Measuring the Power of Learning.®

Research Report
ETS RR–17-34

Articulation of Cut Scores in the Context of the Next-Generation Assessments

Priya Kannan

Adrienne Sgammato

December 2017

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

Articulation of Cut Scores in the Context of the Next-Generation Assessments

Priya Kannan & Adrienne Sgammato

Educational Testing Service, Princeton, NJ

Logistic regression (LR)-based methods have become increasingly popular for predicting and articulating cut scores. However, the precision of predictive relationships is largely dependent on the underlying correlations between the predictor and the criterion. In two simulation studies, we evaluated the impact of varying the underlying grade-level correlations on the resultant bias in cut scores articulated using the LR method. In Study 1, we compared different articulation methods (LR and equipercentile smoothing), and in Study 2, we evaluated different criteria for linking (e.g., adjacent grade or end of course). The collective results indicate that as correlations became smaller, cut scores articulated using LR-based predictions became increasingly biased when compared to a true value obtained under perfect correlation. The predicted values are significantly biased for lower achievement levels, irrespective of the linking criteria used. Results from these studies suggest that the LR method must be used with caution, particularly when articulating cut scores for multiple achievement levels.

Keywords Cut-score articulation; next-generation assessments; on-track cut scores

doi:10.1002/ets2.12162

Federal accountability mandates since the No Child Left Behind Act of 2001(2002) (NCLB) have required states to administer standardized tests and report results for all students (disaggregated by subgroups) enrolled in Grades 3–8 and in one high school grade level. This federal reporting mandate continues with the new Every Student Succeeds Act of 2015. In the 2015–2016 academic year, approximately 22 million students were projected to be enrolled in Grades 3–8 (Hussar & Bailey, 2014) in public schools across the United States. Each of these ∼22 million students will take a minimum of two (English language arts and mathematics) standardized assessments, and results must be reported for these students at multiple achievement levels (ALs). The federal mandates have required states to demonstrate growth or progress for students at each of the multiple ALs. However, for results on growth/progress to be reported, at the very least, the performance standards across the grades should be articulated (see Kannan, 2016). Unfortunately, this was not the case prior to NCLB.

In addition, concerns about the international competitiveness and college readiness of high school graduates (Barton, 2009) led to a rethinking of methods across the entire spectrum of the assessment system. Efforts at conceptualizing a set of common curriculum standards across states resulted in the Common Core State Standards (CCSS; see Council of Chief State School Officers & National Governors Association Center, 2010), which were written with the goal to ensure that students are "college and career ready" (CCR) at the end of high school. Since then, multistate assessment consortia (e.g., the Partnership for Assessment of Readiness for College and Careers [PARCC] and Smarter Balanced Assessment Consortium) have developed (and recently administered) assessments with the goal of identifying students who are on-track to being CCR at each grade level, wherein on-track cut scores at each lower grade level are directly articulated with being CCR at the end of high school. To establish CCR (and on-track) cut scores for these new assessments, there is an increasing focus on evidence-based standard setting (EBSS; Haertel, 2002; McClarty, Way, Porter, Beimers, & Miles, 2013; Miles, Beimers, & Way, 2010) that integrates evidence from multiple criteria by linking educational data with various national (e.g., National Assessment of Educational Progress) and international (e.g., Trends in International Mathematics and Science Study) benchmarks to broaden the definition of high school performance (e.g., Betebenner, 2012; Haertel, Beimers, & Miles, 2012; Phillips, 2012).

