Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 8, Number 10, April, 2003 ISSN=1531-7714

Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability

Robert W. Lissitz, University of Maryland
Huynh Huynh, University of South Carolina

(The authors contributed equally to the paper and their names were listed randomly.)

Of all the provisions of the federal No Child Left Behind (NCLB) legislation, the definition and determination of adequate yearly progress (AYP) is perhaps the most challenging. NCLB requires states to administer reading and mathematics assessments at least once each year to students in grades 3 through 8 (and once more within grades 10-12) by 2005-06, and adds a science assessment administered at least once in each of three grade spans by 2007-08. States may select their own assessments and define their own proficiency levels, but they must submit plans to the U.S. Department of Education to establish goals for what percentages of students in various subgroups (e.g., low income, minority, limited English proficient) will meet or exceed proficiency levels on the state's assessments each year.

This article describes AYP and some of the psychometric issues it raises. It examines scaling as a means to equate tests as part of a process to confirm educational gains. Vertically moderated standards are recommended over vertical equating of state assessments to measure annual progress and provide useful instructional information.

ADEQUATE YEARLY PROGRESS

In a paper titled "Making Valid and Reliable Decisions in Determining Adequate Yearly Progress" (Marion et al., 2002), the Council of Chief State School Officers summarizes AYP as follows:

> *Each of at least 9 subgroups of students must reach proficient or advanced achievement levels in reading or language arts and mathematics by 2013-2014 (Uniform progress is required beginning in 2002-03.) AYP determinations are based solely on student achievement results on State assessments. At least 95% of the students in each subgroup must participate in the assessments and all must meet the State's performance target in another academic indicator as prescribed by the law* (p. 5).

Further,

> *The NCLB Act requires States to determine the number of students in a group necessary to yield statistically reliable information as well as the number of students required to be in a group to ensure that the results will not reveal personally identifiable information about an individual student* (p. 12).

To briefly summarize the challenge, NCLB requires states to develop a system that tracks students' (by defined subgroups) success in reading/language arts and mathematics (with science coming on board soon) as the students progress through school, and the data associated with these adequately sized subgroups must show at least minimum levels of proficiency. Some additional indicators must be provided, as well, but the primary focus will be on the determination of proficiency and the success of most students over the time span of schooling. The purpose of the AYP is to allow the state to monitor progress and to identify problem schools, and low performing subgroups and to prescribe remediation that will result in No Child Left Behind.

PSYCHOMETRIC ISSUES RELATED TO AYP DETERMINATION

Since test scores are going to be the major source for determining student progress and school accountability under NCLB, it is critical that test scores be comparable from test to test and year to year. Scaling is a measurement technique that can facilitate test score comparability.

Description of Scaling

A scaling process, in general terms, is one in which raw scores (usually calculated as the total number of correct responses) are transformed to a new set of numbers with certain selected attributes, such as a particular mean and standard deviation. For example, the Scholastic Aptitude Test has scores that range from 200 to 800 and result from a scaling process that transforms the number correct score that a student has obtained. Some scaling procedures are non-linear transformations of the raw scores and some are linear. The particular approach used depends upon the purpose of the scaling and the properties that we want in the resulting scale.

One of the most common purposes of scaling has to do with equating two or more tests. The tests to be equated might be given at different times, so that the purpose of the scaling would be to arrive at comparable scores for tests across time. The tests might be given to different groups, as well. The most common application for scaling involves equating different forms of the same test. In any case, the rescaling of the students' raw score performance level has the following advantages:

- Regardless of changes in the test from year to year, the scores reported to the public are always on the same scale. This makes it easier for teachers and principals, as well as students and parents, to learn to interpret the results of testing.
- If several related tests need to be available for use, transforming each one to the same scale allows them all to be interpreted in a similar way. Again, this helps the problem of communication of test results.
- Equal raw scores from different forms will not usually express the same amount of ability because one form might be easy but the other form might be more difficult. Scaling allows us to "equate" the two forms for purposes of reporting.

