of proportions in reporting changes in school performance with achievement levels. Holland (2002) proposed two measures of distance to examine the difference between two cumulative distribution functions: the vertical (difference in percent at the same score) and horizontal (difference in percentiles for the same percentage) distances. Tong and Kolen (2005) used effect sizes. Other studies have used cross-validation (e.g., Holmes, 1982), "reasonableness," such as grade-to-grade growth (e.g., Karkee, Lewis, Hoskens, Yao, & Haug, 2003), and first-order equity (e.g., Harris, 1991). Simulations have been used, but as the data are simulated to fit a particular model, recovery of "truth" might be a more useful criterion for examining issues such as the effect of concurrent or separate calibrations than in evaluating the resulting scales themselves. Yen (1986) argued that "clearer criteria are needed for judging the appropriateness and usefulness of alternative scaling procedures and more information is needed about the qualities of the different scales that are available" (p. 299).

Criteria need to be determined that will be generally accepted as a way to evaluate the acceptability of a vertical scale. Two primary criteria were used in evaluating placing PLAN on the existing ACT scale: how closely the same-scale property was met (meaning an obtained PLAN scale score can be interpreted as approximately the ACT scale score that an examinee would have obtained if he/she has taken the ACT at the same time that the PLAN was taken) and how equal the conditional SEM was across the score scale range. Other factors, such as gaps in the reported scale, were also considered.

13.4. Implementation Issues

Many issues arise in the construction of vertical scales, which might be loosely grouped under the umbrella of "technical issues." These include scale indeterminacy, calibration method (concurrent, separate, etc.), choice of item parameter linking (mean-sigma, a curvilinear method, etc.) for placing separate item parameter calibrations on the same scale, choice of model (classical, IRT, testlet, polytomous, number of parameters, etc.), choice of item parameter estimation procedure, and so on. Much of the vertical scaling literature that does exist compares and contrasts scales created using different technical methods. However, there is no definitive comparison study (it is unlikely that there could be), and the practitioner does not have any unequivocal guidelines to follow.

There are a multitude of methodologies and variations on these methodologies that can be used to create vertical scales. If an assessment includes both constructed response and multiple-choice types of items, they might be scaled in a single calibration run, or scaled separately and combined later. Examinee raw scores might be computed by using a number correct score, a corrected-for-guessing score, or an IRT-based score (typically, theta). Different items or contents or sections can be weighted differentially, and combined in various ways, to form raw scores.

Item calibrations can be conducted concurrently, or separately. Fixed item parameters can be used, or various item parameter linking procedures, such as item characteristic curve methods or mean-sigma, have been used to place item parameter estimates from separate calibrations on a common scale. Different approaches exist to chain different calibration runs together. Different "bases" can be used, such as scaling through a calibrated item pool or a base form approach. For, say, a K-8 battery, any grade test from K to 8 could be used as the base form to create the scale. No single combination of methodology, data collection design, and sample has been found to be superior to others to a generalizable extent; most designs seem to work well in at least some of the settings studied.

New, innovative methods are also being explored, such as the hierarchical and multivariate modeling approaches discussed in Patz, Yao, Chi, Lewis, and Hoskens (2003) and Patz and Yao (Chapter 14). The hierarchical multigroup method allows the functional form of growth to be explicitly estimated, whereas the multidimensional multigroup model can consider the dimensionality differences that occur at different levels. Although the authors presented these models as exploratory, it is clear that they address some additional issues related to vertical scaling that bear further research.

Research summaries should be created (along the lines of meta-analyses?) to summarize when particular methods appear to work well. Research comparing the results of applying different combinations of methods should be continued. One of the best exchanges I am aware of were the IRT versus classical scaling exchanges: There were IRT advocates implementing classical methods and classical advocates implementing IRT methods, different data, different implementation decisions, inconsistent results, and so on. It was a relatively open exchange of impact (results of the two approaches) and we all benefited from it. For the PLAN–ACT example, details, including the strong true-score model used, specifics regarding the examinee samples, the formulas used in computing the SEM and the same-scale property are provided in ACT (1999). Note that not all operational vertical scalings are this well documented in the public domain.

One implementation issue that is especially important is the choice of software. Although some vertical scaling can be done by hand, virtually all research and operational scaling makes use of computer programs. Most software programs make numerous options available, although many users likely implement only default settings. Although programs certainly differ in the extent of documentation and the ease of implementing alternatives, users frequently lack the knowledge to make an informed decision. For example, a smoothing program might offer degrees of .05 and .10, as defaults, yet provide no guidance to the user for determining which of these would be a better alternative. Some programs provide limited information as to what algorithms are used, how to interpret output, and how truncation, interpolation, extrapolation, smoothing, and so on are handled, which can impact the final reported scale values.

