Measuring the Power of Learning.™

Research Report
ETS RR–16-29

Vertical Articulation of Cut Scores Across the Grades: Current Practices and Methodological Implications in the Light of the Next Generation of K–12 Assessments

Priya Kannan

August 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Vertical Articulation of Cut Scores Across the Grades: Current Practices and Methodological Implications in the Light of the Next Generation of K–12 Assessments

Priya Kannan

Educational Testing Service, Princeton, NJ

Federal accountability requirements after the No Child Left Behind (NCLB) Act of 2001 and the need to report progress for various disaggregated subgroups of students meant that the methods used to set and articulate performance standards across the grades must be revisited. Several solutions that involve either *a priori* deliberations or *post-hoc* adjustments have been offered over the years. In this paper, I provide a methodological review of the alternative cut-score articulation methods, including some novel solutions (e.g., using predictive methods) that have been proposed in the context of the next-generation K–12 assessments. In systematically evaluating these methods, I focus on the psychometric challenges they might present and the practical feasibility of their operational implementation. In addition, results from a survey of several state departments of education help to provide information on the prevalence of these methods across the states. Overall, this review shows that none of the alternative methods is completely free of limitation; yet, each method provides solutions that are appropriate for addressing certain methodological and practical requirements. Therefore, in the context of the next-generation assessments and the need to identify students who are on track to being college and career ready, practitioners are advised to consider a combination of methods and cautioned against overreliance on any single method.

Keywords Standard setting; large-scale assessment; accountability testing

doi:10.1002/ets2.12115

Accountability mandates have become central in reforming and restructuring methods for teaching and learning (Koretz, 2008, 2010; Linn, 2008; Shepard, 2008). These methods include ones used to set performance standards and classify students (O'Malley, Keng, & Miles, 2012). Novel solutions for recommending and articulating cut scores across the grades were necessitated by the No Child Left Behind (NCLB) Act of 2001 (NCLB, 2002). This federal reporting mandate required that multiple performance levels be established at Grades 3 through 8, and it also required states to demonstrate annual progress for various disaggregated student subgroups at each of the multiple performance levels. Therefore, states were faced with the need to revisit the ways in which performance standards and cut scores were articulated across the grade levels. Several novel solutions were offered. In this paper, I provide a methodological overview of the alternative cut-score articulation methods that have been proposed over the years. The prevalence of these methods across the states is taken into account in evaluating their practical feasibility within the current policy context.

Policy Context That Steered the Need for Articulation Across Grades

The NCLB Act of 2001 (NCLB, 2002) required states to establish multiple performance levels at Grades 3 through 8 and one high school grade level and to demonstrate 100% proficiency for all student subgroups in English language arts (ELA) and mathematics by 2014. However, results from large-scale cross-state analyses soon began to reveal that increasing accountability does not necessarily increase student achievement or retention (Carnoy & Loeb, 2002; Center on Education Policy, 2007; Nichols, Glass, & Berliner, 2006). Moreover, several attempts to map state proficiency standards onto a common scale (e.g., National Assessment of Educational Progress [NAEP]) revealed considerable variability in these standards, such that being "proficient" did not mean the same across states (Bandeira de Mello, 2011; Barton, 2009; Linn, Baker, & Betebenner, 2005). Overall, controversies over the effectiveness of the NCLB and concern about the international competitiveness (Barton, 2009) and college readiness of our high school graduates resulted in efforts to establish common

Corresponding author: P. Kannan, E-mail: pkannan@ets.org

standards and expectations for students across the country (i.e., the Common Core State Standards, see Council of Chief State School Officers [CCSSO] and the National Governors Association [NGA], 2010).

However, although the NCLB focus on universal proficiency has been criticized (e.g., Ho, 2008) for its inaccurate interpretations of achievement, the federal accountability requirement to demonstrate adequate yearly progress (AYP) and growth continues, with some executive orders for waivers. For example, in August 2013, the United States Department of Education (USED) invited state departments of education to request flexibility and waivers on the provisions of the NCLB. The flexibility plan offered a reprieve from the 2014 deadline for demonstrating the 100% proficiency referred to above and allowed states to continue to move forward in developing their "ambitious, but achievable" (USED, 2013) annual measurable achievement objectives (AMAOs) that provide meaningful goals to guide achievement. Overall, such changes in the accountability mandates (i.e., need for annual measurement and reporting of student achievement and progress at all grade levels) prompted states to rethink the ways in which their performance standards are set and articulated across the grades.

In the pre-NCLB era, standard setting had predominantly been conducted as a grade-specific activity (Cizek & Agger, 2012). Subject matter experts (panelists) in a content area were recruited for grade-level meetings and were responsible for recommending the cut scores for the respective grade-level assessment. Each of these groups of panelists developed performance-level descriptors (PLDs) that were not necessarily linked across the grades, and sometimes even used different standard-setting methodologies across grades, which led to a unique set of cut scores for each grade and content area (Cizek & Agger, 2012). The cut-score decisions for each grade were sometimes influenced by discussions held across the grade levels, but not necessarily required by design. Therefore, when the need to demonstrate annual progress within student subgroups for multiple performance levels became required by federal law, school districts were unable to explain the fluctuations in the percentages of students classified at each performance level across the grades (Cizek, 2005; Lewis & Haug, 2005).

In order to make meaningful inferences about annual progress when independent cut scores are set for each grade, the content standards should have been developed to reflect an increasing level of complexity along an underlying developmental continuum, and the tests across grade levels should similarly have been designed to measure these developmentally progressive content standards (Lewis & Haug, 2005). However, because this was typically not the case, cut scores set independently across the grades did not support the intended inferences about annual progress required by the federal mandates. As a result, states began to realize that grade-level tests could no longer be considered in isolation when cut scores are recommended. Rather, the cut scores across content areas and grade levels needed to be consistent. This resulted in the advancement of several novel standard-setting solutions for vertically articulating performance standards across the grades. Moreover, the federal law, NCLB, was soon extended to incorporate individual student growth in accountability calculations, which led to the introduction of the growth model pilot program in 2005 (Ho, Lewis, & MacGregor Farris, 2009; Hoffer et al., 2011). Schools were required to demonstrate grade-to-grade growth relative to the grade-level standards (Cizek & Bunch, 2007), and one of the first solutions to demonstrating longitudinal growth for individual students was based on vertically equated score scales across grades (Yen, 2007).

Vertical Scales: A Satisfactory Solution?

Vertical scales can provide an ideal solution and offer several practical benefits (e.g., Petersen, Kolen, & Hoover, 1989; Yen, 2007) for tracking student growth and progress over the grades. But they are challenging to develop (e.g., Kolen, 2011; Patz & Yao, 2007; Yen, 2007). However, if they are successfully implemented, vertical scales could be used to accurately track individual student growth on a common scale by comparing scores for the same student from 1 year to the next, and they have been deemed more efficient than any alternative solution (Kolen, 2011; Patz & Yao, 2007).

A vertical scale may be used to articulate the cut scores across the grades. If a vertical scale is successfully established, then cut scores may simply be articulated based on concurrent or chained calibrations of linked items (see Ito, Sykes, & Yao, 2008) across grade-level tests (e.g., Delaware and Florida have developed assessments based on vertical scales and calibrations based on vertically linked items to develop articulated cut scores). Moreover, with the upsurge of interest in evidence-based standard setting (McClarty, Way, Porter, Beimers, & Miles, 2013), a trend has been observed toward obtaining cut scores using external benchmarks — for example, comparisons to performance on other related assessments such as freshman college courses, NAEP, Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) — for matched students as the criteria for recommending the high school cut

scores, and then articulating cut scores for lower grades based on a vertical scale. Texas and West Virginia are examples of some states that have attempted this approach (Kannan, 2014).

However, in order for these advantages to be realized, a vertical scale should first be established in which the grade-level assessments are linked to each other using common items. Repeatedly, researchers (e.g., Kolen, 2011; Patz & Yao, 2007; Yen, 2007) have acknowledged that vertical scales are challenging to develop. The underlying progression of content strands across the grades, the choice of common items, the choice of grade spans on which the vertical scale is developed (i.e., a vertical scale that spans adjacent grades or all grade levels), and design decisions about using common items and/or equivalent groups are all factors that have methodological and practical implications and influence the successful development and implementation of a vertical scale (see Kolen, 2011, for a clear exposition of the challenges and considerations in developing vertical scales).

Finally, there are several practical challenges to successfully developing a vertical scale (e.g., Harris, 2007; Kenyon, MacGregor, Li, & Cook, 2011; Martineau, 2006; Yen, 2007). Primary among these is the unavoidable construct shift in certain domains that results in multidimensional latent constructs. For instance, prior to the Common Core standards, one primary issue with content standards across the states was that the content across the grades was not developmentally integrated and vertically articulated. The content of fifth and sixth grade mathematics, for example, might be totally unrelated, such that students from subsequent grades do not learn any more of the previous grade's content. Even when a conscious effort is made to articulate the content standards, such as with the Common Core, the constructs in certain domains (such as mathematics and science) tend to vary greatly from middle to high school (Martineau, 2006) and even from grade to grade in high school (e.g., from a focus on algebra in Grade 9, to trigonometry in Grade 10, to precalculus in Grade 11). In addition, other overwhelming challenges in developing a vertical scale, such as problems with obtaining a matched sample and appropriateness of content of the off-grade items, deter most practitioners from embarking upon the challenging task of creating vertically scaled assessments.