Two primary situations exist for scaling multiple sets of tests to a common scale or equating them. Horizontal equating is designed to test different groups of students that are assumed to be at approximately the same level. It occurs within grade, where multiple forms are used to test the same general content. Vertical equating may be used when testing students who are at different levels of education. It entails across-grade testing of the same general content. Each type is discussed in more detail below.

Within-Grade (Horizontal) Scaling. An example of the horizontal equating situation is the case in which a school system has a test for graduation and students are allowed to retake the test if they fail. The retakes are on different forms of the test that are all equated to provide comparable scores. The cut-off for failing is set at the same scale score level no matter how often a student retakes the test, thus ensuring a constancy of the standard for passing from administration to administration. The table of specifications (i.e., the test blueprint) for each test is also the same, thus ensuring that content is comparable and that the dimensions of knowledge that underlie the test are the same in each case. The difficulty level will be approximately the same for each form of the test, as well. Occasionally, horizontal equating is used to allow for comparison of groups of students that are different in some fundamental way that requires modification of one form of the test. For example, comparisons of recent immigrant students who speak only Spanish with those who are fluent in English will require tests that are equated, yet differ in the language of the test items.

Across-Grade (Vertical) Scaling. One of the common ways that psychometricians have approached the AYP problem is to develop a single (unidimensional) scale that summarizes the achievement of students. This scale is then used to directly compare the performance level across grade levels. For example, TerraNova K-12 (CTB/McGraw-Hill, 1997, 2001), the Stanford Achievement Test from Harcourt (1996), and the recent work in Mississippi (Tomkowicz and Schaeffer, 2002) present scales that are purported to allow for the meaningful, continuous, tracking of students across grades.

A classic example of the vertical equating situation is that of a test of mathematics that is used to track expertise across middle school. In this scenario, the tests at different grade levels are of differing content, but still focus on the same general concept, say, mathematics fluency. The students are expected to show performance improvements at each year, and these improvements should be reflected in a steady increase in their ability to do mathematics. The tests for grades 7 and 8 should be linked so that scores are directly comparable along a common continuum or dimension. Sometimes this approach is used for tests of literacy, as well.

The content must have some sense of commonality across grades in order to be meaningfully equated across grade levels. These scales are often considered developmental, in the sense that they encourage the examination of changes in a student's score across grades that indicate the improvement in that student's competency level. Sometimes the equating is only for adjacent grades and sometimes equating is across the whole school experience.

Major Assumptions for Horizontal and Vertical Scaling

The major assumption for equating is that the tests are assessing the same general content. In other words, a psychometric model will be appropriate for each test being scaled or equated and it will develop the modeling relating the two tests using a single or a common set of dimensions. In the case of horizontal scaling, this is not usually a problem. Since each form of the test is designed to examine the same curriculum material, a model that works for one test will usually work for all the forms of that test. Naturally, we are assuming that the tests are not only designed with the same table of specifications, but are using the same mix of test item types. For example, each test would have approximately the same mixture of performance items and selected response items. The English language demands would be about the same, as well. In situations such as these, we have had considerable success with modeling and achieving quite accurate equating (i.e., successful scaling).

Vertical scaling has the same assumption of comparable content. It is assumed that the same basic dimension or dimensions are being assessed in each grade for which we are developing a test to be equated to other grades. This implies that the same dimensions are the focus of the teacher's efforts in each grade, as well. This is usually a problem if the goal is to scale across more than two adjacent grades. Even with two adjacent grades, it is not usually clear that the same dimensions are being assessed. If you are trying to scale two or more tests and the tests are really not assessing the same content, you are actually predicting one from the other, rather than equating the two.

