Measuring Growth with the MCAS Tests:
A consideration of vertical scales and standards[1]

Charles A. DePascale
National Center for the Improvement of Educational Assessment
October 2006

A consequence of the No Child Left Behind (NCLB) requirement for annual statewide testing in reading and mathematics at grades 3 through 8 is an increased focus on student growth from one grade to the next. When students are tested each year, there is a natural inclination among educators, parents, and students to ask questions such as:

- *Does Johnny read better this year than last year?*
- *Does Jane know more math than she did last year?*

These rather general questions may be followed by more specific and evaluative questions such as:

- *How much better did Jane and Johnny do this year?*
- *How much more do they know?*
- *Did they improve as much as expected?*

Finally, information about current performance may lead to questions about future performance:

- *Is Jane on track to be successful next year? The following year?*
- *How much should we expect Johnny to improve next year?*

These are not unreasonable questions to ask. At issue, however, is how best to answer such questions based on results from a large-scale assessment system designed specifically to measure performance against an established standard within a given year.

As interest in measuring student growth over time has grown, two terms and concepts have received considerable use in both the professional literature and among policymakers: vertical scales and vertically-moderated standards. The discussion of scales based on vertically-moderated standards as an alternative to vertical scales (e.g., Lissitz and Huynh, 2003; Schafer, 2006) has led some to the erroneous conclusions that

a) vertical scales and vertically-moderated standards are two approaches designed to answer the same set of questions;
b) vertical scales and vertically-moderated standards are mutually exclusive; and
c) vertical scales or vertically-moderated standards are necessary to measure student growth.

In reality, vertical scales and vertically-moderated standards are intended to answer very different questions about student growth and may both be used within a single assessment system. Additionally, growth models exist that rely on neither vertical scales nor

[1] This paper was prepared under a contract with the Massachusetts Department of Education. The views expressed are solely those of the author and do not necessarily reflect the views, policies, or positions of the Department of Education or the National Center for the Improvement of Educational Assessment.

vertically-moderated standards to represent and make predictions about student growth over time.[2]

In this paper, the concepts of vertical scales and vertically-moderated standards as well as student growth, in general, are considered in the context of the grade 3 through 8 MCAS tests in English Language Arts and Mathematics. Part I of the paper provides a description of relevant characteristics and design features of the MCAS tests. Part II contains a discussion of student growth distinguishing among three conceptions of growth: growth relative to self (i.e., growth on the construct), growth relative to others (e.g., the student's cohort), growth relative to an established standard (e.g., the grade level achievement standard for proficiency). Following the discussion of student growth, the concepts of vertical scales and vertically-moderated standards are defined and described in relation to student growth, each other, and the MCAS tests in Part III of the paper. Part IV summarizes the discussion on the usefulness and practicality of vertical scales and vertically-moderated standards for measuring growth with the MCAS tests.

Part I: The MCAS Tests

Since the initial MCAS administration in 1998, the Massachusetts Department of Education has consistently described the purposes of MCAS and the MCAS tests in terms of improvement and accountability:

> MCAS has three primary purposes: (1) to inform and improve curriculum and instruction; (2) to evaluate student, school, and district performance according to *Curriculum Framework* content standards and MCAS performance standards; and (3) to determine eligibility for the high school Competency Determination requirement. (MA DOE, 2006, p.9)

This description also illustrates the dual focus of MCAS on student and school/district performance and accountability. That dual focus characterized the Massachusetts Education Reform Law of 1993 and distinguished MCAS from the previous state assessment program which was designed to provide only school/district level results.

The emphasis on individual student performance is also reflected in the range of student-level reports provided to educators and parents with the results of each MCAS administration. Given the existing level of attention devoted to individual student results, the interest in using the MCAS tests to measure student growth from grade to grade is compatible with the overall goals and purposes of the program.

The MCAS tests and reporting structure, however, were not designed to support the measurement and reporting of growth from grade to grade. From the initial administration of the MCAS tests through the Spring 2000 administration, MCAS tests were administered to students in grades 4, 8, and 10. An individual student enrolled in a Massachusetts public school from grades K-12 would participate in MCAS testing once at the elementary school, middle school, and high school levels. This design, of course, is not conducive to the measurement of individual student growth. Consequently,

[2] This paper does not address the topic of growth models that are being developed for NCLB and other accountability purposes. The focus of this paper is on the reporting of assessment results in a manner that supports interpretations of individual student growth across grade levels.

measuring and communicating information regarding student growth was not a high priority.

Subsequent to the Spring 2000 administration and prior to NCLB, there were changes to the original MCAS test administration schedule. The Spring 2001 administration changes included the introduction of additional tests (i.e., grade 3 Reading, grade 6 Mathematics) and the shifting of the grade 8 English Language Arts test to grade 7. These changes, however, did not impact the fundamental design of the MCAS tests and reports.

Decisions regarding the design of the MCAS reporting scales reflect both the emphasis on achievement standards as well as a focus on performance within a single grade (or grade span) rather than across grades. The MCAS reporting scale was designed to maximize interpretability in relation to the achievement standards established at each grade level tested. Results of each MCAS test are reported on a scale that ranges from 200 to 280 with fixed points of 220, 240, and 260 representing the thresholds dividing the four MCAS achievement levels: *Warning (Failing), Needs Improvement, Proficient, Advanced*. Tradeoffs to the usefulness of the reporting scale for computing mean scaled scores or examining scaled score growth over time were made to enhance the ease with which MCAS results could be interpreted relative to the achievement standards. This emphasis on achievement levels rather than scaled score growth is also reflected in the Composite Performance Index (CPI) computed for the state's accountability system. The design of the existing MCAS reporting scale will have to be considered in any plans to measure student growth across grade levels.