Vertical Articulation of Cut Scores: A Methodological Overview

Although research on overcoming the challenges in developing and using vertical scales continues (Briggs, 2013; Briggs & Weeks, 2009; Dadey & Briggs, 2012; Kolen, 2011), immediate remedies were deemed necessary to meet the federal accountability mandates post-NCLB. States were compelled to revisit their cut scores so that the consistency in the strictness or lenience (i.e., rigor) of cut scores is maintained across the grades. As a result, Lissitz and Huynh (2003) first offered vertically moderated standard setting (VMSS) as a criterion-referenced alternative to assessing individual student growth. VMSS was offered as an alternative to vertical scaling because moderation makes weaker assumptions than equating or scaling (Kolen & Brennan, 2014), can be used when the tests are not assumed to be measuring the same construct, and is appropriate when no statistical data on their relation can be obtained to support a projection. Statistical moderation was therefore recommended by Lissitz and Huynh as a way to achieve consistency in the percentages of students classified in each performance level across the grades. Such an attempt at statistically moderating, however, does assume that instructional efforts remain relatively uniform and that student achievement does not fluctuate too much across the grades (Huynh & Schneider, 2005). Ferrara, Johnson, and Chen (2005) suggested that, in general, vertical articulation of cut scores should be built upon a two-step process: First, the content standards should be articulated, and second, the instruction in each grade should not only build upon the previous grade's instruction, but also be developed with the intention of preparing students to succeed in the subsequent grade.

In response, several researchers (e.g., Ferrara et al., 2005; Huynh, Meyer, & Barton, 2000; Lewis & Haug, 2005; Lissitz & Huynh, 2003) proposed a model of standard setting that involved such a step: identifying an across-grade alignment model. They argued that the cut scores should be comparable across grade levels within a content area, across content areas within a grade level (where feasible), and across standard-setting methods used. If this model is not followed in practice, it would result in inconsistent standards across the grades and send mixed messages to various stakeholders that might prompt them to question the validity of the testing program and the educational system in general (Lewis & Haug, 2005). Researchers (e.g., Ferrara et al., 2005; Ferrara et al., 2007; Lewis & Haug, 2005) therefore recommended that developmental expectations for growth across the grades should guide the standard-setting and cut-score articulation processes across the grades.

Over time, with the increasing popularity of theoretical perspectives such as learning progressions (Smith, Wiser, Anderson, & Krajcik, 2006) and Evidence Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003), some authors

(e.g., Bejar, Braun, & Tannenbaum, 2007) have offered holistic design frameworks that integrate educational policy, learning theory, and curriculum design in the development of content standards and cut scores (or performance standards). Bejar et al. (2007) argued that performance standards should be prospective (developed a priori in an iterative fashion, such that they influence the test development process), progressive (articulated across the grades), and predictive (indicative of performance in higher grades and explainable in terms of sound scientific constructs, that is, developmental learning progressions).

The authors (Bejar et al., 2007) further argued that the link between the assessment and the standard-setting method should happen *a priori*; that is, the performance standard should be developed first, and these standards should dictate the inferences to be made from the assessment results. However, vertical articulation of cut scores started as more of a *post-hoc* solution to federal reporting requirements. These *post-hoc* solutions enabled states and testing programs to make cross-grade adjustments to recommended cut scores that could help meaningfully explain the percentages of students who are classified within each performance level.

Irrespective of whether the solution provided is holistic or *post hoc*, vertical articulation of cut scores would help achieve consistency in the strictness or lenience (i.e., rigor) of cut scores across the grades (Cizek & Bunch, 2007; Ho et al., 2009; Lissitz & Huynh, 2003; Lissitz & Wei, 2008). Consistency in the rigor with which cut scores are established across the grades has important implications for the ability to predict which students will attain a required performance level (*proficient*/*on track* to being college or career ready) in subsequent grades. This ability to predict could also have a substantial impact on annual progress results that have become mandated post-NCLB. Alternative cut scores can dramatically change the percentage of students who are predicted to be proficient (or on track) according to Ferrara et al. (2005) and affect accountability decisions for a large percentage of schools and districts (e.g., Ho et al., 2009). Moreover, if the cut scores are not set to the same rigor, an individual student can move back and forth in performance-level classifications across the grades. In the past decade, several alternative cut-score articulation methods have been proposed and tried out by the states. Throughout this paper, these methodological alternatives are reviewed and evaluated. However, it is not sufficient just to understand the array of solutions available to articulate cut scores across the grades; it is also important to understand the prevalence of these methods across the states and the practical implications of these alternative methods when implemented by the states. Therefore, in order to better understand current state practices for articulating content and performance standards (or cut scores) across the grades, a survey of the 50 state departments of education was conducted; 35 states responded to the survey. States were specifically asked to describe the challenges they experienced when implementing these methods. Detailed results from this survey are described elsewhere (Kannan, 2014). Selected results from this survey relevant to the main thesis of this paper are summarized in Table 1 and are discussed throughout this paper to illustrate the prevalence of the various methods across the states. Such a detailed synthesis and evaluation of the alternative methods will help inform the ways these methods may be combined optimally to produce efficient standard-setting solutions for the next generation of K – 12 assessments.

Alternative Approaches to Articulating Cut Scores

Vertical articulation solutions that have been proposed over the years can be classified in various ways. For instance, Cizek and Agger (2012) introduced several methods of classifying these procedures: for example, as front-end or back-end procedures; as empirical or policy-based procedures. Most classification methods employed, however, resulted in some degree of overlap in the overall procedure used, and a combination of these procedures is, in practice, typically implemented to result in a complete solution. For instance, Cizek and Agger classified each of the following as an independent step in the vertical articulation procedure: (a) establishing the assumptions about underlying trajectories of development (e.g., Ferrara et al., 2005; Ferrara et al., 2007; Lewis & Haug, 2005); (b) articulating content standards (e.g., through a learning progressions framework); (c) articulating PLDs based on the articulated content standards (e.g., Egan, Ferrara, Schneider, & Barton, 2009; Huff & Plake, 2010) by developing some policy-level PLDs that generalize across the grade spans (as proposed by Lissitz & Huynh, 2003, and elaborated by Egan, Schneider, & Ferrara, 2012); (d) including cross-grade discussions (e.g., Buckendahl, Huynh, Siskind, & Saunders, 2005; Lewis & Haug, 2005); and finally by (e) incorporating the specific type of statistical cut-score articulation method, such as statistical interpolations (e.g., Huynh et al., 2000) or impact-percentage smoothing (e.g., Buckendahl et al., 2005; Ferrara et al., 2005; Lewis & Haug, 2005).

Although all of the previously mentioned procedural components aid in the articulation process, they have to be used collectively in order to effectively implement and obtain vertically articulated cut scores. Moreover, some combinations

Table 1 Summary of Prominent Approaches to Vertical Articulation and Current Prevalence of These Methods Across States

Cut score articulation method	Description of general method	Alternative solutions, if any	States using this method[a]	Methodological/practical challenges
Statistical interpolations	Cut scores for the end grades (e.g., Grades 3 and 8) are recommended based on a panel meeting. The cut scores for intermediate grades (e.g., Grades 5, 6, and 7) are derived through statistical interpolations, based either on proportions of students classified or the point on the scale score. A straight line interpolation, typically based on historical data, is imposed to derive cuts for the intermediate grades. In some applications, a reactor panel composed of panel representatives and/or a TAC would review the cut scores from each grade to evaluate the appropriateness of the cut scores for intermediate grades.		AZ, DE, HI, LA, MA, MN, NC, WI	- Without longitudinal data, making assumptions about the rationale for the trajectories of development to interpolate data may be challenging. - Lack of face validity due to the absence of grade-level meetings for the intermediate grades.
Impact-percentage smoothing	To the extent that operational test administration data are available for a representative group of examinees, the impact-percentage information would consist of percentages of examinees who would be classified into each of the performance levels (e.g., *Basic, Proficient*, and *Advanced*) given the recommended cut scores. This family of methods employs a smoothing procedure used *post hoc* to decrease the differences in the percentages of students classified across the grade levels to result in reasonably comparable percentages of students at each performance category.	A standards-validation meeting is conducted after the standard-setting meetings for each grade is concluded. A reactor panel composed of policy makers, TAC, or grade-level panel representatives considers the impact data available and smooths out the differences in the percentages of students classified across the grade levels.	DE, FL, GA, MA, MN, NE, ND, VT, WV, WI	- Assumptions about the rationale for the trajectories of development. - Justifications for adjusting the percentages of students classified in each performance category, when large adjustments are needed. - Fluctuating patterns established at grade-level meetings never really alleviated.
		At the end of a standard-setting meeting, a joint (meta) panel composed of panel representatives from grade-level meetings reviews the cut scores from each grade to smooth out the differences in the impact percentages for the intermediate grades.	AZ, CT, DE, FL, GA, HI, IN, NE, OH, OR, TN, TX, WI	
Vertical scale-based methods	A vertical scale is established in which the grade-level assessments are linked to each other using some common items between the assessments. Subsequently, concurrent or chained calibrations of linked test items across the grade levels may be used to estimate test taker ability on a single underlying continuum, and cut scores may thereby be established on a single scale point across the grade levels.		DE, FL, ID	- Dimensionality of the underlying construct - Developing assessments with common items spanning grades - Relevance of off-grade items when several grades apart
Benchmarked cut scores for high school used to predict cuts at lower grades	Benchmarks are obtained for the high school assessments based on external criteria such as performance on college courses, college entrance exam scores (e.g., SAT and ACT), and scores on national and international assessments (e.g., NAEP, TIMSS and PISA). The cut scores obtained for the high school are back-translated to the lower grade levels using a variety of articulation methods.	Articulated at the same scale point using concurrent calibrations for the lower grade levels using a vertical scale	OR, TX, WV	- Correlation between predictor and criterion - Correlation between grade-level assessments - Same challenges as using vertical scales
		Logistic regressions are performed to predict the point on the adjacent lower grade where students have a specified probability (e.g., 66%) of achieving the cut at the higher grade.	MI, NC, OR, NE[b]	- Correlation between predictor and criterion - Correlation between grade-level assessments - Purely statistical, and therefore a need exists for substantiation with additional panel to validate the cut scores and supplementary articulation methods.