The equating of two tests in the horizontal scaling context is fairly easy using an item response theory (IRT) approach (e.g., Stocking and Lord, 1983). If one believes that the content dimensionality assumption in vertical equating is met, then a variety of approaches can be adopted to accomplish the task of equating across several grades. The paper by Tomkowicz and Schaeffer (2002) about implementing a strategy in Mississippi provides one example. Vertical equating also has been carried out on a trial basis for the South Carolina PACT assessments in reading and mathematics. Generally, the procedures focus on adjacent grades since these are usually the most instructionally similar and more likely to be content similar, as well. The successful equating across grades also involves careful design of each grade's test so that overlap across grades will be more systematically achieved. For example, to vertically equate grades 3 through 8, the design for each test will involve carefully crafted subtests (one for grades 3 and 8 and two for the other grades). This will provide enough overlap in difficulty level to allow scaling adjacent grades.

Major Problems with Vertical Equating

Vertical equating is useful mainly in reading and mathematics, the two subjects that are taught and learned continuously through the schooling process. A vertically equated scale cannot be reasonably constructed for subjects like science (e.g., trying to equate physics and geology) or social studies, and issues arise even in scaling mathematics or reading/language arts. A vertical scale captures the common dimension(s) across the grades; it does not capture grade-specific dimensions that may be of considerable importance. The instructional expectations for teaching and learning reading/language arts and mathematics may not really be summarized by one (or even a few) common dimensions across grades.

The assumption of equal-interval measurements within a grade is not easily met either, and across grades it is very hard to justify, so the comparison of growth at different times in a student's life or comparisons of different groups of students at different grades cannot be satisfactory made. Since the typical motivation for vertically equated scales revolves around capturing the developmental process, this difficulty is a serious issue for schools wishing to implement vertical equating.

Going to a single dimension to capture a very rich assessment environment encourages simplifications that lose the very insights that the assessments were done to illuminate. As Haertel (1991) noted with regard to the decision to abandon across-grade scaling for grades 4, 8, and 12 for the National Assessment of Educational Progress, "In fact, it is very difficult to say anything useful about the fact that eighth graders outperform fourth graders by more points than twelfth graders outperform eighth graders" (p. 13).

Since the nature of the items and the assessment process often changes over grades, vertical equating mixes or confounds content changes with method changes. This makes interpretation of results difficult and violates the assumption of comparable assessment across grades. Further, capturing the span of test difficulty within a single scale is very difficult.

Creating the vertical scale is also a technically difficult task, even with (perhaps because of) the use of IRT models. Artificial adjustments must be made to smooth out the results. As Camilli (1999) indicates, "Dimensionality remains a concern, though investigation of its interaction with equating is significantly complicated by indeterminacy of the latent scale" (p. 77). In simplest terms, performance and learning are essentially multidimensional activities.

AN ALTERNATIVE APPROACH: VERTICALLY MODERATED STANDARDS

After examining the problems related to vertical scaling, it is reasonable to conclude that the construction of a vertical scale to equate state assessments is difficult to accomplish, difficult to justify, and difficult to utilize productively. Even if a satisfactory vertical scale could be constructed, it makes little sense to report reading and mathematics on one vertical scale and science and social studies on a different scale. Within-grade scales could be useful in themselves; however, there is another approach that could be even more beneficial to help teachers and principals use state assessment results as they try to comply with the NCLB legislation.

We recommend vertically moderated standards—a primary focus upon the categories of performance that a given state department of education has determined (e.g., advanced, proficient, basic, below basic) and efforts to relate these explicitly to adequate yearly progress through a carefully crafted judgment process. In other words, we recommend defining AYP in terms of adequate end-of-year performance that enables a student to successfully meet the challenges in the next grade. Vertically moderated standards call for state departments of education to implement a judgmental process and a statistical process that, when coupled, will enable each school to project these categories of student performance forward to predict whether each student is likely to attain the minimum, or proficient, standard for graduation, consistent with NCLB requirements.

With the focus of assessment necessarily upon classroom instruction and teachers' adaptation to student needs, changes in the specific scale scores should not be the focus. Rather, the focus should be upon each student meeting the achievement categories at a level that predicts adequate (i.e., successful) achievement in the next grade. Particularly in a large state assessment, it is important to use a common reporting system for all students and this approach will accomplish that.