Part II: Student Growth

When considering student growth, it is essential to identify and understand what is meant by the term "growth" and the type(s) of interpretations that will be made. In educational contexts, three general classifications of student growth are common: growth relative to self, growth relative to others, growth relative to a standard. Each type of growth has specific requirements for the measurement of student performance and for the measurement and reporting of a measure of student growth.

Growth Relative to Self

The most fundamental classification of growth can be described as growth relative to self or growth relative to the construct being measured. That is, the question of interest is whether the individual student possesses more of the attribute being measured at Time B than he or she possessed at Time A. In simplest terms, growth is defined as the difference between the measures at two points in time. For example, an individual measured at 63 inches tall in January and 65 inches tall in June has grown 2 inches during the period between January and June.

Paradoxically, in many ways growth relative to the construct is both the simplest and the most complex conception of growth – particularly with regard to complex constructs such as English language arts and mathematics measured across multiple years. As described in the previous example, the basic concept of growth is straightforward. Given two accurate measures or student performance, the growth measure is a simple comparison of the two performance measures.[3] However, among the three conceptions of student growth, measuring growth relative to the construct has the most stringent requirements for the measurement of student performance.

Measurement of growth relative to a construct requires both a well-defined construct and a measurement instrument with enough precision to detect the desired level of growth. As an example, consider again the measurement of a student's height. Units of measurement such as inches or centimeters are well-defined and standardized. Similarly, measuring devices, accurate to the nearest fraction of inches (e.g., 1/8, 1/16) or millimeters are commonplace, easy to use, and interchangeable. Unlike height, constructs such as English language arts and mathematics are more complex and much less well-defined – even within the context of the *Curriculum Frameworks*. Both English language arts and mathematics are a combination of discrete, but related and/or interdependent learning standards. In English language arts, these learning standards are clustered in the broad categories of writing, language, and literature. In mathematics, learning standards are clustered within five content strands: number sense; patterns, relations, and functions; geometry; measurement; data analysis, probability, and statistics. Within each content area, acquisition of knowledge and skills (including the enhancement of current knowledge or skills) within each learning standard is interpreted as an increase in the amount of the overall construct possessed. The combination of knowledge and skills that define a particular level of English language arts achievement may not be unique and may not be consistent from grade to grade.

In addition to being less well-defined than a physical construct such as height, English language arts and mathematics are also more difficult to observe. Levels of achievement in English language arts and mathematics are measured indirectly and inferred through the use of tests such as the MCAS tests. Unlike rulers, tests are more difficult to standardize and are seldom interchangeable.

The lack of a well-defined construct and the use of indirect measurement instruments are two factors that result in a lack of precision in the measurement of achievement in constructs such as English language arts and mathematics. Ultimately, this lack of precision must be considered and accounted for in the measurement of student growth.

Growth Relative to Others

The comparison of the performance of an individual student to the performance of a predefined group is common in the context of education. Student performance may be computed and/or interpreted in relation to a set of classmates (e.g., grading on a curve); a cohort within a school, district, or state (e.g., local user norms); or a previously tested group of students (e.g., a state or national norm group). In such contexts, student growth may be defined as improvement in the student's relative standing within the group from one year to the next. That is, a student scoring at the 50th percentile at Time A and at the

[3] For the purposes of this discussion, issues of measurement error and reliability that will not be addressed. Those factors impact the calculation and interpretation of specific scores, but not the interpretation and use of the general classification of growth scores.

60th percentile at Time B would have demonstrated growth across the time period. Conversely, a student performing at the 50th percentile or lower at both Time A and Time B would be classified as showing no growth across the time period.

This conception of growth makes no assumptions about growth relative to a construct. Growth relative to the construct is treated as irrelevant. In the example above, a student maintaining performance at the 50th percentile in mathematics from grade 5 to grade 6 is likely to “know more math” at the end of grade 6 than she or he did at the end of grade 5. Regardless of the change in the student’s absolute mathematics performance from grade 5 to grade 6, however, the student has shown no growth relative to others. In fact, this conception of student growth imposes few, if any, requirements on the construct being measured at Time A and Time B. Although there are obvious benefits to maintaining some consistency in the construct being measured across years, the degree of consistency needed is more dependent on the particular situation and desired interpretations.

Growth Relative to a Standard

The measurement of growth relative to a standard was not as common as growth relative to others in K-12 large-scale assessment, but has become more commonplace since the advent of standards-based reform, the assessment requirements of the Improving America’s Schools Act of 1994 (IASA), and the accountability requirements of NCLB. The basic growth question addressed is whether a student’s performance is closer to an established standard at Time B than her or his performance was at Time A. Performance closer to the standard demonstrates growth and performance that remains the same or declines relative to the standard demonstrates a lack of growth. As was the case with growth relative to others, stability of the construct from Time A to Time B is desirable but is not a requirement. However, growth relative to the construct may be a more critical factor in evaluating growth relative to a standard than it is in measuring growth relative to others.