As the next-generation assessments (based on the CCSS or other state standards linked to CCR) are being administered in several states, there is an imminent push to ascertain if students are on-track to being CCR as early as Grade 3. And with

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

the goal of predicting students who are on-track at earlier grades, predictive methods (as an extension of the EBSS) have gained popularity in recent years (cf. Boyd, Davis, Powers, Schwartz, & Phan, 2014; Michigan Department of Education [MDE], 2011). However, the appropriateness of the cut-score articulation methods will be fundamental to making valid interpretations of the results from these assessments and ensuring that consistent decisions are made across the grade levels. It is well known that the ability to predict a criterion is dependent on the underlying correlation between a predictor and the criterion (see Cohen, Cohen, West, & Aiken, 2003). Therefore, in the context of the increased use of predictive methods in cut-score articulation, in this report, we present the results from two simulation studies in which we evaluated the robustness of the logistic regression (LR) method for articulating cut scores across the grades when the underlying correlations are varied.

Background

Ever since the initial push for reporting adequate yearly progress under NCLB legislation, several alternative cut-score articulation methods (e.g., statistical interpolation, impact percentage smoothing) have been offered (see Cizek & Agger, 2012; Kannan, 2016). In addition to statistical methods used to articulate cut scores, in practice, many states use a vertical articulation or smoothing committee, composed of panelists from grade-level standard-setting committees, to review the grade-level standards. This committee typically uses a combination of impact percentage smoothing and a review of the demands of the content standards at each grade level to articulate cut scores across the grades (for a full review of various vertical articulation methods employed by states, see Kannan, 2014).

Moreover, in the context of the next-generation assessments, several researchers (e.g., Bejar, Braun, & Tannenbaum, 2007; Huynh, Barton, Meyer, Porchea, & Gallant, 2005; Lewis & Haug, 2005) have recommended the use of holistic approaches that integrate educational policy, learning theory, and curriculum design in developing content and performance standards. Two such approaches, which involve the articulation of curriculum and test design, serve as promising solutions for the articulation of performance standards: (a) curricular (or content standard) articulation and (b) use of vertical scales in developing tests across the grades. The learning progressions framework (Smith, Wiser, Anderson, & Krajcik, 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. Use of vertical scales in developing cross-grade tests can provide an ideal solution for articulation of performance standards and offer several practical benefits (e.g., Petersen, Kolen, & Hoover, 1989; Yen, 2007) for tracking student growth and progress over the grades. However, these approaches are challenging to develop (e.g., Kolen, 2011; Patz & Yao, 2007; Yen, 2007) and have not been implemented widely (e.g., while the Smarter Balanced tests were based on vertical scales, the other large multistate consortia, PARCC, did not use vertical scales because of several practical limitations; for additional explications on the challenges in implementing vertical scales, see Kannan, 2016).

Despite the variety of methods available, in practice, post hoc performance standard articulation methods that employ some variant of impact percentage smoothing (i.e., using a post hoc smoothing procedure to minimize grade-level differences in percentages of students classified in each performance category) remain extremely popular among the states (see Kannan, 2014, 2016). This method does not make as many assumptions about the underlying scale and/or the comparability of the tests across the grade levels (see Kannan, 2016), which explains the popularity of this method, both in empirical evaluations (e.g., Foley, 2014) and operationally among the 50 states surveyed (see Kannan, 2014). Nevertheless, the impact percentage smoothing method makes important assumptions about the trajectories for student development that form the basis for the trajectory employed for smoothing (see Kannan, 2016). These assumptions about developmental trajectories are often derived from historical longitudinal data on student performance, but determining such a trajectory ahead of time is not feasible when historical data are unavailable (such as with the next-generation assessments that were only operationally administered for the first time in Spring 2015). Therefore, despite its inherent appeal, and the demonstration of its application in various contexts by researchers, this method has some notable limitations.

More recently, LR-based methods have become increasingly popular (e.g., Bay, Dunn, Kim, McGuire, & Sukin, 2012) for predicting and articulating on-track cut scores at lower grade levels by back-translating the benchmarked CCR cut scores to the lower grades. Additionally, ACT used LR-based methods to update its college readiness benchmarks on the ACT and Explore assessments (Allen, 2013). Specifically, hierarchical LR was conducted for each course and institution to determine the ACT score associated with the college readiness benchmarks. Similarly, the College Board used LR-based

methods to determine college readiness benchmark scores on the *SAT*® Reasoning test based on first-year college grade point average (Kobrin, 2007).