General Considerations for Vertically Moderated Standards

Mislevy (1992) and Linn and Baker (1993) defined four types of linking: equating, calibration, projection, and moderation. These are listed in decreasing order in terms of the assumptions required, with equating requiring the strongest assumptions and moderation the weakest. The ordering of the four types is also in decreasing order in terms of the strength of the link produced.

Under the best conditions, vertical scaling would fall under the category of "calibration." Given misgivings about the feasibility and usefulness of vertical scaling for state assessments, we believe that a procedure that combined the major features of "projection" and "moderation" should be considered and a reporting system that emphasizes achievement levels (e.g., similar to the NAEP categories of advanced, proficient, basic, below basic) would provide information that is easier to understand. This may necessitate that states undertake a new round of standard setting for their assessments. We recommend that cut scores for each test be set for all grades such that (a) each achievement level has the same (generic) meaning across all grades, and (b) the proportion of students in each achievement level follow a growth curve trend across these grades.

The first criterion may be referred to as "policy equating" in the sense that a common meaning is attached to each achievement category. Thus, in some sense, the term "equating" is used in the context of a qualitative (i.e., having to do with quality) interpretation of test score. The second criterion is similar to the "linear statistical adjustment" (Mislevy, 1992) that imposes some level of consistency in the normative data of all grades under consideration. This type of consistency is based on the belief that current instructional efforts and expectations are approximately equivalent in all grade levels, so there should not be wild and unpredictable variations in student performance across grades for an entire state.

An Example of Vertically Moderated Standards

The 1999 standard setting for the South Carolina 1999 PACT assessments (Huynh, Meyer, & Barton, 2000) produced standards that may be described as "vertically moderated." The South Carolina process followed three basic steps:

- A common set of policy definitions for the achievement levels was agreed upon for all grades in each area.
- Cut scores were initially set for grades 3 and 8 only.
- Once the final cut scores for these grades were adopted by the state based upon a technical advisory committee's recommendation, cut scores for grades 4 through 7 were interpolated from those of grades 3 and 8. A simple growth curve trend line was used in the interpolation.

Procedures for Developing Vertically Moderated Standards

Setting vertically moderated standards for several grades requires adopting a forward-looking orientation regarding proficiency, examining curriculum across grades, considering smoothing procedures for the statistical process, and paying special attention to issues related to at-risk students. States should also conduct annual validation studies to guide their assessment programs. These procedures are introduced below.

Forward-looking Definition of Proficient. The complexity of this judgment process would indicate that most states would requires two groups—one for mathematics and one for reading/language arts--to provide advice that determines the cut-points that would be used to define levels of achievement on a test. The definition of proficient should be forward-looking; that is, students who achieve that category should be understood to be proficient on the material from the grade covered by that year's end-of-grade testing and also judged to have made adequate yearly progress at a level that will enable them to likely be successful in the context of the next school grade. In other words, students who score at the proficient level on an assessment should have the educational background from that grade to succeed in the next.

Use of Content Scatter Plots. The judgment process for determining cut-points for the categories of performance on a state's end-of-year exam will involve examining certain relevant data, in addition to the test items from the end-of-year test. In the new process, the judges will need to see the test that will be used in the next grade's end-of-year exam, as well as a description of the relevant curriculum for both years.

The new process will also require that the judges become familiar with a grade-to-grade scatter plot of the two test blueprints, a so-called assessment scatter plot. The scatter plot presentation will be a comparison of the assessment design from the current grade to the coming grade. This curriculum/test assessment blueprint scatter plot will provide an indication of the topic areas that are found on both exams (for example, mathematics at grade 7 and mathematics at grade 8) as well as the topic areas that are unique to each exam (i.e., that material which has no overlap across grades). Taking into consideration the content scatter plot will help maintain the common qualitative interpretation of the achievement levels across grades.