Perhaps one of the less considered decisions is that of which IRT calibration program to use. Several authors found the particular software used for IRT parameter estimation could have an impact. In addition to the more obvious estimation method differences, (e.g., Hendrickson et al., 2004, looked at three IRT proficiency estimation procedures: expected a posteriori [EAP], maximum a posteriori [MAP], and maximum likelihood estimate [MLE], even issues as subtle as the number of EM cycles in BilogMG or whether to use default settings might have an impact.

Programs are complex and the manuals are often obscure about what computations are actually being done, and for proprietary reasons, source code is generally not available. When a publisher uses a program developed in-house, there is generally even less information about the program made available, making it difficult to know the effect of the program (what options were used, how calculations were done, etc.) on the final scale. One solution is for the test developer to do comparison studies, although, admittedly, a case could be made that a disinterested party would be preferred. Fitzpatrick (1994), for example, compared parameter estimates produced by PARDUX and BIGSTEPS.

Way, Twing, and Ansley (1988) compared Bilog and Logist using two different calibration procedures, as did Omar and Hoover (1997). Omar (1997) followed up on the previous study, examining BilogMG. Childs and Chen (1999) described obtaining comparable item parameter estimates from MULTILOG and PARSCALE. Pomplun, Omar, and Custer (2004) compared WINSTEPS and BilogMG, finding that WINSTEPS tended to result in more accurate individual and mean measurement, whereas BilogMG resulted in more accurate standard deviations. Hendrickson et al. (2004) compared MULTILOG and ICL and found that the computer program/estimation method used impacted the resulting vertical scale.

Limitations, such as the number of categories allowed for polytomous items, or the size of a data matrix that can be input, might also affect the final vertical scales, as they require collapsing of data categories or the winnowing of data. Bishop and Omar (2002) mentioned that in their study, a number of decisions had to be made, such as collapsing categories of data, because of limitations in the software used. Writing one's own programs might eliminate this problem, but this leads to the issue of potential lack of comparability with other investigators, making consistencies and inconsistencies in different methods of scaling, and so forth more difficult to discern.

Most of the studies reported in the literature do not provide much detail on how computer runs were conducted, although some exceptions exist (e.g., Jodoin, Keller, & Swaminathan, 2003, provided information on the optional commands they used). Proprietary software was used in the scaling of PLAN and ACT; information on the algorithms used is provided in ACT (1999).

To summarize, there is no clear guidance to a practitioner on what software to use in vertical scaling. When new versions of software appear, it is up to the practitioner to determine, for example, if parameter estimates calculated under the new and old versions are comparable. It is suggested that more use be made of open-source software, where, for good and bad, how calculations are done is publicly available.

13.5. Scale Maintenance Issues

One issue that has not been addressed much in the literature is that of maintaining vertical scales over time and over new forms. For example, should new grade 3 forms be equated to the original grade 3 form or should there be an attempt to link the entire range of, say, grade K to grade 8 forms to the original set of forms on which the scale was set? What types of drift, or error, are we apt to see over time? How often should a vertical scale be monitored? Reevaluated? Reconstructed? Because of the different results that different procedures have lead to, what are the dangers of "mixing and matching" procedures over time? Also, what is the trade-off between what is practically possible and what is best from a consistency standpoint?

Issues such as data collection designs, equating methodologies, and examinee sample characteristics need to be considered in equating new forms to a vertical scale (see Kolen, Chapter 3 for additional discussion of these issues). How equating is defined, whether by Lord's (1980) equity definition, Angoff's (1971) equipercentile definition, Divgi's (1981) two approaches, Morris' (1982) method including conditional variance, an IRT true-score definition, or some other definition, should guide the equating of new forms (see Harris & Crouse, 1993). A choice of equating methodology needs to be made, which might or might not correspond to the methodology used to scale. For example, IRT could be used to create the vertical scale, but classical methods, such as equipercentile methods, could be used to equate new forms. However, if the assessment is constructed using IRT procedures, equating (and scaling) the test using IRT could take advantage of the test development procedures.

An equating is always referenced to a particular population of examinees (Flanagan, 1964). The data collection design/samples combination is the most important part of any equating study (assuming, of course, that the characteristics of the instruments make equating defensible). No equating methodology exists that can counteract bad data. One of the most important sample characteristics (in addition to size of the sample, motivation, and appropriateness for the test being equated) is that the sample be representative of the population in which one is interested.