However, the process of using LRs in the articulation of cut scores across the grades in the K–12 assessment context is not very clear. Some solutions have been offered in state applications. One approach, as applied in Texas (see Boyd et al., 2014), is to perform a series of LRs to establish cut scores in an upper grade given student performance in the lower grade and present this as one piece of evidence to the standard-setting panelists to assist them in determining cut scores. Another method (e.g., Hao & Wyse, 2015; MDE, 2011) is to use the end-of-course (EOC) or high school performance standard for CCR as the criterion to predict on-track cut scores at all lower grade levels. A brief review of its application in a couple of state assessment contexts helps illustrate how the LR method may be applied.

For example, in Michigan, the MDE first established EOC cut scores for the Michigan Merit Examinations (MME) at the 11th grade by linking EOC test scores to course grades for freshman-year college students enrolled in Michigan public postsecondary institutions (MDE, 2011). *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 a 33% and 67% chance, respectively. The MDE used a series of LR analyses to vertically articulate the EOC cut scores to determine on-track performance at the lower grades for the Michigan Educational Assessment Program (MEAP). In addition, it used a signal detection theory (SDT)-based method to articulate the cut scores. The two methods had somewhat different but related aims: The LR method aimed to identify the score that provides a fixed probability of success, whereas the SDT method aimed to maximize consistent classifications for any given criterion—in other words, the articulated cut score in this method would be the score that maximizes the true positives and the true negatives. When monotonicity can be assumed, the articulated cut scores using LR and SDT should be equivalent. Therefore the MDE validated the results obtained from the LR method against the SDT method and found that the results were largely comparable (MDE, 2011). Yet, the MDE decided to pursue the SDT method for its future evaluations (see Hao & Wyse, 2015).

In the Texas application, Boyd et al. (2014) reported the results of an EBSS process (which included empirical studies, policy considerations, and face-to-face panel-based standard-setting meetings; see McClarty et al., 2013) to articulate cut scores for the State of Texas Assessments of Academic Readiness (STAAR). One component of that EBSS process was to present panelists with a host of information to assist them in establishing the cut scores. One such source of evidence included empirical studies using LRs that linked performance standards to postsecondary success, used to estimate the potential EOC cut scores. Next, Grade 8 reading and mathematics and Grade 7 writing assessments were linked internally to the EOC assessments and internally for Grades 3–8 in a pairwise fashion. Additionally, the Grade 8 and Grade 7 assessments were linked to external benchmarks (e.g., EXPLORE and ReadiStep). LRs were used to make all of these linkages. The articulated (internal) linkages predicted the probability of attaining an AL on the STAAR EOC assessments, given a student's performance on the STAAR Grade 8 reading and mathematics assessments or Grade 7 writing assessment. To identify a range of cut scores, the authors operationalized a 60% probability as having a *reasonable likelihood* and a 75% probability as having a *high likelihood* of success in attaining Level II (or satisfactory) performance on the EOC assessment. Similarly, they used LRs to identify a range of cut scores for each of Grades 3–7 using the next subsequent higher grade as the criterion. STAAR Grades 3–8 reading and mathematics assessments were also aligned using a vertical scale. The results from the linking studies provided the initial ranges for the AL cut scores, which were then evaluated for reasonableness in a panel-based standard-setting meeting.

Several measures were taken in the Michigan and Texas applications to alleviate cumulative measurement error in making multiple linkages. For example, the MDE obtained matched data sets for cohorts of students ranging across 6 years. The MDE included at least two cohorts for each back-mapping (i.e., regressions that link each higher grade cut score to the lower grade cut scores). In addition, to minimize the number of links and the resultant cumulative measurement error, a stepwise linking procedure was adopted, and the number of links per grade level was minimized to no more than three. 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 have already been linked to MME [i.e., Grade 11]). The MDE evaluated the classification consistency rates from year to year and found the lowest classification consistency for MME (Grade 11) to college grades, and the remaining consistency rates showed a high degree of stability from grade to grade. They reasoned that the lower classification consistency from MME to college grades was due to the largest construct shift (in terms of what is being measured in Grade 11 and in college) for this group.