Note. SAT = Scholastic Aptitude Test; ACT = American College Testing; NAEP = National Assessment of Educational Progress; TIMSS = Trends in International Mathematics and Science Study; PISA = Program for International Student Assessment; TAC = Technical Advisory Committee.

[a] Only 35 of the 50 states provided responses to the survey (see Kannan, 2014).

[b] NE reported using benchmarks to obtain high school cuts, but the articulation method for lower grades is unclear.

of solutions attempted in the past could be perceived as more holistic (e.g., those that start with the development of articulated content standards and curriculum across the grades or those based on the development of vertically scaled assessments) than others (e.g., *post-hoc* statistical solutions such as smoothing) in their approach. My focus in this paper is on the actual procedures (interpolation, smoothing, vertical scaling, logistic regressions, and the like) used to articulate the cut scores across the grades. Results from a survey of the 50 states (Kannan, 2014) shed light on the prevalence of these methods across the states.

First Generation of Statistical Approaches

The series of methods first proposed in the 2005 *Applied Measurement in Education* special edition (e.g., Buckendahl et al., 2005; Ferrara et al., 2005; Lewis & Haug, 2005), which include statistical interpolation and smoothing, are grossly classified here as the first generation of statistical approaches to articulating cut scores. A pivotal first step to this class of methods is establishing the assumptions about the underlying trajectories of development (e.g., Ferrara et al., 2005; Lewis & Haug, 2005). These assumptions are typically based on historical student performance to derive the expectations for student progress or growth over time. One example of establishing alternative trajectories for developmental expectations was presented by Lewis and Haug, who categorized the assumptions about growth over time as *equal*, *approximately equal, smoothly increasing,* or *decreasing over time*.

Similarly, Ferrara et al. (2005) provided in their simulation study another illustration of applying developmental expectations. Student performance data were generated under various types and amounts of growth in order to determine the accuracy of Grade 2 on-track performance level classifications for predicting Grade 3 proficient performance. The authors compared three types of growth: (a) linear growth: increase in proficiency by a fixed amount for all test takers; (b) remediation: more rapid growth for students placed under intensive remediation when compared to all students; and (c) rich get richer: more rapid increase in proficiency for students who were already above the on-track level in Grade 2. The three types of growth are crossed with four amounts of growth: (a) negative growth: students are placed at a theta metric one *SD* lower in Grade 3 than in Grade 2, and because the Grade 3 proficient cut score was set 1 *SD* lower than Grade 2, this would indicate that the same percentage of students will be classified as proficient or on track in both grades; (b) no growth: all Grade 3 students are set to the same theta metric as in Grade 2; (c) low growth: Grade 3 cut score is set 1 *SD* higher on the theta metric than Grade 2; and (d) moderate growth: Grade 3 cut score is set 2 *SD* higher on the theta metric.

Cizek and Agger (2012) reviewed the developmental trajectory approach in more detail, and therefore just a succinct summary is presented here. They noted that these developmental expectation models are typically employed at the front end as a precursor to cut-score articulation and should typically inform not only the articulation of cut scores but also the development of content standards and the entire test creation process. For instance, a refined application of developmental expectations in the design of content standards is illustrated in the learning progressions framework (Smith et al., 2006) and its realization in the form of the Common Core state standards (CCSSO and NGA, 2010). With or without the implementation of the Common Core, building on the basis of some form of underlying developmental expectations is critical to the successful implementation of cut-score articulation methods. Nevertheless, developmental expectations, as described here, were first employed in the implementation of the first generation of statistical articulation methods, which are broadly classified into statistical interpolation methods and statistical smoothing methods.

Statistical Interpolation of Cut Scores at Intermediate Grades

In this family of methods, cut scores for the end grades (e.g., Grades 3 and 8) are recommended based on panel meetings. Subsequently, statistical interpolations are used, based either on the proportions of students classified or directly based on the scaled cut scores, to determine or adjust the cut scores for intermediate grades (e.g., Grades 4, 5, 6, and 7). For instance, NAEP has historically used a monotonic across-grade trend line in the percentage of students classified in each performance level (Huynh & Schneider, 2005).

As typically applied within a K – 12 standard setting, cut scores for the intermediate grades are developed through statistical interpolations, and no grade-level meetings are conducted at these grades. The original cut scores for the end grades are typically reviewed by a technical advisory committee (TAC; e.g., Huynh et al., 2000). In some cases, the interpolated cut scores may also be reviewed by a metapanel in a standards-validation meeting (e.g., Louisiana Alternate Assessment Program; see Louisiana Department of Education, 2010). This metapanel is composed of panel representatives

from the grade-level meetings for the end grades and possibly a secondary reactor panel (of policy makers and TAC members).

Examples from two states (technical report of the South Carolina 1999 PACT assessments, Huynh et al., 2000; technical report for Louisiana Alternate Assessments, Louisiana Department of Education, 2010) that illustrate how this interpolation approach may be applied are presented here. Huynh et al. applied the statistical interpolation method at the 1999 standard setting for the South Carolina PACT assessments. Louisiana implemented a modification of the statistical interpolation method to recommend cut scores for their state's ELA and mathematics alternative assessments. In both these examples, cut scores were set for only the end grades and then interpolated for the intermediate grades using statistical interpolations. For instance, Huynh et al. reported the cut-score articulation process for multiple achievement levels (ALs) for the 1999 South Carolina PACT assessments based on statistical interpolations. Grades 3, 6, and 8 were considered benchmark years when a student's performance was measured, and that student's progress toward achieving the standards by Grade 12 was determined. However, panel-based standard-setting meetings were conducted for only Grades 3 and 8. The Bookmark (RP67) method (Cizek, Bunch, & Koons, 2004; Karantonis & Sireci, 2006) and three rounds of judgments were used to establish the ALs for the Grades 3 and 8 ELA and mathematics assessments. First, a common policy definition of the achievement categories was developed and used across the grades in order to articulate the cut scores for the intermediate grades. Once the final cut scores for Grades 3 and 8 were adopted by the state of South Carolina (based upon the TAC's recommendations), cut scores for Grades 4 through 7 were interpolated from those for Grades 3 and 8. The common policy descriptors and a simple growth curve trend line were used in the interpolation of the cut scores for the intermediate grades (Huynh et al., 2000).

Similarly, in Louisiana, the cut scores for Grades 4, 8, and 10 ELA and mathematics alternate assessments were established in panel-based meetings using a modified Bookmark procedure in 2006. The following year, statistical interpolations were used to set cut scores for the intermediate grade levels (Grades 5, 6, 7, and 9), and no specific grade-level panel meetings were conducted for the intermediate grades. The interpolations were followed by panel discussions using a reactor panel that reviewed the appropriateness of these cut scores for the knowledge and skills measured by the assessments in these grade levels. Time constraints and federal reporting requirements for all grades in the current year were provided as a rationale for not designing independent meetings for each grade level. After test administration, a standards-validation study was conducted to verify the recommended cut scores against actual student performance data (Louisiana Department of Education, 2010).

In general, although the statistical interpolation method has its intuitive appeal, a large portion of the success in using statistical interpolation relies on meeting certain assumptions about the comparability of the tests and the content standards. Statistical interpolations on scale scores tend to rely heavily on the comparability of the underlying scales and are therefore hard to justify in the absence of an underlying vertical scale. Furthermore, many practitioners (based on self-reports from states and testing contractors for states; see Kannan, 2014) are often hesitant to discuss the face validity of this method due to an absence of grade-level representations at intermediate grades. Therefore, despite being one of the first offered solutions to articulating cut scores (Huynh et al., 2000; Lissitz & Huynh, 2003), statistical interpolations are rarely used independently in practice. If used, they typically tend to be supplemented with results obtained from other cut-score articulation methods (see Table 1).

Smoothing Cut Scores at Intermediate Grades Based on Impact Percentages

Since the initial conception of VMSS by Lissitz and Huynh (2003), which used a combination of professional judgments and statistical interpolations to achieve across-grade consistency, several alternative methods of vertical articulation have been offered, implemented, and evaluated. However, articulation methods that employ some variant of impact-percentage smoothing across the grades remain extremely popular among the states (see Table 1). To the extent that operational test administration data are available for a representative group of test takers, the impact-percentage information would consist of percentages of test takers classified into each of the performance levels (e.g., *basic*, *proficient*) given the recommended cut scores. This family of methods employs a smoothing procedure, used *post hoc*, to minimize the differences in the percentages of students classified across the grade levels, resulting in reasonably comparable percentages of students at each performance category. Impact-percentage smoothing is employed following rounds of discussions in one of two ways (see Table 1). In one variant, a joint (meta) panel composed of panel representatives from grade-level meetings that reviews and adjusts the cut scores recommended for each grade to smooth out the differences in the impact percentages

for the intermediate grades (e.g., Lewis & Haug, 2005). Alternatively, panel meetings at intermediate grade levels are used to develop preliminary cuts, which are then followed by a standards-validation meeting involving a secondary or reactor panel of policy makers or grade-level panel representatives. This panel considers the impact data available to smooth out the differences in the percentages of students classified across the grade levels (e.g., Buckendahl et al., 2005). In practice, states sometimes use these two methods together, in which panel representatives review and smooth cut scores at the end of grade-specific meetings, and a reactor panel is also employed in a standards-validation meeting to consider the grade-level information available and smooth out the differences in percentages of students classified in each performance level across the grades (see Kannan, 2014).

