ED 474 270 TM 034 805

AUTHOR	Lissitz, Robert W.; Huynh, Huynh
TITLE	Vertical Equating for the Arkansas ACTAAP Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability. A Report Submitted to the Arkansas Department of Education.
PUB DATE	2003-01-21
NOTE	25p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Accountability; *Equated Scores; *Scaling; Standards; State Programs; *Testing Programs
IDENTIFIERS	*Arkansas; *Vertical Equating

ABSTRACT

This paper addresses issues of vertical equating for the Arkansas Comprehensive Testing, Assessment and Accountability Program (ACTAAP) assessments as they relate to school accountability and determination of Adequate Yearly Progress (AYP) as required by the recent federal legislation, the No Child Left Behind Act. The paper first provides a brief statement of the problem, followed by a review of the testing in Arkansas, and related policies. It also examines some of the issues raised by this testing for AYP determination. The paper explores one commonly suggested solution to such a problem, scaling to arrive at vertically equated tests, and explains why this approach is not recommended in Arkansas for AYP determination. The paper suggests a solution, which is referred to as Vertically Moderated Standards. It is suggested that AYP be defined in terms of adequate end-of-year performance that enables a student to meet the challenges of the next grade successfully. (Contains 10 references.) (Author/SLD)

VERTICAL EQUATING FOR THE ARKANSAS ACTAAP ASSESSMENTS:
Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability

A Report Submitted to
The Arkansas Department of Education

By

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

For Deliberations By
The Technical Advisory Committee

William Brown
Gregory Cizek
Huynh Huynh
Robert Lissitz
William Mehrens
Roger Trent

This paper addresses issues of vertical equating for the Arkansas *ACTAAP* assessments as they relate to school accountability and determination of Adequate Yearly Progress (AYP) as required by the recent federal legislation (2001 ESEA reauthorization act) known as No Child Left Behind (NCLB). The paper first provides a very brief statement of the problem, followed by a review of the testing in Arkansas, and related policies. We also provide an examination of some of the issues raised by this testing for AYP determination. We then look at one commonly suggested solution to such a problem (scaling to arrive at vertically equated tests) and why we do NOT recommend its use by Arkansas for AYP determination. Finally, we present our solution, which we will refer to as Vertically Moderated Standards..

Note: The authors made equal contributions and their names are listed in random order.

January 21, 2003

PART I

THE PROBLEM OF ADEQUATE YEARLY PROGRESS

The paper by the Council of Chief State School Officers (CCSSO), titled Making Valid and Reliable Decisions in Determining Adequate Yearly Progress (Marion at al., 2002) provides considerable guidance regarding the problem of determining AYP, including the following selected quotes:

1. "While expectations for the proficient level will vary by state, AYP is based on the percent of students meeting proficient and the expected percentage increases over time (p.1)."

2. "This submittal must, "provide the State's definition of adequate yearly progress. The definition must include: i. For the percentage of students meeting or exceeding the State's proficient level for reading/language arts and for mathematics—the starting point percentage; The intermediate goals; The timeline; and Annual objectives. ii. The definition of graduation rate.... iii. One academic indicator applicable to elementary schools. iv. Any other (optional) indicators... (and) identify the minimum number of students that the State has determined, based on sound methodology, to be sufficient to yield statistically reliable information for each purpose for which disaggregated data are used and justify this determination" (U.S. Department of Education, 2002, pp. 12-13) (p.3)."

3. "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 states performance target in another academic indicator as prescribed by the law (p.5)."

4. "States may select any academic indicator for the elementary school level...(p. 8)."

5. "Graduation rate is required for the secondary school level indicator....(p. 8)."

6. "....additional academic indicators .... Cannot be used to "compensate" for other missed performance targets... (p. 9)."

7. "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)."

8. ".... Assessments must be administered at least once annually in each grade, 3 through 8 by 2005-06 (and once within grades 10-12) and with science administered at least once in each of the three grade spans by 2007-08 (p. 5)."

To briefly summarize the problem, Arkansas must develop a system that

- tracks students' (by defined subgroups) success in
- reading/language arts and mathematics (with science coming on board soon)
- as the student progresses 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.