Similarly, Boyd et al. (2014) used coarsened exact matching to match students across two subsequent grades. Because 2012 marked the first operational administration of the STAAR, they did not have access to a single cohort of students that had taken two subsequent-year assessments. Therefore, to link students from two subsequent grade levels, the student cohorts were matched based on their prior-year test scores on the Texas Assessment of Knowledge and Skills in the same content area. Boyd et al. used adjacent-grade linking to minimize cumulative measurement error, and only one link from a lower to an immediate higher grade was made to identify the initial range of cut scores for each grade level. They reasoned that at each grade level, the criterion of interest was successful performance on the subsequent grade level, and they were not interested in linking back to the EOC criterion. The initial range of cut scores identified using the adjacent-grade linking method was then evaluated for reasonableness in a panel-based standard-setting meeting using a Bookmark method.

Despite the cautionary steps taken in these applications (Boyd et al., 2014; MDE, 2011), Ho (2013) has effectively expressed a caveat to predictive standard setting (i.e., using LRs). In predictive standard setting, empirically defensible predictive statements are attached to the ALs classified on a score scale. Ho has argued that an appropriate distinction needs to be made between a cut score that is somehow linked to future performance and a cut score that in fact predicts future performance (when the predictor and indicator variables are not highly correlated, the predictive variance is likely very small). In not making such a distinction, interpretations of stringency in classifying students become confounded with the predictive utility of the test. For example, for two grades that have a correlation of 0.3, using regression-based articulation methods would result in a regression that explains less than 9% of the variance in the outcome. The standards established for on-track performance can only support the necessary inferences to the extent that the test can predict the outcome (in this case, the readiness for college or careers). In other words, to the extent that the higher grade criterion (e.g., CCR) used is not highly correlated with the lower grade predictor (e.g., on-track cut score), the resultant predictive relationship, though mathematically valid, might lead to incorrect interpretations due to the lack of predictive power.

The Current Studies

In the series of two simulation studies reported here, we evaluated the impact of varying the underlying correlations between grade-level assessments on the articulated cut scores. Two different articulation methods (LR and equipercentile smoothing) and three different LR linking procedures (i.e., EOC as criterion, adjacent grade as criterion, two-step linking) were evaluated in Study 1 and Study 2, respectively. In Study 1, we investigated the impact of varying the underlying correlations on the cut-score recovery of the LR and equipercentile methods. As pointed out previously, despite its inherent appeal, the equipercentile method does make some big assumptions about the equality of the trajectories of development across grades. Yet, because this method is likely not going to be influenced by the underlying correlations, in Study 1 we used the equipercentile smoothing method as a comparative criterion to evaluate the bias in cut-scores articulated using LRs. In Study 2, we investigated the impact of using various criteria for linking (i.e., EOC as criterion, adjacent grade as criterion, two-step linking) on the cut-score recovery of the LR method. In both studies, we evaluated the average absolute bias in recovering a "true" value. The true value was estimated under conditions of perfect correlation across all grades and by using the EOC cut score as the criterion for all linkages.

Study 1

Previous literature (e.g., Ho, 2013; Zwick, 2013) has indicated that the correlation between a predictor and a criterion will influence the validity of the predicted relationship such that in the absence of a perfect correlation ($r = 1.0$), a regression-based relationship will result in a prediction biased toward the mean of the future cut score (Ho, 2013). Furthermore, bias increases as correlation decreases (Ho, 2013). In the context of K–12 assessments, cut scores are determined for multiple ALs at each grade, and it is important that the cut scores for each of these ALs be articulated across the grades. It is not clear from previous research how the correlation between grade-level assessments is likely to impact the cut-scores articulated for multiple ALs. Therefore, in this study, we evaluated the impact of varying the underlying correlations for seven grades and three ALs within each grade level. The following specific research question was investigated in this study: When cut scores for multiple ALs are articulated across the grades, how does varying the correlation across grade-level assessments affect the articulated cut scores for each AL using two cut-score articulation methods (i.e., LRs and equipercentile smoothing)?