To illustrate an empirical evaluation of the application of the smoothing method, Lewis and Haug (2005) presented the results of an operational implementation of this design in Colorado in which they discussed the processes used to determine a cross-grade alignment model and articulate the cut scores across the grades. In their first attempt at adopting an across-grade alignment model for reading from Grades 4 through 10, they used the reading standards that were established previously for Grades 4 and 7 and used an equipercentile model to make sure that an equal percentage of students were classified as proficient across the grades in reading. It was, however, very challenging for them to justify and communicate this procedure to the media and the public. Therefore, in determining the cross-grade alignment model for writing and mathematics, they decided to integrate a cross-grade alignment model into the judgmental standard-setting process and to use the existing cut scores as a frame of reference for the participants. This cross-grade alignment model was implemented for Grades 3 – 10 (writing) and Grades 5 – 10 (mathematics). Grade-level panels included teachers from the given grade and from the grades immediately above and below the grade level. All grades within a content area worked in the same room at separate tables to enable cross-grade discussions. Previous cut scores in writing (from 2001) established for Grades 4, 7, and 10 were presented as a frame of reference to the writing standard-setting committee in 2002, but panelists were free to make new recommendations within the constraints of the alignment model (an equipercentile model was adopted for writing in consultation with the Colorado Department of Education). Data based on the 2001 administration of the mathematics assessments indicated that the equipercentile model was not appropriate for mathematics based on previously established cut scores for Grades 5, 8, and 10. The data showed a sharply declining percentage of students at or above proficient at higher grades. In response, a committee of mathematics education leaders recommended that a model of smoothly decreasing percentages of students (at or above proficient) be adopted for mathematics. Cross-grade discussions were facilitated after both Rounds 1 and 2 of the process for each performance level. The recommended cut scores resulted in impact data (percentage classified as proficient) within 1% of the final adopted cut scores on average for writing and 3% of the adopted cut scores on average for mathematics. The authors recommended that policies should therefore be adopted prior to standard setting in order to constrain participants to a predetermined alignment model that produces rational across-grade cut scores (Lewis & Haug, 2005).

Overall, the only assumption made in using impact-percentage smoothing is the comparability of the student cohort from grade to grade, and this method makes the least number of assumptions about the underlying scale and the comparability of the underlying cross-grade-level tests. The minimal assumptions made by this method could partially explain the popularity of this family of methods, both in empirical evaluations (e.g., Buckendahl et al., 2005; Ferrara et al., 2007; Lewis & Haug, 2005; Lissitz & Wei, 2008) and operationally, among the 50 states surveyed (see Table 1). However, even though these first-generation methods (both interpolation and smoothing) were originally proposed as an alternative to problems encountered in vertical scaling, a major problem with these initial methods was an inherent dependence on assumptions about trajectories for student development. These assumptions were derived either from historical longitudinal data on student performance (which was difficult to obtain), on vertically scaled assessments (which often have its own array of problems, as reviewed previously), or were determined *ad hoc*. Furthermore, determining a developmental trajectory ahead of time might not always be feasible when historical data are unavailable, such as with the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced assessments, which were field tested only in 2014. Even when they are available, using historical data based on cut scores that are not articulated (and known to be noncomparable across grades) would result in trend data that are not very accurate. Therefore, despite its inherent appeal, the successful demonstration of its application by several researchers (e.g., Buckendahl et al., 2005; Ferrara et al., 2007), and its popularity across the states (see Table 1; Kannan, 2014), this method is not free of limitations. Moreover, the process of deriving the developmental expectations based on historical performance as a step in the vertical articulation

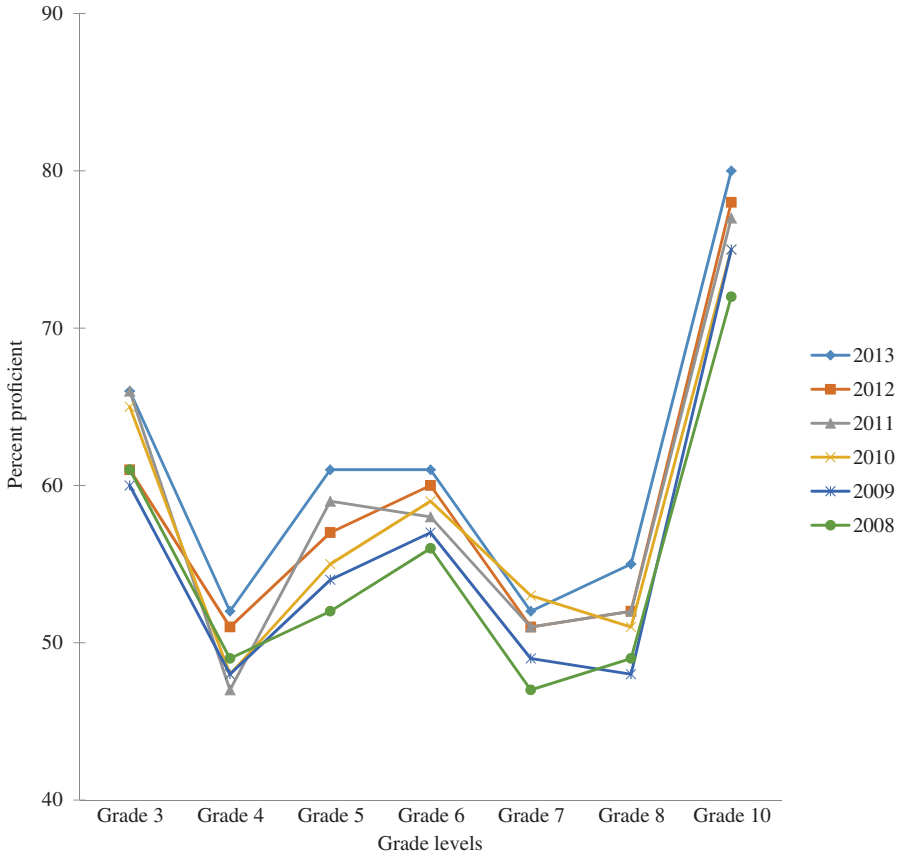

Figure 1 Percentages of students scoring *proficient* or higher in the Massachusetts Comprehensive Assessment System (MCAS) statewide assessments across the years. Data obtained from Table E-1, showing ELA, mathematics and STE results at all grade levels for each test since its inception, in the *Spring 2013 MCAS Tests: Summary of State Results* (Massachusetts Department of Elementary & Secondary Education, 2013).

process is considered to be time consuming and arduous. This step never appears to have picked up momentum among the states and is not widely practiced.

For the most part, states reported that they did not make any assumptions about the underlying developmental expectations in making the statistical adjustments in articulating the cut scores across the grades (see Kannan, 2014). Instead, the application of this family of methods generally relied on smoothing out differences in classifications across the grades without explicitly determining the extent to which the standard-setting methodology applied to the base year in their models resulted in accurate classifications. Furthermore, the adjustments made to cut scores based on the impact percentages were typically within one or two standard errors of the original panel-recommended cut scores. This practice is recommended (Foley, 2014) because it is important not to disregard the original panel-based recommendations. However, due to this practice, sometimes the fluctuating patterns established based on the grade-level meetings are not completely ironed out. Historical percentage-proficient data retrieved from the Massachusetts report of Spring 2013 results for the statewide assessment (Massachusetts Department of Elementary & Secondary Education, 2013) help illustrate this point (see Figure 1). It can be seen that, despite the state's attempts at smoothing the proportions of students classified as proficient across the grades, the fluctuating patterns established based on grade-level meetings have not been completely alleviated. The percentages of proficient students in Grades 3 and 10 remained somewhat higher than the other grade levels, and this pattern appears to have persisted across several years of test administration.

Student Growth Percentiles to Articulate Cut Scores at Intermediate Grades

Growth reporting at the individual student level, federally mandated in the context of NCLB, resulted in a flurry of state responses to the requirement (see report on state growth models by Blank, 2010). However, measuring growth has been

challenging for most states. As discussed earlier, and due to some of the inherent complications, vertical scale-based models are far from the ideal solution for growth reporting. VMSS-based approaches offered as an alternative for AYP reporting requirements helped resolve that problem but are not ideal for measuring growth. Value-added models (or VAMs) were offered as yet another solution to growth reporting. With statistical packages such as the Education Value-Added Assessment System (EVAAS) made available by the SAS Institute, this family of solutions soon became popular. These sophisticated statistical techniques brought with them the possibility (or danger) of being able to justifiably (or unjustifiably) isolate and quantify the effects of a single teacher (and possibly school) on a student's achievement and growth (Braun, 2005; Raudenbush, 2004). Despite their intuitive value, these models are susceptible to a number of biases (AERA Council, 2015; McCaffrey, Lockwood, Koretz, & Hamilton, 2004) and make inaccurate causal attributions of student achievement to a single teacher or school. Such attributions can only be reasonably made when all relevant student, school, and community factors such as ethnicity, parental education, income, and other characteristics, and also school stability, location, and so forth, have been controlled for (Lockwood et al., 2007; McCaffrey et al., 2004). However, the numerous contributing factors cannot reasonably be identified, let alone controlled for by any existing model. Therefore, the validity of the conclusions drawn from such models becomes questionable (Lockwood et al., 2007; Martineau, 2006; McCaffrey et al., 2004), particularly for students from low-status schools who demonstrate insufficient growth (Dunn & Allen, 2009). Consequently, the utility of these models is currently controversial at best.