The next section of our report seems to indicate that the State has initiated the necessary testing program to meet the federal requirements, assuming that a measure of science achievement is established for each of the grade levels (elementary, middle, and secondary).

The CCSSO document (2002) was written to help state accountability programs. Unfortunately, it is not prescriptive, but instead it has attempted to clarify the requirements and the goals of the NCLB act. Therefore, Arkansas is left to its own devices to determine how it will define the AYP process and how it will implement this process. The purpose of our document is to provide that guidance to the Arkansas Department of Education (ADE).

PART II

A SUMMARY OF THE ACTAAP STATE ASSESSMENTS

Purposes and Goals

The *Arkansas Comprehensive Testing, Assessment & Accountability Program* (*ACTAAP*) has developed in response to Arkansas Legislative Act 1172, which requires the State Board of Education to develop a comprehensive testing program that includes performance assessment of the core concepts, abilities, and thinking and problem-solving skills defined by the *Arkansas Curriculum Frameworks*. The original legislative intent of these examinations was to promote the development and local implementation of the *Arkansas Curriculum Frameworks*, and the development and use of assessment in accordance with defined statewide goals for instructional change.

ACTAAP is a comprehensive system that focuses on high academic standards, professional development, student assessment, and accountability for schools. More specifically, the goals for the *ACTAAP* are to:

- improve classroom instruction and learning;
- support public accountability by exemplifying expected achievement levels and reporting on student and school performance;
- provide program evaluation data; and

- assist policymakers in decision making.

General Descriptions on Types of Tests

The *ACTAAP* encompasses the state's Smart Start Initiative, which focuses on Grades K-4; the state Smart Step Initiative, which focuses on Grades 5-8; and education for Grades 9-12. *ACTAAP* represents the culmination of extensive planning and discussion by Arkansas educators, policymakers, and school patrons.

The authority to implement *ACTAAP* is firmly established in legislation by Act 999 of 1999. Current law and State Board of Education regulations require the administration of criterion-referenced tests (CRTs), and a norm-referenced test (NRT). CRTs are administered at Grades 4, 6 and 8 (Benchmark Exams), End-of- Course Exams in Algebra I and Geometry, and a Literacy Exam at Grade 11. The CRTs are aligned to the Frameworks and were developed by Arkansas teachers and the ADE (Arkansas Department of Education). The state's norm-referenced test (NRT), presently the Stanford Achievement Test, Ninth Edition (SAT-9), is administered at Grades 5,7, and 10.

Criterion-Referenced Tests

While the *ACTAAP* has evolved with changes in curricular and program focus, the primary purpose of *ACTAAP*'s criterion-referenced instruments has remained the same. As Arkansas moves toward the implementation of a statewide accountability system, the criterion-referenced testing component of *ACTAAP* continues to be both a basis for evaluating student performance relative to state curriculum-based goals for student success and an impetus for instructional change.

The criterion-referenced tests developed for and administered in Arkansas are specifically designed to align with the *Arkansas Curriculum Frameworks* to assess student performance relative to the mathematics, reading, and writing curricula prescribed by the Arkansas State Board of Education. In addition, student performance results are related to specific standards of performance set by Arkansas educators. Unlike norm-referenced tests, the criterion-referenced testing portion of the *ACTAAP* is specifically designed to measure and evaluate student performance relative to what students are to be taught in the state of Arkansas.

The *ACTAAP* criterion-referenced part is comprised of the following assessments:

- *Benchmark Examinations* for grades four, six, and eight, (mathematics and literacy),
- *End of Course Examinations* for *Algebra I*, *Geometry*, and *Literacy (Grade 11)*,
- *Alternate Portfolio Assessments for Students With Disabilities* for grades four, six, and eight, (mathematics and literacy) and for *End of Course Literacy (Grade 11)*, and
- *Alternate Portfolio Assessments for Students With Limited English Proficiency* for grades four, six, eight (mathematics and literacy), and grade 11 (literacy).