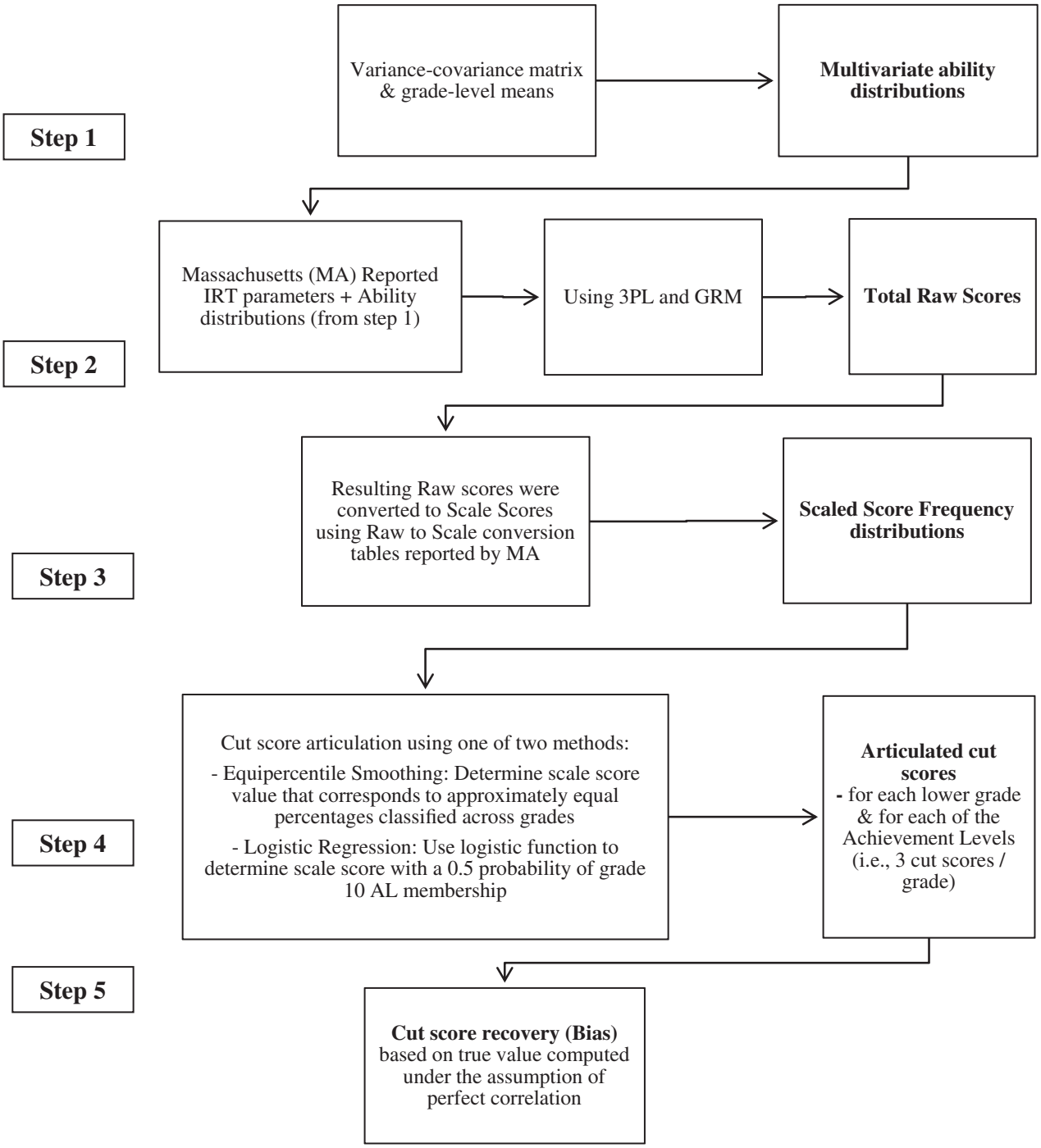

Figure 1 Simulation procedure: Study 1. *Note.* IRT = item response theory; 3PL = three-parameter logistics; GRM = graded-response model; AL = achievement level.