Student growth percentiles (SGPs; Betebenner, 2008, 2012) were proposed as a solution to measuring individual student growth. The SGP model does not make the same causal attribution error as its predecessor (i.e., VAM), yet at the same time provides a description of individual growth, which provides stakeholders the opportunity to understand an individual student's growth (or progress toward the desired criterion performance level) without the avoidable attribution of the responsibility for that growth to another entity (e.g., teacher, school; Betebenner, 2009). Betebenner conceived of SGPs as a unifying framework that bridges the gap between a norm-referenced growth analysis technique, the criterion-referenced standards, and the accountability system (Betebenner, 2012; Betebenner & Linn, 2009). SGPs are similar to growth percentiles used by pediatricians and are used to normatively compare a student's level of achievement to that student's academic peers based on prior scores. An SGP describes a student's achievement *change* from one year to the next compared to other students with similar prior test scores (the student's academic peers). A student's current level of achievement is compared to that student's previous level of achievement in order normatively determine the rate of achievement growth. The resultant percentile reflects the likelihood of a student achieving a certain outcome, given the student's prior achievement. The relationship between prior and current achievement scores for cohorts of students in the norm group can be used to generate growth trajectories based on historical and anticipated rates of growth to predict the likelihood of future achievement for students statewide (Betebenner, 2008, 2009) and may thereby enable assumptions regarding growth over time.

The SGP approach acknowledges that the rate of growth necessary to reach or maintain proficiency can be different for each student and will be based on the student's current level of achievement and predicted or projected future rate of growth. Typically, each student's most recent scores are conditioned against that student's prior scores to quantify the adequacy of growth. Student growth is quantified in a norm-referenced fashion to determine the rate of growth necessary to achieve target ALs (such as proficient) by each student. Furthermore, the likelihood of such an event occurring can be calculated based on projected growth trajectories for each student (Betebenner, 2008, 2009). However, criteria that quantify the adequacy of this growth for individual students would also be necessary. Betebenner, Shang, Xiang, Zhao, and Yue (2008) suggested that stakeholders may use a standard-setting procedure to compare the levels of growth to the state cut scores so that the growth in any given year may be categorized as inadequate, satisfactory, or good. Based on the NCLB-mandated achievement outcomes of universal proficiency, standards of growth necessary to reach prescribed levels of achievement for certain students might be unreasonable. Betebenner (2009, 2012) recommended that achievement mandates that are currently stipulated, based on the moral imperative of high standards for all and without heed to the likelihood of achieving this by individual students, should definitely be revisited.

Betebenner's (2009) original growth and achievement chart is presented here to illustrate this point further (see Figure 2). The white lines in this figure depict the norm-referenced achievement progression across grades, and the shaded background regions depict the criterion-referenced achievement progression across grades, that is, based on students classified at the various performance levels (let us use the *below basic*, *basic*, *proficient*, and *advanced* performance-level classifications here for illustration). In addition, five normative percentile growth trajectories (at

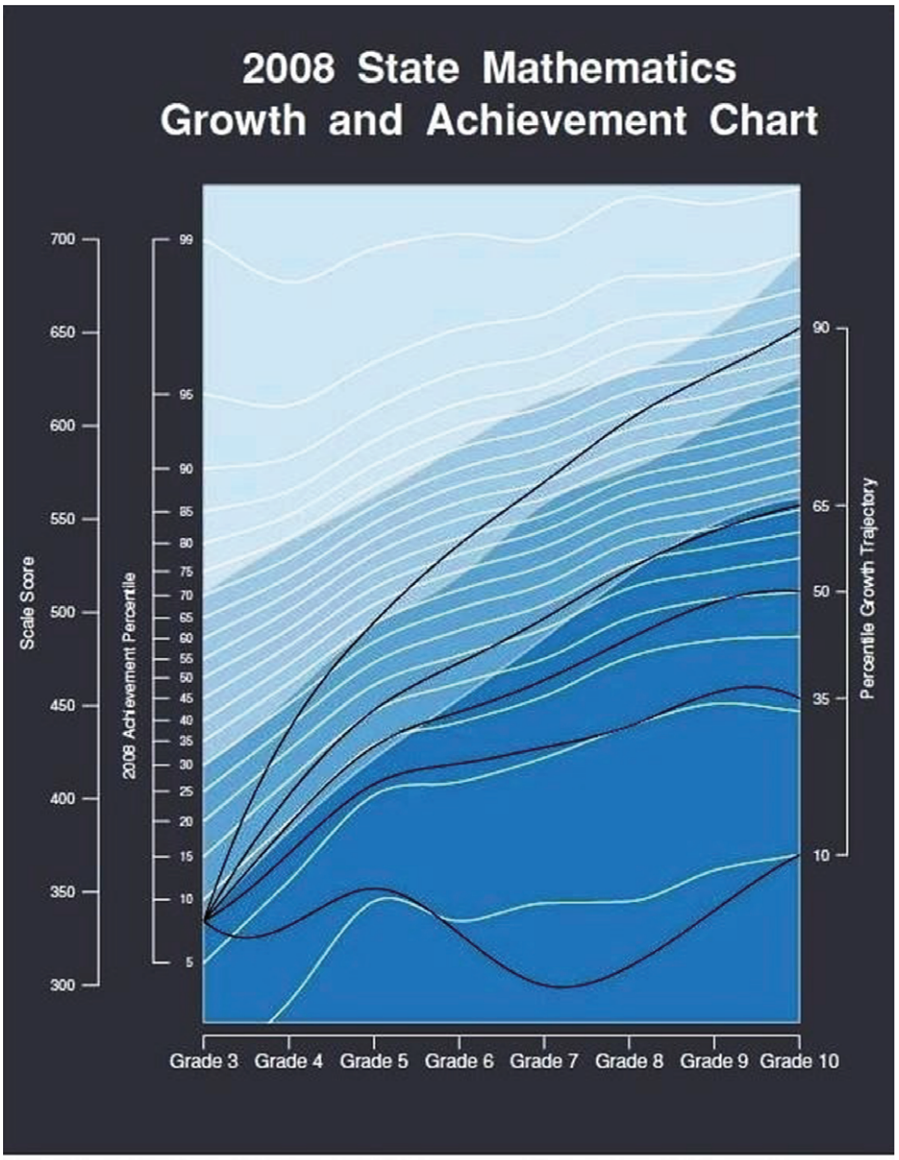

Figure 2 Growth and achievement chart showing various percentile growth trajectories. From D. W. Betebenner, 2009. Reproduced with permission of John Wiley and Sons.

the 10th, 35th, 50th, 65th, and 90th percentiles) are superimposed (dark lines). This reflects various normative growth trajectories for a student who is at the cut score between performance levels 1 and 2 (let us assume that this student is "just barely basic"). This figure is very effective in showing the unreasonable growth expectations for various students under the NCLB-mandated universal proficiency requirements. Notice how this hypothetical student, who is just barely basic at the third grade, has to demonstrate a 90th percentile growth in order to be classified as proficient by the time that student is in high school. Betebenner (2009) pointed out that when such a high rate of growth is required to reach an achievement target, the student's chances of reaching that outcome become proportionally very small (e.g., if a 90th percentile achievement is required, only 10% of students having the same required percentile growth as that student would be expected to reach that achievement target).

Finally, SGPs may also be used to articulate the cut scores for the intermediate grades. Even though the SGP approach has become the most popular growth model currently used by states for AYP reporting, this model has not been used

to articulate cut scores across the grades by any of the 35 states that responded to the survey (see Kannan, 2014). Nevertheless, these models provide the necessary data, based on historical growth for students in a state, that might help make assumptions about the trajectories of student development in articulating cut scores across the grades. Such historical data would be helpful to make the required assumptions about articulation trajectories employed by other methods.

Articulation in the Context of the Next-Generation K–12 Assessments

Some early proponents of vertical articulation methods (e.g., Bejar et al., 2007; Huynh, Barton, Meyer, Porchea, & Gallant, 2005; Lewis & Haug, 2005) have recommended a holistic approach from the beginning. They argued that the following steps in test development should be considered from a holistic perspective, and when attention is paid to each step, the steps logically support articulation of cut scores. First, the content standards should be articulated. Next, the curriculum and assessments should be based on these articulated and logically progressing content standards. Next, students at each performance level should be classified based on the expectations of the knowledge and skills (K/S) required at the given grade and an expectation of growth from the K/S required at the previous grade. Finally, from a holistic viewpoint, the cut scores across the grade levels should also be logically articulated based on these underlying expectations. Such a holistic approach is exactly where the next-generation assessments are headed.

The learning progressions framework (Smith et al., 2006) offered a solution to a need for articulating content standards and has been influencing the development of curriculum and instruction based on scaffolding of content in a developmentally appropriate sequence. Much effort has gone into the creation of lesson plans, instruction, development of formative assessments, and content standards based on learning progressions (Alonzo, 2010, 2011; Black, Wilson, & Yao, 2011; Corcoran, Mosher, & Rogat, 2009; Heritage, 2008, 2011). The Common Core State Standards initiative (see CCSSO and NGA, 2010), launched in 2009, provided a consistent and clear understanding of what students are expected to learn at the forefront and therefore offers a holistic approach to curriculum development and assessment design. These core standards have driven states to rethink methods across the entire spectrum of the broader assessment system (Forgione, 2012), which includes methods used to recommend cut scores. Innovative standard-setting solutions such as benchmarking were offered. These novel solutions were geared to the holistic assessment system and were intended to span the grade levels and content areas (O'Malley et al., 2012).