The *End of Course Algebra I, Geometry,* and *Literacy (Grade 11) Examinations* were all field-tested in the spring 2000 administration and pilot-tested in the Midyear 2001 and/or spring 2001 administrations. They became operational in the Midyear 2002 and spring 2002 administrations. These examinations have been designed to assess student performance in algebra or geometry (mathematics) and literacy (reading and writing) to determine the level of proficiency exhibited by students in these content areas.

Achievement Level Reporting

Keeping with the NAEP tradition, four levels of student achievement are used for the Benchmark examinations. Following are their descriptions.

Advanced: Advanced students demonstrate superior performance well beyond proficient grade level performance. They can apply Arkansas established reading, writing, and mathematics skills to solve complex problems and complete demanding tasks on their own. They can make insightful connections between abstract and concrete ideas as well as provide well-supported explanations and arguments.

Proficient: Proficient students demonstrate solid academic performance for the grade tested and are well prepared for the next level of schooling. They can use
Arkansas established reading, writing, and mathematics skills and knowledge to solve problems and complete tasks on their own. Students can tie ideas together and explain the ways their ideas are connected.

Basic: Basic students show substantial skills in reading, writing, and mathematics; however, they only partially demonstrate the abilities to apply these skills. They demonstrate a need for some additional assistance, commitment or study to reach the proficient level.

Below Basic: Below basic students fail to show sufficient mastery of skills in reading, writing, and mathematics to attain the basic level.

Major Use of *ACTAAP* Data

ACTAAP data are reported back to parents, teachers, and appropriate local school authorities for various instructional planning purposes. In addition, the data are used for school improvement planning, determining rewards and sanctions, and creating accountability indicators and the goals associated with them. The following is a brief summary of these, but the reader should note that they are subject to change. The ADE is currently engaged in a process to consolidate its assessment and accountability plans, while insuring that *ACTAAP* meets all requirements of NCLB.

School Improvement Planning

Each school must have a school improvement plan that specifies priorities and the relevant indicators from the student assessment and other data. The emphasis is consistent with and even exceeds the requirements of AYP, in that it is designed to insure that all students demonstrate proficiency on all portions of the statewide CRT examinations. The school improvement plan also calls for disaggregated data analysis, again consistent with Federal guidelines. In addition, the schools must identify performance based benchmarks and intervention and remediation strategies. The strategy requires intervention and remediation that is based on scientific study of proven success.

Rewards

Rewards are based on a system that recognizes schools that demonstrate and maintain high performance over time on state and school-selected indicators. Rewards will be recognitions of 1) absolute levels of student achievement and other indicators, and 2) recognized improvement in student achievement and other indicators, over time. Funds will be used to improve the capacity of the schools to better serve their students.

Sanctions

Sanctions are for the purpose of improving teaching and learning. Based upon a lack of success on the indicators selected, the ADE may designate a school as in 1) high priority status in year one, 2) alert status in year two, 3) low performing status in year three, and 4) academic distress in year four.

Continued lack of success may cause a school to be considered in Academic Distress Phase I, II, or III.

Accountability Indicators

Attached is a table titled Tier I indicators and verbiage associated with that table which define the indicator, the goal and the grades for which they apply. The table titled Tier II relate trend and improvement goals and generally add to the goals identified in Tier I. Here the ADE is clear about the need for disaggregation (see AYP discussion above) of student groups.

Summary

In summary, the Arkansas Department of Education has already established a statewide testing program. Various elements of the assessment program can be used in a manner that is consistent with the NCLB legislation and the requirements for determining AYP. The next step is to clarify this process so that it can be implemented to the state and federal government's satisfaction.

PART III

PSYCHOMETRIC ISSUES RELATED TO AYP DETERMINATION

This section of the paper starts with a brief discussion of the purpose of scaling and what some psychometricians hope it will permit schools to do, and then proceed to the notion of a vertically equated scale and the problems associated with that approach to the AYP determination.

General Description of Scaling

A scaling process, in general terms, is one in which raw scores 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 topic of equating and scaling can get very complex very quickly and this report will avoid doing that. The particular approach used depends upon the purpose of the scaling and the properties that we want in the resulting scale.