The definition and measurement of the concept “closer to the standard” is complex. Growth toward a standard may be defined either in terms of growth relative to the construct or growth relative to others. A common application of the latter involves the use of standard units (e.g., z-scores, standard deviations) to describe closeness to the standard. By definition, such measures are grounded in the relative performance of the group rather than in absolute performance relative to a construct. Less common, are attempts to define closeness to the standard in terms of distance from the construct. A performance index such as the type used in the Massachusetts accountability system and other state accountability systems is an example of such an approach. Growing closer to the standard is defined as progress across one of the achievement level thresholds established at achievement levels below the proficiency standard. This approach is not a direct measure of growth on the construct, makes no claims of precision beyond the gross performance levels, and relies on the assumption of some consistency in the relative meaning of the achievement standards across grade levels.

Regardless of the approach used, defining growth in terms of performance relative to a standard from one grade to the next ignores absolute growth in relation to the construct.

This is true because in the context of annual testing the standard that performance is compared against changes from Time A to Time B. That is, as a student moves from grade 5 to grade 6, performance at grade 6 is compared to the sixth grade standard rather than the fifth grade standard. A student performing below the proficient standard at the end of grade 5 may well be performing above that grade 5 standard by the end of grade 6 – growth relative to the construct. However, the student's performance may not be closer to the sixth grade standard at the end of grade 6 than it was to the fifth grade standard at the end of grade 5. If so, the student did not demonstrate growth.[4]

Summary

The preceding discussion of growth highlights the complexity in conceptualizing, measuring, and interpreting student growth. None of the approaches described represent the "correct" way, or even the preferred way to consider growth. Norm-referenced measures or interpretations reported alone, however, may be regarded as less acceptable in this era of standards-based reform and accountability. In practice, most educators, parents, and policy makers probably have interest in some of the measures; and some probably have interest in all of the measures. The task, therefore, becomes to determine the extent to which it is possible to derive one or more of those growth measures from a series of large-scale assessment such as the MCAS tests.

Part III: Vertical Scales & Vertically-Moderated Standards

Both vertical scales and vertically-moderated standards are tools for showing the progress of students over time. Beyond that general similarity, however, the two concepts have very little in common and serve very different purposes. In this section of the paper, vertical scales and vertically-moderated standards are defined and discussed in terms of their relation to a) the three conceptions of student growth described in the previous section, and b) the goals and purposes of MCAS. An overview of the assumptions and processes involved in developing vertical scales and vertically-moderated standards is also provided.

Vertical Scales

Vertical scales are a common and longstanding component of K-12 large-scale assessment. A core component of norm-referenced standardized tests (NRT), Lissitz and Huynh (2003) define a vertical scale as "a single (unidimensional) scale that summarizes the achievement of students." The scale is used to directly compare the performance of students through scores earned on different grade level tests and allows for the monitoring and tracking of student performance and progress across grades (Lissitz and Huynh, 2003; Jorgensen, 2004; Young, 2006). An additional feature of vertical scales is that they are designed to be equal-interval scales – meaning that an increase of *x* points at one point on the scale is equivalent in terms of distance on the scale to a change of *x* points anywhere on the scale. That is, a 50-point increase from 200 to 250 is considered equivalent to an increase from 375 to 425 or 650 to 700.

[4] Alternatives to annual testing, such as fall (pre-test) and spring (post-test) testing within the same school year would allow comparisons against the same standard. However, this doubling of the number of test administrations, would demand a significant investment of time and money.

The appeal of vertical scales is their seeming simplicity. As Andrew Porter explains in a 2004 interview (Lockwood, 2004)

> If children are tested in the spring of the third grade, and tested again in the spring of the fourth grade with a different test, but that yields scale scores on a vertical scale, it is easy to compare the scores from one grade to another…. "While it can get technically very complicated, at its simplest it is helpful to think of it this way. If a student tests in the third grade and gets 300, and tests again in the fourth grade and gets a 350 on the test, one can say the student gained 50 points. In another third-grade-fourth-grade situation, the student may have a 300-325 gain. The comparisons make sense."

A student who earns a higher score on the mathematics test in the fourth grade than in the third grade has increased her or his achievement in mathematics. Students whose fourth-grade scores are equal to or lower than their third grade score have not increased their achievement in mathematics.

Scaled score comparisons with vertical scales make sense because they are the type of comparisons that we are comfortable making. Whether with regard to simple units of measure such as inches, pounds, dollars, temperature; or more complex units such as miles/hour, dollars/gallon, or earned runs per nine innings pitched, it is understood that larger numbers mean more of the attribute being measured. A number that is twice as large means twice as many, twice as fast, twice as expensive, twice as bad a pitcher.[5]

Vertical scales also provide a common language for discussing student achievement – what Jorgensen (2004) describes as "a common vocabulary and metric for describing a student's progress throughout his or her educational journey." Unlike other educational currency such as course grades or grade-point-average, test scores on a common vertical scale allow for the direct comparison of the achievement of a single student across teachers, grades, schools, or states. The value of a common language for discussions of student achievement should not be underestimated.

In the context of our discussion of student growth, vertical scales describe growth relative to the construct (or relative to self). Considered by itself, a 50-point gain from 300 on the third grade test to 350 on the fourth grade test tells us that the student's achievement has increased, but provides no information about the student's performance relative to other students and no information about the student's performance relative to a standard. In short, we know that the student has gained 50 points in mathematics achievement, but what does that mean? How can scores of 300 and 350 or a gain of 50 points be interpreted? What information do they provide and is it the information that people want to know?