The assessment consortia, namely PARCC and Smarter Balanced Assessment Consortium (Smarter Balanced), have developed assessments with the goal of identifying students who are on track to being college and/or career ready (CCR) at each grade level. With the push for ascertaining whether students are on track to being CCR with these new standards, the need to articulate standards across grade levels with equal rigor becomes all the more important. Cut-score articulation therefore remains central to the consortia assessments. For instance, PARCC, as outlined in its *Request for Proposal* (PARCC, 2013), explicitly addressed the importance of being able to predict whether students at lower grades are on track to meet evidence-based benchmarks for college and career readiness and to meet established international benchmarks, and it called for external validation studies to establish these benchmarks. PARCC defined students who are CCR in ELA and mathematics as those who would have a 0.75 probability or higher of obtaining at least a *C* in mapped entry-level college courses, for example, college English composition for ELA or college algebra for mathematics. PARCC was particularly requesting studies that established the benchmarks as defined above using various external criteria (e.g., NAEP scores, American College Testing (ACT) scores, Scholastic Aptitude Test (SAT) scores, ASVAB scores, scores on international assessments, and scores for matched students in at least three credit-bearing higher education institutions each from at least 10 different PARCC states).

In general, to establish the CCR cut scores for the high school assessments, research has focused on methods that synthesize the benefits of linking K–12 educational data with external benchmarks using various national (e.g., NAEP) and international (e.g., TIMMS, PIRLS, and PISA) comparisons to broaden the definition of grade-level performance. Furthermore, there is a trend toward collecting longitudinal student data to evaluate whether students are on track to being successful at the next and subsequent grade levels (O'Malley et al., 2012). In response, several novel solutions to standard setting in general, and to articulating cut scores across the grades in particular, have been advanced. For example, establishing benchmarks (Phillips, 2010, 2012) geared specifically toward identifying students who are CCR (Miles, Beimers, & Way, 2010; O'Malley et al., 2012) at the end of high school and articulating cut scores at the lower grades to predict on-track performance to being college and career ready at higher grades are some of the more recent solutions being offered

to vertically articulate cut scores. Although promising, these methods also raise some additional methodological concerns that must be reviewed carefully. Some of these methods and the methodological challenges they pose are reviewed next.

Benchmarking and Predictive Standard Setting

Ever since Haertel's (2002) article introduced the idea of validating established cut scores in a participatory standard-setting process, the idea of including data from multiple sources of evidence (i.e., benchmarks) in validating the cut scores has become very popular. This idea culminated in several efforts that suggested comparisons to external reference points using criterion-referenced validation studies (e.g., Haertel, Beimers, & Miles, 2012; McClarty et al., 2013) or benchmarking (Phillips, 2010, 2012). In general, data from multiple sources, such as linking studies with other related national or international tests (e.g., NAEP, TIMSS, and PISA) and correlational studies with performance in college courses, are provided to panelists as preliminary values for recommending the high school or end-of-course (EOC) cut scores. Panelists then use the collective evidence in evaluating the reasonableness of various alternative cut scores and make a final recommendation (e.g., Haertel et al., 2012; Miles et al., 2010). However, Cizek and Agger (2012) pointed out that the use of external criteria for establishing valid cut scores rests on two very strong assumptions. One is that the external criteria (tests) used are directly relevant to the purpose of the test and are aligned to the content standard measured by the test. Second, if these external criteria are introduced at various grade levels (e.g., NAEP at the fourth, eighth, and 12th grades) for articulation of cut scores across grades, then they are themselves based on relevant and well-articulated content standards and performance expectations. Cizek and Agger recommended that information from external references should be used as a moderating influence rather than as the sole influencing criterion for establishing standards across the grades. In the end, panels are still tasked with interpreting the empirical data to inform decisions (judgments) about where to locate the cut scores.

Logistic Regressions to Articulate Cut Scores

Once the CCR standards are established for the high school assessments, these standards would then have to be articulated such that consistent on-track decisions are made across the grade levels and students who are not on track can be identified and targeted for remediation. One novel suggestion that has been offered to articulate on-track cut scores is the use of a logistic regression procedure (e.g., as applied by the Michigan Department of Education, 2011) to establish the cut scores at the lower grade levels. Logistic regression-based approaches were offered as a solution to back-translating the benchmarked EOC cut scores to lower grades, first suggested by Miles et al. (2010) and implemented by the Michigan Department of Education in the context of the Michigan Educational Assessment Program (MEAP). The process of using logistic regressions in the vertical articulation of cut scores is not clear. One solution that has been tried (e.g., by the Michigan Department of Education) is to perform a series of logistic regressions using the standards set for the higher grade as the criterion for establishing cut scores for the lower grades.

As an illustrative example, the study conducted by the Michigan Department of Education (2011), where the logistic regression procedure was used to vertically articulate cut scores, is briefly described here. In Michigan, the CCR cut scores were first established for the EOC Michigan Merit Examinations (MMEs) at the 11th grade by relating the 11th-grade scores to the course grades from first-year college students enrolled in Michigan public postsecondary institutions. *Proficient* was defined as having a 50% chance of obtaining a *B* or higher in selected freshman college courses. *Partially proficient* and *advanced* were defined as having 33% and 67% chances, respectively. Separate analyses were run for 2-year and 4-year institutions, but the results were within one standard error of each other and therefore combined. Subsequently, a series of logistic regression analyses were used to vertically articulate the CCR cut scores to determine on-track performance in one or more of the lower grades for MEAP. They also used a signal detection theory (SDT) – based method and equipercentile cohort matching for articulating the cut scores. The results obtained from these additional methods were validated against those obtained using the logistic regressions.

In their application, Michigan obtained matched datasets for cohorts of students ranging across 6 years such that datasets were matched for some students from Grades 3 through 8 and for others from Grades 8 through 12 (see Table 6, Michigan Department of Education, 2011). However, they included at least two cohorts for each back-mapping (i.e., the regressions that link each lower grade cut score to the higher grade cut scores). In addition, to minimize the number of links and the resultant cumulative measurement error, the number of links per grade level was minimized such that not

more than three links were made to establish the cut scores at each grade level. For instance, to establish the Grade 3 cut scores, a link was made from Grade 3 to Grade 7 (where the cut scores for Grade 7 had already been linked to the EOC). They evaluated the classification consistency rates from year to year and found the lowest classification consistency for MME to college grades and the remaining consistency rates showed a high degree of stability from grade to grade. They reasoned that the smaller classification consistency from EOC to college grades was due to the largest construct shift for this group. This basically indicated that the criterion that was being predicted (college performance) was not well correlated with the predictor (EOC or high school performance), and these results bring into question the validity of using college performance as a criterion in predicting the CCR standards at the high school level.

Moreover, such a conclusion somehow does not reconcile with findings from the predictive validity literature that indicated that high school grades are the best predictors of college grades (e.g., Atkinson & Geiser, 2009; Soares, 2012). One possible explanation for this discrepancy could be that high school performance-level classifications (as opposed to high school grades) are perhaps not very good predictors of college grades. But in fact, a deeper analysis to understand the squared correlations reported across these predictive validity studies helps resolve these contradictory conclusions. For example, in a large-scale study spanning several student cohorts at the University of California, Geiser and Studley (2004) found that the squared correlation between high school GPA (HGPA) and freshman-year college GPA (FGPA) was .15. When compared to a squared correlation of .13 for SAT scores and FGPA, the conclusion that HGPA as the best predictor is perhaps mathematically accurate, but evaluating predictors in terms of rank-ordered correlations can be quite misleading. Even though HGPA was the best predictor of FGPA in the above example, it still accounted for only 15% of the variance in college grades. In addition, the small magnitude of this relationship is particularly concerning in predicting the on-track cut scores at lower grades. The back-mapped cut scores at lower grades, regardless of their correlations with each other, are likely to be even less correlated with the final criterion (i.e., freshman-year college performance). Therefore, it might be advisable to use freshman college performance with caution (and perhaps in combination with other criteria) in predicting college-ready standards at the end of high school. One solution might be to use multiple external benchmarks; for example, it might be prudent to combine information from multiple criteria such as links to performance on freshman college courses in association with links to performance on nationally and internationally benchmarked assessments (e.g., NAEP, TIMSS, and PISA) and perhaps links to performance on related tests that comprehensively measure knowledge and skills attained at the end of high school (e.g., SAT and ACT). Such an approach might somewhat alleviate the negative consequences of the small correlation between the predictor and a single criterion.

A Paradigmatic Solution or an Inadequate Solution?

Despite the cautionary steps taken by the Michigan Department of Education, a caveat for predictive standard setting was pointed out by Ho (2013). In predictive standard setting, empirically defensible predictive statements are attached to the performance levels classified on a score scale. Ho argued that an appropriate distinction needs to be made between a cut score that corresponds to future performance and a cut score that predicts future performance—in not doing so, interpretations of stringency in classifying students become intermingled with the predictive utility of the test. The standards established for on-track performance can support the necessary inferences only to the extent that the test can predict the outcome. In other words, to the extent that the criterion used is not highly correlated with the predictor (in this case, the grade-level assessments used to make the linkages), the resultant predictive relationship, though mathematically valid, might lead to incorrect interpretations of the predictive relationship and, therefore, misuse.