One common purpose of scaling, as indicated above, is to transform raw scores from a test to a new set of scores with special properties. Another, and 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 in that case, 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 (usually represented by the sum of the number of items found to be correctly answered) has several advantages, including the following:

- 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 state 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.

- The *ACTAAP* scale for each test is defined such that 200 is considered to be "Proficient" and 250 is considered to be "Advanced". This is another good example of the use of scales to maintain some uniformity of communication.

Within-Grade (Horizontal) Scaling

There are two primary situations for scaling multiple sets of tests (also referred to as the equating of tests) to a common scale. These are horizontal (within grade, multiple forms testing of the same general content) and vertical (across grade testing of the same general content) equating. In other words, the one approach is designed to test different groups of students that are assumed to be at approximately the same educational level. The other approach is for the testing of students who are at different levels of education. 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 insuring a constancy of the standard for passing from administration to administration. The table of specifications (i.e., the test blue-print) for each test is also the same, thus insuring comparability of content 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 only speak 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 they claim allow for the meaningful, continuous, tracking of students across grades (vertical equating).

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 are of differing content, but still focus on the same general concept of mathematics fluency, say. 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 two tests (7th, and 8th grade) should be linked for each grade 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 communality 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. In other words, this assumption of common content is really critical for the success of equating two or more tests.

Implementation through IRT

The equating of two tests in the horizontal scaling context is fairly easy using an 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) implementing a strategy in Mississippi, is one example. Vertical equating also has been carried out (on a trial basis) for the South Carolina PACT assessments in reading and mathematics. As we indicated above, several national testing organizations have done this as well. 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, although 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.

NAGB Positions on Across-Grade Scaling

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.

NAEP has abandoned across-grade scaling for grades 4, 8 and 12 for a variety of reasons.
The reasons for the National Assessment Governing Board decision, summarized from Haertel (1991), and some of our own comments directly relate to the current Arkansas and AYP issues, and are presented next:

- … "scales defined separately for the three grade levels, would be more valid, more accurate, or less likely to be misinterpreted or misused (p1)."

Haertel is summarizing a position held by many psychometricians and is important for Arkansas because the state faces many of the same issues that the federal government does. For example, both face the difficulty of explaining to the teachers and principals and the press what a vertical scale means. 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.

- Since the nature of the items and the assessment process often changes over grades, vertical equating mixes (i.e., confounds) content changes with method changes. This makes interpretation of results difficult and violates the assumption of comparable assessment across grades.

- NAEP is moving away from across grade scales and comparisons, thus lending credence and legitimacy to Arkansas doing the same thing.

- Capturing the span of test difficulty within a single scale is very difficult. As noted earlier, assessments at different grades can differ considerably, thus violating assumptions of vertical scaling.

- Haertel presents the difficulty of explaining the results, as follows: "In fact, it is very difficulty to say anything useful about the fact that eighth graders outperform fourth graders by more points than twelfth graders outperform eighth graders (p. 13)."

- Haertel also notes that " This (note: Haertel is referring to interpretations such as those using "grade equivalents") along with several others, depends critically on the linearity of the cross-age scale (p. 12)." Again, we note that linearity would be very difficult to guarantee and being wrong leads to misinterpretations.

- Haertel also makes clear that using a vertical scale to make comparisons of subgroup performance differences at one grade level to those at another grade level, is not a good idea. The linearity assumption is critical and not justified here, either.

- Haertel, near the end of his paper, makes a particularly important point: "A final objection to this form of interpretation is that it is useless for informing curriculum, instruction, or educational policy. Reliance on within-grade scales would direct attention to the more productive and meaningful comparisons among groups of children of the same age (p.14)." We think this is especially important for ADE to consider. In other words, it is not the job of the federal government to tell Arkansas what to teach or how to teach it, but that is the job of Arkansas and assessments that give such insights to teachers are critical. If the NAEP testing does not do that when it uses vertical equating, it is unlikely that Arkansas will get such insights if it were to use vertically equated scales.