It is likely that there are two types of information that educators and parents would use to place those scores in context and make use of them. One type of information is a direct

[5] Note that we are also likely to say that it is twice as hot, although the temperature scale lacks the absolute zero point common to the other referenced scales.

interpretation of the scale: a description of the mathematics knowledge and skills that are represented by scores of 300 and 350. Another type of information is normative or comparative. It is reflected by questions such as:

- What percentage of students scored 350 or higher?
- What percentage of student increased their achievement at least 50 points?
- Does a score of 350 meet the proficiency standard for fourth grade mathematics?
- Does a 50-point gain from third grade to fourth grade place the student on track to be proficient by the end of the sixth grade?

Those normative questions shift the focus from growth relative to the construct to growth relative to others and growth relative to the standard. It is readily apparent that the vertical scale alone is not sufficient to provide the information needed to answer those questions. It may also be apparent that a vertical scale is not necessary to answer most of the questions posed. (The exception being that determining the percentage of students with a 50-point increase in achievement across grade levels requires a vertical scale.)

To those familiar with vertical scales, however, it may also not be apparent how the vertical scale will provide the direct interpretable information about mathematics achievement. Historically, vertical scales have not been used in K-12 assessment to provide absolute information about student achievement. In fact, although vertical scales were constructed for virtually all K-12 NRT, the scales played a relatively minor role in the reporting, interpretation, and use of results from those testing programs. The primary scores reported for students were norm-referenced scores such as percentile ranks, grade equivalents, or stanines. Group scores commonly reported were median percentile rank, percentile rank corresponding to the mean scaled score, or distributions of students across percentile ranks. Among these statistics, only grade equivalent scores (perhaps the most misunderstood and misinterpreted of the statistics) required the use of the vertical scale.

Ironically, one use of NRT in which the scaled scores from the vertical scale were commonly used by educators is now prohibited under NCLB. Results of students administered "out-of-level" tests could be reported as scaled scores and then translated to the appropriate grade-level norm-reference scores, if desired.

The most likely opportunity for the use of the vertical scale on NRT was in the reporting of gain scores (growth scores) to meet federal Title 1 reporting requirements. Title 1 reporting required the aggregation of gains across the various grade level forms of the test administered in the school. The use of the underlying vertical scales for this purpose, however, was eschewed in favor of the development of a new statistic- the normal curve equivalent (NCE). NCE scores offered the advantage of a common language to be used across NRT. Also, with a mean of 50 and an effective range from 0 to 99 NCE provided an equal-interval scale that maintained the look-and-feel of percentile ranks – the coin of the NRT realm.

Deriving Meaning from Vertical Scales

As described previously, the siren song of simplicity is the allure of the vertical scale. The construction of a scale to provide precise information about student achievement across grade levels in mathematics and English language arts, however, requires a Herculean feat of test development that may exceed the current state of the art of educational measurement. The constructs of mathematics and English language that are defined through the state's *Curriculum Frameworks* and measured through the MCAS tests may be too broad, complex, or multidimensional to yield a true vertical scale.

Vertical scales in education are also commonly referred to as "developmental scales" (Jorgensen, 2004; Lissitz and Huynh, 2003; Florida Department of Education, 2002; Young, 2006). The implication is that scores on the scale reflect a particular level of development in the construct being assessed (i.e., mathematics or English language arts). For test scores to reflect a particular level of development, however, there must be an underlying theory of learning or student cognition that defines development across the construct (Pellegrino et al., 2001). Kennedy (2005) in an overview of the BEAR Assessment System succinctly provides a description of the role of a theory of learning in the development of assessments and scales:

> The BEAR Assessment System includes four building blocks for constructing quality assessments: Construct Maps, Items Design, the Outcome Space, and the Measurement Model…. A *construct map,* which defines a latent variable or construct, is used to represent a cognitive theory of learning consistent with a developmental perspective. This building block is grounded in the principle that assessments are to be designed with a developmental view of student learning. This means that the underlying purpose of assessment is to determine how students are progressing from less expert to more expert in the domain of interest, rather than limiting the use of assessment to measure competence after learning activities have been completed. (pp 2-3)

In this model, higher scores on the vertical scale reflect "increasing sophistication in learning" and lower scores reflect "decreasing sophistication in learning."

Current state assessments, including the MCAS tests, fall short of Kennedy's description of a "quality assessment" for two important reasons. First, the primary purpose and use of the MCAS tests are, in Kennedy's terms, "to measure competence after learning activities have been completed" or at least at the point where learning activities are supposed to have been completed. That is, although performance on the MCAS tests are reported along a continuum 200 to 280 or from *Warning(Failing)* to *Advanced*, the focus of the test (and the *Curriculum Frameworks*) is on knowledge and skills that should be acquired at the time of testing. The test is not designed to measure the developmental progressions that a student follows from September through the time of testing. Lower test scores reflect less proficiency, but not a neat mapping back to a clearly defined developmental perspective of learning. Second, the constructs of Mathematics and

English Language Arts across grades K-12 or even 3-8 are too broad and have not been defined to the extent needed to build a developmental vertical scale. Attempts to define clear developmental sequences in education have generally been limited to well-defined, narrow tasks measuring skills developed over a relatively short period of time. What people are really interested in, however, is some type of generalizable skill and knowledge (e.g., the ability to do new tasks and problems, not just those exactly drawn from what was taught) (Gong, 2006). Attempts at developmentally defining broader constructs acquired over longer periods of time across years still focus on aspects of the curriculum such as spelling or arithmetic reasoning, that would be considered smaller components of the larger constructs of mathematics or English language arts. The current standards for test development and current application of learning theory do not support the construction of the type of developmental assessments and scales described by Kennedy (Pellegrino et al., 2001). Significant changes to the current standards for test development and the current level of application of learning theory to test development are needed to construct tests that could support developmental scales across multiple grade levels – and perhaps even within a single grade level.