There is no easy mechanism to apply to determine which equating method is preferable in any given situation. Additionally, there is no universally accepted criterion to know if an equating is best or even acceptable. When the new forms are part of a vertical scale, the issues are much more complex. Also, whether, say, new forms at one level are equated separately from new forms at a different level depends in part on how new forms are introduced. For example, if a new battery is introduced at a single point in time, and not very frequently, equating the new forms to the previous scale simultaneously might be done. However, if new forms are introduced frequently, and at different times, equating the forms separately is more practical.

Hoskens, Lewis, and Patz (2003) looked at maintaining a vertical scale over time, examining several approaches, including equating within each grade, an augmented approach that used both vertical and horizontal anchors, and a concurrent and a separate method of setting a new vertical scale for all grades concurrently and linking it to the previous vertical scale. They found that the method chosen had an impact, with the horizontal and augmented methods indicating grade-to-grade variability was relatively flat, and the other methods indicating an increase in variability.

There are additional practical issues that might also affect the stability of scales, such as changes in software used to calibrate items (e.g., a change from Bilog to BilogMG) or a change in a vendor (e.g., when a state department moves its test development from one testing company to another).

In our example, new forms of the PLAN and ACT mathematics assessments are equated. The stability of the PLAN-ACT scale over time was checked in 1995, using a scaling test design. Both the original methodology (equal SEM method) and IRT methodology were used to create PLAN scores, which were then compared to the existing PLAN scores. The resulting scales were somewhat different, which was expected because of the different design (there was a test length adjustment used in the 1988 scaling, as well as a difference from a random groups to a scaling test design) and a slight change in the test specifications for PLAN between the form used for the 1988 scaling and the form used in the 1995 scale. It was determined, however, that the differences were not compelling. It should be noted this was not a traditional examination of scale drift, but a comparison of an entirely new scale created under different circumstances, to the original scale. A change in administration policy to allow the use of calculators on the mathematics assessments in 1996 also led to a reexamination of the PLAN-ACT scale.

13.6. Other Issues

Issues that arise in other contexts (e.g., single-grade-level testing) might be magnified in a vertical scaling context. Some of these issues were mentioned earlier, such as content dimensionality issues, but other issues, such as moving a paper-and-pencil test to a computer-based test, were not. (See Eignor, Chapter 8, for a discussion of issues related to moving from a paper mode to a computer mode.) The fact that multiple levels need to be considered simultaneously increases the complexity of dealing with issues such as these. Issues arise when not all examinees answer all items, whether from a matrix sampling design or from an examinee choice model, where an examinee chooses, for example, which two questions to answer of the five questions available. These issues become more complex when the consistency of scores needs to be maintained vertically (across grades) as well as horizontally (within a grade). This is also true for issues such as modifications in test specifications, conducting standard settings (assuming there is a desire for continuity across grades), translating the test into other languages, preequating test forms, pretesting items, and dealing with test speededness and guessing issues. Technical issues, such as establishing validity for score use or computing reliability coefficients, as well as operational issues such as training raters to grade essay responses are more problematic in a vertical scaling context. Although it is possible to establish a scale initially and then subsequently treat each grade separately, there still needs to be monitoring across the entire range of grades to ensure reasonableness (e.g., that a cutoff for adequate performance is not set at a score of 130 for grade 3 and at 120 for grade 4).

Although these issues might (e.g., dimensionality) or might not (e.g., translation issues) directly impact the setting of the vertical scale, they all might impact the usefulness of the scale as it is put into operational use.

13.7. Summary

This chapter presents issues that a practitioner would encounter when developing a vertical scale for an operational testing program: using a framework of conceptual issues, technical issues, implementation issues, maintenance issues, and other issues. The scaling of the PLAN and ACT mathematics tests is used as an example to demonstrate some of the issues underlying vertical scales. The practitioner who chooses methods with careful attention to his/her purpose of the scale (i.e., how the resulting scores are intended to be used) and to the literature and current practices of other test publishers is more likely to create a scale that will lead to scores on which appropriate decisions can be made.

Vertical scaling is a complex process, involving philosophical, technical, and practical issues. Although it can be disconcerting that there is no consensus on the best way to create a scale, it is also comforting. Many assessments, such as ITBS (Iowa Test of Basic Skills, 2003), Stanford Achievement Test (Harcourt Educational Measurement, 1985), and EPAS (ACT, 2000), state-specific tests, and so on, have created vertical scales in different ways, yet all of those scales appear to be functioning adequately for some of the same purposes. Perhaps there are many roads to Rome. However, that does not mean that all roads lead to Rome, or that all implementations of vertical scaling lead to acceptable scales for all purposes. Instead of arguing which single scaling method is the best, we might do better to see which slate of options work for which purposes, under which conditions.