Method

The simulation procedure, including data sources and methods of data analysis, for Study 1 are described here.

Simulation Procedure

All data were generated using SAS 9.3, and 1,000 replications were performed across grade levels for each articulation method. The simulation procedure is presented in Figure 1 and fully described here.

Table 1 Correlation Matrix Used to Generate Grade-Level Data

	Grade level						
	10	8	7	6	5	4	3
Grade level							
10	1.00	.80	.70	.60	.50	.40	.30
8	.80	1.00	.80	.70	.60	.50	.40
7	.70	.80	1.00	.80	.70	.60	.50
6	.60	.70	.80	1.00	.80	.70	.60
5	.50	.60	.70	.80	1.00	.80	.70
4	.40	.50	.60	.70	.80	1.00	.80
3	.30	.40	.50	.60	.70	.80	1.00
Mean	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The multivariate normal ability distribution for 20,000 students at each grade level was generated in Step 1 of the simulation. Table 1 presents the correlation matrix and grade-level means used to generate a multivariate normal ability (theta) distribution across the grades. Correlations between the grade levels were varied such that adjacent grades were more highly correlated than grades that were further apart. Moreover, we wanted to evaluate the impact of various degrees of underlying correlation between predictor and criterion and therefore used correlation values ranging from .3 all the way to .8 in our simulated data set.

Data Sources

To generate simulated data sets for seven grade levels in Step 2, we used publicly available item parameters reported for the Massachusetts Comprehensive Assessment System (MCAS) mathematics assessments. The item response theory (IRT) parameters for Grades 3 – 8 and 10 mathematics assessments for the 2012 MCAS administration are available from the Massachusetts Department of Elementary and Secondary Education (MDESE; 2012). The multivariate normal ability distributions generated in Step 1 were used with the grade-level item parameters (for both the dichotomous and polytomous items) to generate student data for each grade level; the three-parameter logistic IRT model (Lord, 1980) was used to generate the dichotomous item responses, and the graded-response model (Samejima, 1969) was used to generate the polytomous item response data. Total raw score was computed for each simulated student by summing scores on all items.

In Step 3, raw scores for each student were converted to scale scores (on a scale of 200 – 280) using the raw-to-scale conversion tables provided in the Massachusetts 2012 technical report (MDESE, 2012).

Articulation Methods

In Step 4, two different articulation methods (i.e., LR and equipercentile smoothing) were used to articulate cut scores for the lower grade levels. Cut scores for all grades in Massachusetts are set at a scaled score of 220, 240, and 260 for the *basic*, *proficient*, and *advanced* level, respectively (see MDESE, 2012). To truly determine on-track to CCR cut scores (as delineated in the CCSS), we used the cut scores for EOC (i.e., Grade 10) as the criteria to obtain articulated cut scores for all lower grade levels in this study.

Equipercentile smoothing. The scaled score frequency distribution for each of the lower grade levels was used to obtain cut scores that corresponded to the same percentages of students classified in the three ALs as in Grade 10. The final articulated cut score was the scaled score value that corresponded to the smoothed cumulative percentage value for each AL and each grade.

Logistic regression. An LR analysis was conducted for each cut score using the continuous scale score distribution to predict a dichotomous AL membership. The Grade 10 cut scores were used as the criteria that are predicted from each lower grade scale score distribution. The logistic function was used to find the scale score at which the probability of each AL membership in Grade 10 was equal to .5. The corresponding scale score value was used as the cut point for the lower grade level between adjacent ALs.