There is no question that the MCAS tests in English Language Arts and Mathematics measure student performance on a wide range of learning standards within and across grades. Although there is evidence that performance on these discrete standards is correlated (at least within grades), deriving precise meaning from individual test scores on tests similar to the MCAS tests has proven problematic. Results of recent studies attempting to define the knowledge and skills of students scoring at the Proficient level on state assessments within a grade level demonstrate that at best it is possible to make assertions such as students with a particular scaled score have a certain mid-range probability (e.g., .60 - .75) of possessing particular knowledge and skills (Achieve, 2003). Within a single grade level test, there are many millions of ways for students to achieve the same scaled score. Additionally, analyses of state assessment results have traditionally shown that there is not necessarily a developmental continuum in the knowledge and skills that discriminate among high-scoring and low-scoring students. In English language arts, particular vocabulary items are often found to be the most discriminating items at the highest levels of performance. In mathematics, basic statistical concepts which would be considered at a low level developmentally have been beyond the grasp of high achieving students – likely more a reflection of the focus of instruction than the difficulty of the concept.

The problems described above of deriving meaning from a scale are only exacerbated with vertical scales extending across multiple grade levels. In any application of vertical scaling in large-scale K-12 assessment there is considerable overlap among the range of student performance across grades. That is, it is possible for students across a wide range of grade levels to achieve the same test score on the vertical scale. The variation in content taught and assessed across grade levels makes it impossible to derive a content-based (or construct-based) interpretation of the test score achieved by a student outside of the context of the particular grade level at which the student was tested (Schafer, 2006; Lissitz and Huynh, 2003).

The lack of a solid theory of learning as a foundation for test development is the primary contributor to the difficulty in deriving meaning from vertical scales that can be used to inform instruction for individual students or groups of students. Test development, however, is not the only barrier to the construction of interpretable developmental, vertical scales. The state of the art in the technical areas needed to construct vertical scales is also a major contributor. In the following section, a summary of practical issues in the construction of vertical scales is presented.

Construction of Vertical Scales

Notwithstanding the long history of the development of vertical scales for NRT, the technical criteria for the construction of vertical scales are still emerging and evolving. Young (2006) states, ‘There is no consensus in the measurement community on what constitutes best professional practice for developing and validating vertical scales.’ He indicates that the joint *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) ‘mention vertical scaling only in passing.’ and that although the current literature provides questions to guide scale development, there are not clear choices in the selection of design and methodology. What is clear about the construction of vertical scales is that the different choices in the development process will yield different results – that is, different vertical scales.

Although there may be considerable variation in the details of the methods used to construct vertical scales, there are some common elements to the procedures routinely applied in K-12 large-scale assessment:

- Vertical scales are constructed by linking test forms at adjacent grades. The vertical scale is actually a chain of linked scales rather than a single unitary scale.
- Vertical scales are constructed by “administering an embedded subset of items to different students at two educational levels, typically one year apart, and linking all the items to a common scale through the comparative performance of the two groups of students on the common items” (Schafer, 2006).
- The construction of vertical scales is primarily an empirical process based on student performance on test items.
- The goal of the process is to determine the “distance” between pairs of adjacent grades on the set of common items (i.e., the amount of growth).

These procedures produce scales that appear credible – showing, for the most part, an increasing level of student achievement across grade levels. There are significant issues, however, that must be considered in evaluating the validity of the scales and the practicality of constructing them.

The common practice of administering items to students at adjacent grade levels (i.e., items not assessed at their current grade level) may have to be examined in the current standards-based, NCLB-driven assessment environment. Unlike traditional NRT, current state assessments are built to specific grade-level expectations and there is an intense focus on the alignment of standards and assessments. The inclusion of items measuring standards at a prior grade level or future grade level on an operational test form (even if those items are not included in scoring) might be more problematic both from a policy perspective and from a measurement perspective than in previous years. As Lissitz and Huynh (2003) explain, one appealing solution to this issue is to focus on knowledge, skills, or standards that are common across years. They explain, however, that although such an approach may address the common dimensions across grades, important grade-specific content knowledge and skills may be ignored in describing growth across grades.

The selection of the specific items to embed at adjacent grades and decisions of where to embed items is also critical to defining the scale. The embedding of items at only one adjacent grade (e.g., prior) will results in a very different vertical scale than the embedding of items at both prior and following grades (e.g., including grade 4 and grade 6 items on a grade 5 test) (Vukmirovic, 2003; Hoffman, 2006). The extent to which the embedded items are representative of the content area and the difficulty of those items also will impact the vertical scale. There is wide agreement that very difficult and very easy items can be constructed for most K-12 content standards. The balance of difficult/easy items among the embedded items will also impact the vertical scale.