Technical Difficulty

We would also like to note that creating the vertical scale is 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 indicated by Camilli (1999). As he says on page 77, "Dimensionality remains a concern, though investigation of its interaction with equating is significantly complicated by indeterminacy of the latent scale." This observation that performance and learning are essentially multidimensional activities, whose dimensions change across time, is very much related to a number of the points raised by Haertel, and ourselves and others, and summarized above. Tomkowitz and Schaeffer (2002) also noted their own efforts in regard to developing a well-behaved scale.

Needs for a Common Reporting System for All Tests

Although NCLB regulations currently require testing only in reading and mathematics, many other states assess students on other areas such as science and social studies. Efforts to build vertical scales on these two areas are known to run into considerable difficulty or are not successful. Reporting test data for reading and mathematics on a vertical scale and others on a horizontal scale would certainly create unneeded confusion.
(To the knowledge of the second author, North Carolina has abandoned the vertical scale for its science tests and South Carolina is still debating whether to report reading and mathematics scores on a vertical scale.)

Summary

We should note, in summary, that NCLB regulations put an emphasis on examining success (or lack thereof) in grades 3-8, but even in this limited situation, we recommend against adopting a vertical scaling model. And, NCLB does demand that all states include at least one grade in the upper high school years (10 to 12). The result of examining these many issues, is that the construction of a vertical scale is difficult to accomplish, difficult to justify, and difficult to utilize productively. Also, note that any change in instruction or assessment requires that the whole scale be revised which would add considerably to the workload in Arkansas. Finally, it must be remembered that NCLB does NOT require that vertical scaling be utilized in Arkansas or any other state school system.

PART IV

OUR RECOMMENDED APPROACH:

VERTICALLY MODERATED STANDARDS

General Observations

The above concerns lead us to recommend a very different approach. We want Arkansas teachers and principals to be able to use the assessment results from *ACTAAP* as they try to comply with the NCLB legislation. The within-grade scales will be useful by themselves, but we suggest that the primary focus should be upon the categories of performance that the ADE has determined (below basic, basic, proficient, and advanced) and relating these explicitly to AYP through a carefully crafted judgment process. In other words, we recommend defining adequate yearly progress in terms of adequate end of year performance that enables a student to successfully meet the challenges in the next grade. We believe that ADE can implement a judgmental process and a statistical process that, coupled together, will enable each school to project these categories of student performance forward to predict whether each student is likely to attain the minimum standard (proficient) for graduation, consistent with NCLB demands. While we are using the term "end of year" performance, we note that because of various constraints, the assessment cannot be administered at the very end of the year.

What is Important to the Teacher?

With the focus of assessment necessarily upon classroom instruction and teachers' adaptation to student needs, we argue that changes in the specific scale scores should not be the focus. Rather, we argue that 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. As we indicated above, it does not seem to make sense to report reading and mathematics on one vertical scale and science and social studies on a different scale, even if we were able to construct a satisfactory vertical scale.

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 our misgivings about the feasibility and usefulness of vertical scaling for the *ACTAAP*, we believe that a procedure that combined the major features of "projection" and "moderation" would be in the best interest for Arkansas. We also believe that reporting achievement levels (in the four NAEP categories, say) would provide information that is easier to understand.

Currently *ACTAAP* data are reported using a within-grade scale and the four achievement levels Below Basic, Basic, Proficient, and Advanced. The Technical Advisory Committee (TAC) understands that this recommendation will necessitate a new round of standard setting for the *ACTAAP* assessments. We recommend that new 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 the entire state.

An Example of Vertically Moderated Standards

The 1999 standard setting for the South Carolina 1999 PACT assessments (Huynh, Barton, & Meyer, 2000) produced standards that may be described as "vertically moderated." We are not suggesting that the South Carolina process can be immediately applied without modification to Arkansas, but that it is an interesting example of what can be done. 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 three and eight, only.
- Once the final cut scores for these grades were adopted by the State, upon the TAC's recommendation, cut scores for grades four through seven were interpolated from those of grades three and eight. A simple growth curve trend line was used in the interpolation.