Ho (2013) compared the predictive utility of the logistic regression method to a baseline equipercentile method (which assumes equal percentages of students will be classified as on track in subsequent grades) and showed that the predictive standards are biased toward the future standards, more so than even a simple equipercentile method. Furthermore, when the correlation is small, such as the correlation of assessments at lower grade levels to EOC assessments, Ho pointed out that the stringency or leniency of the cut scores depends on the future cut score. Owing to regression toward the mean, if the future cut score (EOC – CCR cut) is above the mean, then the predictive standard for the lower grades will be more stringent, and if the future (CCR) cut is below the mean, then the predictive standard at lower grades will be more lenient, to the same degree.

In addition, though predictive regression-based methods have been used frequently in predicting the relationship between HGPA and FGPA (e.g., Agronow & Studley, 2007; Geiser & Santelices, 2007), there is the risk of cumulative measurement error (Zwick, 2013), particularly while making multiple linkages. As an underlying property of any regression

method, the residual or prediction error (i.e., the difference between the predicted value and observed value) will average zero across all individuals in the analysis. However, Zwick pointed out that even when there is a strong association between the predictor and the predicted quantity, there could be systematic prediction errors for some student groups, and this is often overlooked in establishing predictive relationships. Furthermore, in the event that the correlation between the predictor and the predicted value becomes smaller (as would be the case between lower grades and future cut scores), the departure from the model-predicted value and actual observed value may be systematically deviant for those lower grades and result in unduly large prediction errors.

No matter the nature of the underlying content standards and the logical progressions used to develop the assessments, the correlation between high school and Grade 3 assessments is likely to be small. It is therefore critical that the impact of using predictive methods be evaluated with simulation studies when the correlation between a predictor and criterion is fairly small. These simulation studies should evaluate the consistency and accuracy with which students are classified when correlations are small and multiple linkages are made. Ho (2013) claimed that using just a simple equipercentile approach for articulating the cut scores might be better than predictive methods. However, the assumption that the equipercentile standards are a baseline is questionable, and the decision consistency in using equipercentile or predictive standards needs to be evaluated, perhaps using simulations. In addition, the relationship between grade-level assessments needs to be ascertained, and the growth trajectory used to operationalize the predictive relationship (equipercentile across grades, increasing percentage, decreasing percentage, etc.) needs to be determined. Furthermore, the validity of established predictive relationships needs to be evaluated for multiple disaggregated student subgroups (see Zwick, 2013, for a review of differential impacts of systematic errors for various subgroups in determining predictive relationships). Finally, the results from preliminary standard-setting meetings should be validated using longitudinal data to evaluate the results suggested from the simulation studies. For example, it would be prudent to determine, using interim/formative assessments, whether students who are classified as on track or ready for Grade 4 are successfully learning content in Grade 4. If these data call into doubt the reasonableness of the standards so established, then the entire process should be revisited so as to avoid negative consequences.

A Cautionary Note

Despite the excitement about the potential of these novel standard-setting methods, one significant uncertainty remains: What is really meant by "college and career ready?" Attempts at defining what it means to be college and career ready (e.g., Camara, 2013; Camara, Wise, Plake, Kolen, & Conley, 2013) have been made, but it is still not clear how to (or whether we should) make the finer distinctions between what it means to be ready for a 4-year institution, ready for a 2-year institution, or ready for professional training in a vocational school. Even though it is apparent that these are not the same thing, for now states and consortia appear to refer to a single standard based on a criterion measure of obtaining a *C* or higher in freshman college courses and not needing remediation as part of the definition of being CCR. However, it is not likely that the same level of performance would be indicative of readiness for each of these end goals. Therefore, the criterion for each type of readiness might have to be more distinctively defined.

Is There Really a Seamless Solution?

Overall, any answer to seamlessly recommending articulated cut scores across the grades is still somewhat arbitrary. Several solutions have been offered in the past decade, but as this review shows none of these methods is infallible or completely free of limitations. Even though predictive standard setting and logistic regression-based approaches to determining articulated grade-level cut scores appear to have a lot of intuitive appeal, one must proceed with caution in classifying students as on track or not on track based on these predictive relationships alone. Predictive relationships established using these methods should be grounded by gathering substantiating evidence based on longitudinal student data and by ascertaining the classification consistency and accuracy of cut scores established using these methods. Using multiple criteria, as illustrated earlier in this paper, might help alleviate some of the concerns of overdependence on a single, less than perfect, relationship between a predictor and a criterion in establishing predictive relationships. Nevertheless, an overreliance on predictive relationships might inevitably place undue emphasis on the predictive utility of the test and therefore lead to negative, unintended consequences.

One possible resolution for seamless recommendation of articulated cut scores would be to use a combination of methods, both data based and judgment based, that complement each other and corroborate each other's results. For example, information from multiple criteria (e.g., performance on the next higher grade, performance on nationally benchmarked assessments at grades where they are available) may be used along with logistic regression-based predictive methods to predict the initial ranges for the articulated cut scores at each grade level. These values could then be used in panel-based meetings at each grade level that would employ a judgmental standard-setting method (e.g., Angoff or Bookmark). At these meetings, panelists could consider the suggested ranges of predicted cut scores as start values but also use their own expertise (as in any judgmental method) to make recommendations about on-track cut scores at lower grades. Subsequently, cut scores recommended at these panel-based meetings may be smoothed using a *post-hoc* statistical approach; when available, it might also be useful to ascertain an appropriate developmental trajectory based on information from historical student classification data.

An illustration of such a balanced approach is provided by the method proposed by one multistate consortium (PARCC, 2013) in determining the EOC – CCR standards. PARCC outlined a possible combination of solutions that might help minimize the potential for systematic errors and biases. The authors highlighted the importance of conducting benchmarking studies using various external criteria prior to the standard-setting meeting to identify initial ranges for the CCR cut score; these cuts could be offered to a panel of judges who would be using a judgmental standard-setting method. In addition, they recommended that performance of college freshmen on the PARCC assessments and an a priori judgmental study using college professors should also inform the ranges of initial CCR cut scores that are offered to the standard-setting panel. Finally, they highlighted the importance of conducting longitudinal external validation studies to evaluate the classification consistency of the recommended CCR cut scores.

The approach offered here for cut-score articulation across the grades follows the same reasoning in its effort to alleviate systematic bias. Nevertheless, from a practical standpoint, it might not be reasonable to use all possible sources of evidence in recommending articulated cut scores. Furthermore, it is unclear whether using a combination of approaches would in fact enhance the achieved outcome or would, instead, provide multiple conflicting directions and compound the shortcomings of each approach. It would be prudent to use simulation studies (reported item parameter estimates from state assessments may be used as start values in such simulations) to evaluate the decision consistency in combining different combinations of methods (e.g., predictions using external criteria combined with smoothing). Such a simulation study should evaluate the degree to which a *post-hoc* smoothing may help alleviate misclassifications based on predictions. Moreover, different trajectories should be employed and corroborated with historical data to determine the appropriate smoothing technique for different grades. Such an evaluation should not directly impact student classifications in a high-stakes testing context, and using simulations would help ensure that, when operationally applied, the proposed solution would result in minimal systematic errors and, to the extent possible, a seamless articulation of cut scores across the grades.

References

Agronow, S., & Studley, R. (2007, November). *Prediction of college GPA from new SAT test scores: A first look*. Paper presented at the annual meeting of the California Association for Institutional Research. Retrieved from http://www.cair.org/wp-content/uploads/sites/474/2015/07/Agronow.pdf

Alonzo, A. C. (2010). Considerations in using learning progressions to inform achievement level descriptions. *Measurement: Interdisciplinary Research and Perspectives*, *8*(4), 204 – 208. doi:10.1080/15366367.2010.526006

Alonzo, A. C. (2011). Learning progressions that support formative assessment practices. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 124 – 129. doi:10.1080/15366367.2011.599629

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665 – 676. doi:10.3102/0013189X09351981

Bandeira de Mello, V. (2011). *Mapping state proficiency standards onto NAEP scales: Variation and change in state standards for reading and mathematics, 2005 – 2009* (NCES 2011 – 458). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC: Government Printing Office.

Barton, P. E. (2009). *National education standards: Getting beneath the surface* (Policy Information Center Report). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/PICNATEDSTAND.pdf

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1–30). Maple Grove, MN: JAM Press.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York, NY: Routledge.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Betebenner, D. W. (2012). Growth, standards, and accountability. In G. J. Cizek (Ed.), *Setting performance standards: Foundation, methods and innovations* (pp. 439–450). New York, NY: Routledge.

Betebenner, D. W., & Linn, R. L. (2009, December). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Paper presented at the Center for K–12 Assessment & Performance Management, Exploratory Seminar: Measurement Challenges within the Race to the Top Agenda. Princeton, NJ

Betebenner, D. W., Shang, Y., Xiang, Y., Zhao, Y., & Yue, X. (2008). The impact of performance level misclassification on the accuracy and precision of percent at performance level measures. *Journal of Educational Measurement*, *45*(2), 119–137.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 71–123. doi:10.1080/15366367.2011.591654

Blank, R. K. (2010, June). *State growth models for school accountability: Progress on development and reporting measures of student growth*. Retrieved from Council of Chief State School Officers website: http://www.ccsso.org/Documents/2010/State_Growth_Models_2010.pdf

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models (Policy Information Center Report)*. Princeton, NJ: Educational Testing Service.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204–226.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3–14.

Buckendahl, C. W., Huynh, H., Siskind, T., & Saunders, J. (2005). A case study of vertically moderated standard setting for a state science assessment program. *Applied Measurement in Education*, *18*, 83–98.

Camara, W. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practice*, *32*(4), 16–27.