There is also the issue of which students' performance contribute to the scale. Unless there is considerable overlap across the individual scales at particular grade levels, it is possible to conceive of the students in the tails contributing most to the construction of the scale. That is, the link in the vertical scale may be formed at the point where measurement is weakest – or least precise. The least able "Grade 4" students who get most of the embedded Grade 3 items incorrect, and the most able "Grade 3" students who get almost all of the Grade 3 items correct will form the basis for the link. In practice with norm-referenced and standards-based tests, however, the overlap in scales across adjacent grades and even non-adjacent grades has been quite considerable – again, raising additional questions about the interpretability of the vertical scale.

It may not be too much of an exaggeration to state that there are an infinite number of vertical scales that could be constructed from a given set of items and students – there are certainly a very large number of them. If we attempted to construct all of the possible scales, perhaps a clear pattern would emerge that would demonstrate that some scales were more likely than others (e.g., perhaps the vertical scales would be normally distributed). It would then be possible to select a scale that made sense based on the empirical evidence. There would be no evidence, however, that the scale selected was the "correct" scale based on the adopted theory of learning.

At this point in K-12 large-scale assessment, without well-defined theories of learning, vertical scales are being developed on the basis of general principles related to the balance of content and growth. The scales that are constructed are probably empirically the most likely scales. Those empirical scales, however, reflect an interaction of intended and enacted curriculum and instruction that may or may not be consistent with a well-defined developmental theory of learning or even loosely-defined expectations of the "difficulty" of a concept.

Vertically-Moderated Standards

Unlike vertical scaling, a longstanding practice with NRTs that is currently being applied to state assessments, the concept of vertically-moderated standards is recent and has been developed for state assessments[6]. Specifically, the practice of developing vertically-moderated standards has emerged as a potential solution to the accountability challenges posed by the assessment and accountability requirements of NCLB (Cizek, 2005). As proposed by Lissitz and Huynh (2003) vertically-moderated standards are defined as follows:

> 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.

Huynh and Schneider (2005) expand the definition to explain the two basic elements of vertically-moderated standards:

> 1. A set of common policy definitions for the ALs [achievement levels] (such as basic, proficient, and advanced) is used for all grades.
> 2. A consistent trend line is imposed on the percentage of students in important performance categories. One such important category is likely the state-defined proficient category in the AYP environment. For assessment programs with a focus on instructional remediation, an important category may be the below basic category. A consistent trend line may reflect no change, a moderate level of increase, or a moderate level of decrease across the grades. (p. 106)

They describe vertically-moderated standards as “essentially a mixture of ‘policy equating’ and ‘linear statistical adjustment’.”

With regard to our three conceptions of growth, vertically-moderated standards clearly shift the emphasis from growth relative to the construct to growth relative to others and/or growth relative to a standard. The extent to which the focus of vertically-moderated standards is on normative performance versus standards-based performance appears to lie within the mixture of policy and statistical adjustment described above. That level of detail is determined at the state level and is not intrinsic to the concept of vertically-moderated standards.

It is also clear that, as defined above, the concept of vertically-moderated standards is a post hoc approach applied to standard setting after the development of content standards and the assessment. Cizek (2005) refers to the concept as “vertically moderated standard

[6] The term, vertically-moderated standards, and the context of high-stakes accountability are recent. The practice of assuming consistency (or constancy) in performance across grade levels or across years is well established in large-scale K-12 assessment.

setting" rather than vertically-moderated standards. The purpose of vertically-moderated standards is to produce a set of coherent performance standard results (i.e., distribution of students within or across achievement levels) across grades 3 through 8 to meet the accountability requirements of NCLB.

To what extent, therefore, is the process of creating vertically-moderated standards tied directly to the content standards, student performance on the assessment, and/or a theory of learning across grade levels? The answer to that question depends on how the process is applied. As an example of the process, Lissitz and Huynh describe the following procedure for setting standards for an assessment program across grades 3 through 8:

1. Traditional standard setting methods are used to establish achievement level cut scores at grade 3 and grade 8.
2. Those cut scores were reviewed and adopted by the state.
3. Cut scores for tests at grades 4 through 7 were interpolated based on the achievement level results at grades 3 and 8. In this case, a "single growth curve line" was used in the interpolation.

There are key points in the process where the balance can shift from norm-referenced to standards-based. One key point is in the design of the standard setting activities at the anchor grades. To the extent that those standard setting procedures are tied directly to the content standards, the results of those procedures will establish a content-based foundation for the remainder of the process. However, to the extent that the initial standard setting is overly influenced by impact data or other policy considerations, a link to the content standards will be weakened – as is the case in any standard setting process. The next key point in the process, obviously, is in the method(s) of interpolation used to derive results at the remaining grades. An interpolation method based on existing knowledge of student progress across grades and knowledge of the content standards across grades would be more directly linked to the content standards than a simple decision to draw a straight line between the anchor points. Extending that decision process across the achievement levels is also critical to the process. The decision whether to apply the same interpolation rule to each of the achievement level should be based in something more than expediency.

The relationship between vertically-moderated standards and vertical scales

Lissitz and Huynh (2003) propose vertically-moderated standards as an alternative approach to vertical scaling in a paper with the pre-colon title 'Vertical Equating of State Assessments.' This may have contributed to confusion among state department personnel, if not among measurement practitioners, about the relationship between the two concepts. In fact, there is little, if any, relationship between the two concepts. With or without vertical scales, states would still be required to establish performance standards for each test at grades 3 through 8. The process of vertically-moderated standard setting can be applied equally well in a system of independent horizontal scales and a system of vertical scales.