PART V

SPECIFIC RECOMMENDATIONS FOR SETTING VERTICALLY MODERATED STANDARDS

The following outlines the procedure that we are recommending. Many additional, specific details will need to be supplied as further recommendations are elicited from the Technical Advisory Committee. Regardless of that need, the following outline summarizes the approach we are recommending.

Broad Representation of Judges

Judges will need to be selected broadly from the various constituents of the State Education System. We assume this selection process will be similar to that used for Bookmark sessions in which the state has already been engaged. We see no need for change in the selection process, as a function of changes outlined below. The complexity of this judgment process does require that two groups make their judgments, one group involving mathematics and one group involving reading/language arts

Forward-Looking Policy Definitions

Judges will ultimately be asked to advise the state on a somewhat different question than they did before. When they provide advice that determines the cut-points that will be used to define below basic, basic, proficient, and advanced, they will be asked to focus on a somewhat different definition of proficient. The new approach to defining proficient gives it a forward-looking orientation. In other words, proficient is now taken to mean that a student who achieves that category is not only proficient on the material that they were learning from the grade covered by that year's end-of-grade testing, but that the student is judged to have made adequate yearly progress at a level to be likely to be successful in the context of the next school grade, as well. In other words, the student has the educational background from that grade to succeed in the next. The judgment could be made using a Bookmark method or a Modified Angoff method and the Technical Advisory Committee (TAC) advice should be sought on the specific method. To capture the fact that this judgment is longitudinal, we suggest a little different terminology, perhaps something like "Stochastic Bookmark" or "Stochastic Modified Angoff" approach.

Use of Content Scatter Plots

The judgment process for determining cut-points for the categories of performance on that 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 (which we are calling an 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 would 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 (VMS) 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 SC PACT assessments), 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 such as the *ACTAAP*, perhaps some attention needs to be put on the students who are at risk of being in a false positive category. These are students deemed marginally proficient in the current year, but 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.

PART VI

ON-GOING DATA COLLECTION AND RESEARCH ACTIVITIES FOR AT-RISK STUDENTS

Data on At-Risk Students

There will also be a collection of supplemental performance data required by NCLB and useful to Arkansas schools to gauge the At-Risk nature of specific students and the student body (i.e., the school). These student data include:

- Grades relevant to mathematics and reading/language arts
- Attendance record
- Whether student was a recent transfer or not
- Any recorded behavior problems (disciplinary actions)
- Special education status, if any, and I.E.P. status
- NAEP scores or any other relevant test information
- Teacher's warnings of potential problems
- Graduation status and relation to earlier assessment results for seniors
- ESL status

These data should be helpful to the state in formulating policy and procedure to deal with At-Risk students in order to improve their chances of meeting the AYP classification for the following year

Annual Validation Study

Every year, the ADE will do a validation study. The purpose of this validation study will be two-fold. In the first year following the implementation of this system and any changes in the system, the ADE will need to identify any problems with the implementation or with the operationalization of the system into school practice. Second and every year, the ADE needs to examine the data that result from the operationalization 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. The ADE will also need to identify schools, as early as possible, that do not seem to be on track to meet the federal guidelines for success (100% of the students achieving proficiency). Arkansas already has a well-articulated system of sanctions and rewards and these will be implemented as a by-product of the analysis.

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 404367): Washington, D.C.: 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.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034805

I. DOCUMENT IDENTIFICATION:

Title: Vertical Equating for the Arkansas ACTAAP Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability	
Author(s): Robert W. Lissitz, Huynh Huynh	
Corporate Source:	Publication Date: January 21, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ XX	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discreta inquiries.

Sign here, → please

Signature: Robert W. Lissitz	Printed Name/Position/Title: Robert W. Lissitz/ Department Chair	
Organization/Address: University of Maryland/ EDMS Department 1230 Benjamin Building College Park, MD 20742-1115	Telephone: 301-405-3624	FAX: 301-314-9245
	E-Mail Address: rl@umail.umd.edu	Date: 3-19-03

ERIC
Full Text Provided by ERIC

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org