Use of Smoothing Procedures for Interpolation and/or Extrapolation. To set vertically moderated standards for several grades (say 3 through 8), there may be no need to conduct the standard setting for all grades. This may be done for two grades at a minimum, but perhaps three grades will be necessary. Interpolation and/or extrapolation would then be used to compute the cut scores for the other grades, with an eye on the proportion of students who are judged to be proficient at each grade. We recommend that cut scores be smoothed out so that the proportion of student in each achievement level is reasonably consistent from one grade to the next. A smoothing procedure may prove satisfactory for the statistical process, which would then supplement the professional judgment involved.

Use of Margin of Error with Focus on At-Risk Students. For many large-scale assessment programs (such as NAEP and the South Carolina state assessment), deliberations regarding the final set of cut scores often take into account the margin of error inherent in any standard-setting process. Judges vary in backgrounds and their individual, recommended, cut scores often vary as well. Therefore it is safe to presume that, over a large pool of judges, the (true) recommended cut score would fall within a reasonably small band, centered at the recommended cut score.

For an assessment program with heavy focus on instructional improvement, some attention may need to be paid to the students who are at risk of being in a false positive category. These are students deemed marginally proficient in the current year, but who may not have acquired the necessary skills needed for learning the material that will be presented next year. They may be at risk of not reaching the proficient level, as required by AYP, at the end of the following year. Supplemental data, such as grades, attendance, special education or limited-English-proficient status, and teacher documentation will aid in the identification, and subsequent remediation, of at-risk students.

Annual Validation Study. Each year, state department of education should also do a validation study in order to identify any problems with the implementation or operationalization of the system into school practice and to see if changes in the level of proficiency are warranted to lead to the overall success of the schools at the end of the 2013-2014 year, as mandated by the NCLB. State department of education will also need to identify as early as possible those schools that do not seem to be on track to meet the federal guidelines for success (100% of the students achieving proficiency within 10 years). Appropriate assistance, sanctions, and rewards can then be offered.

Vertically moderated standards show great promise for state departments of education attempting to track student performance and academic growth on state assessments in a way that is responsive to the NCLB requirements and also yields genuinely useful instructional information. The combination of judgment and statistical analysis described in this article should result in the creation of cut scores that describe proficiency both in terms of a student's mastery of grade-level material and the likelihood that s/he will be ready for the academic challenges of the next grade. Where students score below proficient, appropriate remediation can be offered early so that schools meet annual yearly progress goals and, more important, children are not left behind.

Note:

This paper is based on a report originally prepared for the Technical Advisory Committee of the Arkansas Department of Education.

References

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement,* 36, 73-78.

CTB/McGraw-Hill, (1997, 2001) TerraNova. Monterey, CA: Authors.

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.

Huynh, H., Meyer, P., & Barton, K. (2000). Technical Documentation for the South Carolina 1999 Palmetto Achievement Challenge Tests of English Language Arts and Mathematics, Grades Three Through Eight . Columbia, SC: South Carolina Department of Education.

Linn, R. L., & Baker, E. L. (1993; Winter). Comparing results from disparate assessments. *The CRESS Line*, pp 1-2. Los Angeles: National Center for Research on Evaluation, Standards, & Student Testing.

Marion, S, et. al. (2002). Making valid and reliable decisions in determining adequate yearly progress. Washington, D.C.: Council of Chief State School Officers.

Mislevy. R. J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological*

Measurement, 7, 201-210.

Harcourt Educational Measurement (1996). Stanford Achievement Test Test Series, Ninth Edition. San Antonio, TX: Authors

Tomkowicz, J., & Schaeffer, G. (2002, April). Vertical scaling for custom criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Descriptors: Accountability; Elementary Secondary Education; Federal Government; Federal Legislation; AYP; Methods; Equating

Citation: Lissitz, Robert W. & Huynh Huynh (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10). Available online: http://PAREonline.net/getvn.asp?v=8&n=10.