The “alternative” offered by vertically-moderated standards is that it is not necessary to perform any formal linking (e.g., equating, calibration, etc.) of different grade level tests to meet the requirements of NCLB. In essence, the linking of achievement level results replaces the linking of the assessments. In practice, like many statistical procedures, vertically-moderated standard setting will produce “credible” results almost regardless of the nature of the underlying assessments. It is assumed, however, that states have followed sound practices in developing content standards that are aligned across grades and have developed assessments that are aligned to those standards. Additionally, it is assumed that there will be some relationship between the achievement level descriptors and the performance on the state assessment of students classified at each achievement level.

Vertically-moderated standards, Vertically-articulated standards, Vertically-aligned standards

Consistent with its recency, the terminology surrounding vertically-moderated standards has not yet stabilized. As previously mentioned, Cizek (2005) introduces the term vertically-moderated standard setting. In the same article, he also refers to the process as the “vertical articulation of standards.” The term vertically-aligned standards is also common as a reference to the alignment of achievement standards across grades in discussions of accountability and growth related to NCLB (Marion, 2006). It appears that these terms are used interchangeably and that practice is obscuring a critical component of vertical alignment.

Lost in the lack of precision in terminology is the concept of the a priori establishment of achievement standards based on the content standards and a theory of student learning or student progress across grades. A theory-based concept of vertically-aligned standards eliminates the overt link to accountability and its emphasis on “what is” at the expense of “what should be.” Rabinowitz (2006) described the establishment of achievement standards (i.e., standard setting) as the intersection between assessment and accountability. The requirement of a consistent trend line in vertically-moderated standard setting is a logical reflection of “what is,” and may be a necessity for accountability. For accountability purposes, some might even argue for the more stringent requirement of a flat trend line across grades. A focus on the content and “what should be” may produce inconsistent results across grades, but results that are more defensible in terms of the achievement level descriptions and the state’s expectations for student performance. Wise et al. (2005a, 2005b), in a series of studies conducted in conjunction with efforts of the CCSSO Technical Issues in Large-Scale Assessment (TILSA) SCASS, describe methods for measuring the vertical alignment of content standards rather than the results of achievement standards. In an ideal system, of course, there would be little distinction between the two. In practice, however, states will need to strike a balance between adherence to the intent of the content standards and the demands of accountability.

Part III: Conclusion

The concept of vertically-moderated standards, as defined above, forced a shift in our discussion from assessment to accountability. In our previous discussion of growth and vertical scales, although there was the expectation that results should make sense across grade levels, there was not as direct a link between the process for defining or deriving scores and accountability. In an ideal view, or perhaps it is a utopian view, the focus of education and educational measurement would remain on the content and a theory of student learning for as long as possible.

At some point, however, even without the sword of a high-stakes accountability program hanging overhead, the focus will shift from "simple" measurement to expectations of future performance. Simple measurement can provide information about a student's current performance, past performance, and growth between two points in time. As a tool to inform and improve education, however, simple measurement alone is not sufficient. Action based on information is required and appropriate interpretation and use of information requires an understanding of how much is good enough and what can be expected in the future. The questions posed at the beginning of this paper included questions about whether Jane and Johnny improved as much as *expected* this year and how much can they be *expected* to improve next year. These are not accountability questions as much as they are questions about progress toward "increasing sophistication in understanding" of the construct. Does a vertical scale help or hinder efforts to answer questions regarding expectations?

Even with all of the problems discussed regarding content-based interpretation, the simplicity of the vertical scale is still appealing and tempting. A common, equal-interval scale can make comparisons straightforward – even if test scores are not grounded in developmental theory. A gain of 50 points at one point on the scale is equivalent to a gain of 50 points at any point along the scale. It would be nice to be able to say that we expect a gain of 50 points per year from every student. Unfortunately, once again, vertical scales do not appear to be the solution.

Student progress across grade levels is uneven. One feature of vertical scales as they are currently constructed is that they clearly demonstrate this unevenness (Schafer, 2006). The developmental scale constructed for the Reading tests of the Florida Comprehensive Assessment Test (FCAT) provides a prime example of the issue. The scale ranges from 0 to 3000 across grades 3 through 10 and was anchored with a score of 1300 as the average scaled score at grade 3 in 2001. Student performance on the FCAT is classified into five achievement levels (Level 1 – Level 5). A student performing at the Level 2/Level 3 threshold at grade 3 must increase 258 scaled score points to remain at the Level 2/Level 3 threshold in grade 4. Maintaining that position from grade 4 to grade 5, however, requires an increase of only 54 points. Across grades 5 through 10, the annual increases in scaled score points required to remain at the Level 2/Level 3 threshold are 112, 93, 167, 90, and 96, respectively. Additionally, *expected* gains in performance are not consistent across the scale within a single grade span. From grade 4 to grade 5, as stated above, 54 points are needed to remain at the Level 2/Level 3 threshold. However, remaining at the Level 1/Level 2 threshold requires a gain of 27 points and remaining at the Level 4/Level 5 threshold requires a gain of 94 points. The vertical scale alone does little to facilitate the communication of expected increases in student performance from one grade level to the next.