Camara, W., Wise, L., Plake, B., Kolen, M., & Conley, D. (2013, April). *"College and career ready": Incompatible buzzwords*. Invited debate of the day at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*(4), 305–331.

Center on Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington, DC: Author. Retrieved from http://www.cep-dc.org/displayDocument.cfm?DocumentID=179

Cizek, G. J. (2005). Adapting testing technology to serve accountability aims: The case of vertically moderated standard setting. *Applied Measurement in Education*, *18*, 1–9.

Cizek, G. J., & Agger, C., A. (2012). Vertically moderated standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundation, methods and innovation* (pp. 467–484). New York, NY: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*, 31–50.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report # RR-63). Philadelphia, PA: Consortium for Policy Research in Education. Retrieved from http://www.cpre.org/learning-progressions-science-evidence-based-approach-reform.

AERA Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, *44*(8), 448–452. doi:10.3102/0013189X15618385

Council of Chief State School Officers and the National Governors Association Center for Best Practices. (2010). *Common Core state standards*. Washington, DC: Author. Retrieved from http://www.corestandards.org

Dadey, N., & Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment, Research & Evaluation*, *17*(14). Retrieved from http://pareonline.net/getvn.asp?v=17&n=14

Dunn, J. L., & Allen, J. (2009). Holding schools accountable for the growth of non-proficient students: Coordinating measurement and accountability. *Educational Measurement: Issues and Practice*, *28*(4), 27–41.

Egan, K. L., Ferrara, S., Schneider, M. C., & Barton, K. E. (2009). Writing performance level descriptors and setting performance standards for assessments of modified achievement standards: The role of innovation and importance of following conventional practice. *Peabody Journal of Education*, *84*(4), 552–577. doi:10.1080/01619560903241028

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundation, methods, and innovations* (pp. 79–106). New York, NY: Routledge.

Ferrara, S., Johnson, E., & Chen, W.-H. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education*, *18*(1), 35–59.

Ferrara, S., Phillips, G. W., Williams, P. L., Leinwand, S., Mahoney, S., & Ahadi, S. (2007). Vertically articulated performance standards: An exploratory study of inferences about achievement and growth. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 31–63). Maple Grove, MN: JAM Press.

Foley, B. P. (2014, April). *Evaluating an impact percentage smoothing vertically moderated standard setting design.* Paper presented at the meeting of the National Council on Measurement in Education, Philadelphia, PA.

Forgione, P., Jr. (2012). *Coming together to raise achievement: New assessment for the Common Core State Standards*. Paper presented at the Center for K–12 Assessment & Performance Management, Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, Princeton, NJ.

Geiser, S., & Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes* (Research and Occasional Paper Series: Report No. CSHE 6.07). Berkeley: University of California Center for Studies in Higher Education. Retrieved from http://www.cshe.berkeley.edu/sites/default/files/shared/publications/docs/ROPS.GEISER._SAT_6.13.07.pdf

Geiser, S., & Studley, R. E. (2004). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 125–153). New York, NY: Routledge.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 16–22.

Haertel, E. H., Beimers, J., & Miles, J. (2012). The briefing book method. In G. J. Cizek (Ed.), *Setting performance standards: Foundation, methods and innovations* (pp. 283–300). New York, NY: Routledge.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–252). New York, NY: Springer.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: The Council of Chief State School Officers. Retrieved from http://www.cse.ucla.edu/products/misc/cse_heritage_learning.pdf

Heritage, M. (2011). Commentary on road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*(2–3), 149–151. doi:10.1080/15366367.2011.599647

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*(6), 351–360. Retrieved from http://edr.sagepub.com/content/37/6/351

Ho, A. D. (2013, April). *Off track: Problems with "on track" inferences in empirical and predictive standard setting.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Ho, A. D., Lewis, D. M., & MacGregor Farris, J. L. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice*, *28*(4), 15–26.

Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., Reid-Brossard, P., Ho, A. D., & Furgol, K. E. (2011). *Final report on the evaluation of the Growth Model Pilot Project*. Washington, DC: U.S. Department of Education.

Huff, K., & Plake, B. S. (2010). Innovations in setting performance standards for K–12 test-based accountability. *Measurement: Interdisciplinary Research and Perspectives*, *8*, 130–144. doi:10.1080/15366367.2010.508690

Huynh, H., Barton, K. E., Meyer, J. P., Porchea, S., & Gallant, D. (2005). Consistency and predictive nature of vertically moderated standards for South Carolina's 1999 Palmetto Achievement Challenge Tests of language arts and mathematics. *Applied Measurement in Education*, *18*, 115–128.

Huynh, H., Meyer, J. P., & Barton, K. E. (2000). *Technical documentation for the 1999 Palmetto Achievement Challenge Tests of English language arts and mathematics, grades three through eight*. Columbia, SC: South Carolina Department of Education, Office of Assessment.

Huynh, H., & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education*, *18*, 99–113.

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education*, *21*(3),187–206. doi:10.1080/08957340802161741

Kannan, P. (2014). *Content and performance standard articulation practices across the states: Report summarizing the results from a survey of the state departments of education* (ETS Research Memorandum No. RM-14-09). Princeton, NJ: Educational Testing Service.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4–12. doi:10.1111/j.1745-3992.2006.00047.x

Kenyon, D. M., MacGregor, D., Li, D., & Cook, H. G. (2011). Issues in vertical scaling of a K–12 English language proficiency test. *Language Testing*, *28*(3), 383–400. doi:10.1177/0265532211404190

Kolen, M. J. (2011). *Issues associated with vertical scales for PARCC assessments* [White paper]. Retrieved from http://www.parcconline.org/files/40/TechnicalAdvisoryCommittee/43/Vertical-Scales-Kolen.pdf.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd). Secaucus, NJ: Springer-Verlag.

Koretz, D. (2008). Further steps toward the development of an accountability-oriented science of measurement. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 71–92). New York, NY: Routledge.

Koretz, D. (2010). *Implications of current policy for educational measurement*. Paper presented at the Center for K–12 Assessment & Performance Management, Exploratory Seminar: Measurement Challenges within the Race to the Top Agenda, Princeton, NJ.

Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education*, *18*, 11–34.

Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan, & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3–24). New York, NY: Routledge.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2005). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *33*, 3–16.

Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, *8*(10). [Electronic journal]. Retrieved from http://PAREonline.net/getvn.asp?v=8&n=10

Lissitz, R. W., & Wei, H. (2008). Consistency of standard setting in an augmented testing system. *Educational Measurement: Issues and Practice*, *27*(2), 46–56.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Louisiana Department of Education. (2010). *LEAP Alternate Assessments, Level 2, 2009–2010 annual report*. Baton Rouge, LA: Author. Retrieved from http://www.louisianabelieves.com/resources/library/assessment

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*, 35–62.

Massachusetts Department of Elementary & Secondary Education. (2013). *Spring 2013 MCAS tests: Summary of state results*. Table E-1: 1998–2013 Statewide MCAS Test Results Percentage of Students Scoring Proficient or Higher (pp. 3–5). Malden, MA: Author. Retrieved from http://www.doe.mass.edu/mcas/results.html?yr=2013

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/pubs/monographs/MG158.html

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher*, *42*(2), 78–88. doi:10.3102/0013189X12470855

Michigan Department of Education. (2011). *Establishing MME and MEAP cut scores consistent with college and career readiness: A study conducted by the Michigan Department of Education (MDE) and ACT, Inc. Appendix E: New Developed Cut Scores*. Lansing, MI: Author. Retrieved from https://www.michigan.gov/documents/mde/Appendix_E_-_New_developed_Cut_Scores_451846_7.pdf

Miles, J. A., Beimers, J. N., & Way, W. D. (2010, April). *The modified briefing book standard setting method: Using validity data as a basic for setting cut scores.* Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives*, *14*(1), 1–172.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

O'Malley, K., Keng, L., & Miles, J. (2012). From Z to A: Using validity evidence to set performance standards. In G. J. Cizek (Ed.), *Setting performance standards: Foundation, methods and innovations* (pp. 301–322). New York, NY: Routledge.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2013). *Request for proposals. (RFP# 40-000-13-00027)*. Santa Fe, NM: New Mexico Public Education Department.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253–272). New York, NY: Springer.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 221–262). New York, NY: Macmillan.

Phillips, G. W. (2010). *Integration benchmarking: State education performance standards*. Washington, DC: American Institutes for Research.

Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundation, methods and innovations* (pp. 323–346). New York, NY: Routledge.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121 – 129.

Shepard, L. A. (2008). A brief history of accountability testing, 1965 – 2007. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 25 – 46). New York, NY: Routledge.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic – molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, *4*, 1 – 98.

Soares, J. A. (2012). Introduction. In J. A. Soares (Ed.), *SAT wars: The case for test-optional admissions* (pp. 1 – 9). New York, NY: Teachers College Press.

U.S. Department of Education. (2013, August 29). *States granted waivers from No Child Left Behind allowed to reapply for renewal for 2014 and 2015 school years*. [Press release]. Retrieved from http://www.ed.gov/news/press-releases/states-granted-waivers-no-child-left-behind-allowed-reapply-renewal-2014-and-201

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273 – 284). New York, NY: Springer.

Zwick, R. (2013). *Disentangling the role of high school grades, SAT scores, and SES in predicting college achievement* (Research Report No. RR-13-09). Princeton, NJ: Educational Testing Service.

Suggested citation:

Kannan, P. (2016). *Vertical articulation of cut scores across grades: Current practices and methodological implications in the light of the next generation of K – 12 assessments* (Research Report No. RR-16-29). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12115

Action Editor: James Carlson

Reviewers: Michael Zieky