Another common characteristic of vertical scales is that annual growth in performance tends to decline across grade levels. Growth is highest in the primary grades and begins to flatten as students enter the grades assessed under NCLB. This is evident on scales developed for NRT as well as on the FCAT developmental scales. On the 3000 point FCAT scale, the Level 2/Level 3 achievement level threshold increases by only 500 points from grade 5 through grade 10. On an NRT with a 1000 point scale, the scale score of students at the 50th percentile increases only 37 points from 639 to 676 across grades 4 through 8. This lack of reported growth across grade levels coupled with standard errors of measurement of approximately 10 points per grade level at this point in the scale makes the measurement of individual student growth from grade to grade very difficult.

At the outset, it appeared that vertical scales were best suited to provide information about student growth relative to the construct. However, if there are substantial limitations on the use of vertical scales for both the measurement and interpretation of student growth relative to the construct from grade to grade, then they appear to offer little advantage to separate reporting scales across grade levels. By design, vertical scales offer do not address growth relative to others and growth relative to the standard. In fact, reporting scales such as the MCAS scale with fixed achievement level cut scores may offer a definite advantage in communicating performance relative to the achievement level standard from grade to grade.

With or without a vertical scale, states must do more work to communicate information about the interpretation and use of test scores. Information on the meaning (and lack thereof) of individual test scores must be developed and communicated in a user-friendly manner to appropriate audiences of educators and parents. Studies of *expected* changes in performance relative to others and relative to the standard must be conducted to increase the usefulness of test scores that are being reported annually across grades 3 through 8. An increased understanding of changes in performance relative to the construct and the appropriate role of a state assessment in measuring those changes must be developed. It is not clear, however, that the development of a vertical scale is either necessary or helpful to the process of communicating useful and accurate information about student performance and growth to educators and parents.

References

Achieve, Inc. & The National Center for the Improvement of Educational Assessment, Inc. (2003) *Evaluation of Cut Scores on Indiana's ISTEP+ Assessments.* Washington, DC: Achieve, Inc.

Cizek, G.J. (2005) *Adapting Testing Technology to Serve Accountability Aims: The Case of Vertically Moderated Standard Setting* Applied Measurement in Education. Vol. 18. No. 1 pp 1-9.

Florida Department of Education (2002). FCAT Developmental Score Scale. Memo to District Superintendents. Downloaded from http://www.firn.edu/doe/sas/fcat.htm. October, 2006.

Gong, B. (2006) *Establishing Learning Goals for Formative Assessment*. Presented at the Edward F. Reidy, Jr. Interactive Lecture Series. Nashua, NH. October 2006.

Hoffman, R.G., Ford, L., & Lozzi, D. (2006) *Where is Student Academic Growth? Part One: Vertical Scaling*. Presented at the 36th Annual National Conference on Large-Scale Assessment. San Francisco, CA. June 2006.

Huynh H. & Schneider, C. (2005) *Vertically Moderated Standards: Background, Assumptions, and Practices*. Applied Measurement in Education. Vol. 18. No. 1 pp 99-114.

Jorgensen, M.A. (2004) *The Value of the Stanford Scale as a Common Metric, Revision 1*. San Antonio, TX: Harcourt Assessment, Inc.

Kennedy, C.A. (2005) *The BEAR Assessment System: A Brief Summary for the Classroom Context.* Technical Report Series No 2005-03-01. Berkeley, CA: Berkeley Evaluation & Assessment Research Center.

Lissitz, RW & Huynh H (2003) *Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability*. Practical Assessment, Research & Evaluation vol. 8 no. 10. April 2003.

Lockwood, A.T. (2005) *Expert Opinion: Andew Porter and Value-Added Assessment*. On The Road To Accountability. Northwest Regional Educational Laboratory. Vol 10. No. 5. Summer 2005.

Marion, S. & Gong, B. (2006) *What's a leader to do? State and district leaders' roles in supporting local assessment systems*. Presented at the Edward F. Reidy, Jr. Interactive Lecture Series. Nashua, NH. October 2006.

Massachusetts Department of Education (2006) *Summary of 2006 MCAS State Results*. September 2006.

Pellegrino, J.W., Chudowsky, N. & Glaser, R. (eds.) (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academy Press.

Rabinowitz, S. (2006) *Standard Setting Under NCLB: A 2006 Perspective*. Presented at the 36th Annual National Conference on Large-Scale Assessment. San Francisco, CA. June 2006

Schaefer, W.D. (2006) *Growth Scales as an Alternative to Vertical Scales*. Practical Assessment, Research & Evaluation. Vol 11. No. 4. March 2006.

Vukmirovic, Z. (2003) *Integrating the elements of Adequate Yearly Progress: Vertical scaling and standard setting*. Presented at the 33rd Annual National Conference on Large-Scale Assessment. San Antonio, TX. June 2003

Wise, L.L. & Alt, M. (2005a). *Assessing Vertical Alignment*. Washington, D.C.: Council of Chief State School Officers.

Wise, L.L, Zhang, L., Winter, P, Taylor, L., & Becker, D.E. (2005b). *Vertical Alignment of Grade-Level Expectations for Student Achievement: Report of a Pilot Study*. Washington, D.C: Council of Chief State School Officers

Young, M.J. (2006) *Vertical Scales* in Handbook of Test Development. S.M. Downing and T.M. Haladyna (Eds.) Mahwah, New Jersey: Lawrence Erlbaum Associates.