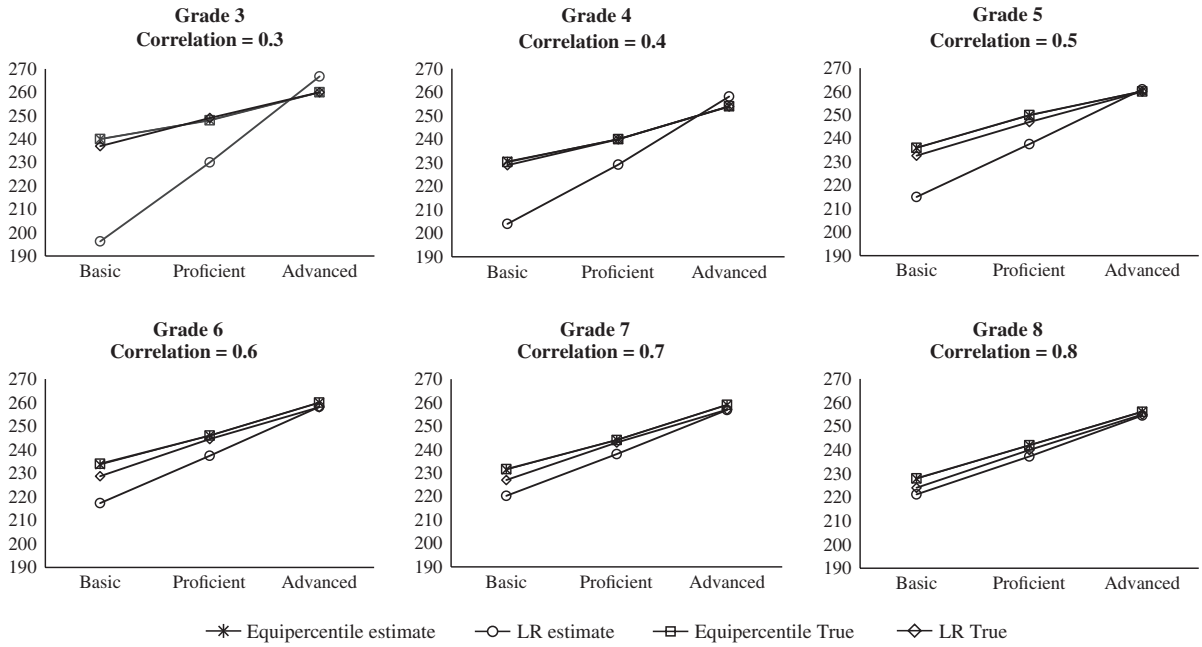

Figure 2 Achievement level cut scores by method and grade compared to true values obtained under a condition of perfect correlation.

Data Analysis

Finally, in Step 5, root mean square deviations (RMSDs) were computed to evaluate the cut-score recovery for each of the two articulation methods in this study. Steps used to compute the dependent variable are described below in the following paragraphs.

Root mean square deviation. True values of the cut scores are not known. Therefore, under the assumption that correlation between the predictor and criterion would be inversely related to measurement error, in this study, true values were computed as the average estimate of cut scores under conditions of perfect correlation. That is, we used $r = 1.0$ for each element in the correlation matrix and the Grade 10 cut scores as the criteria to obtain the articulated cut scores for each lower grade. True cut-score values for each articulation method were computed as the average estimate across 100 replications and are presented in Figure 2. Bias in cut-score estimation for each articulation method was computed, for each replication, as the deviation between the estimated cut score and the true cut score. Additionally, for each condition (i.e., by method, grade, and AL), the RMSD was calculated as follows:

$$\text{RMSD}\ (\theta) = \sqrt{\frac{\sum_{r=1}^{\text{NR}} \left(\widehat{\theta}_r - \theta\right)^2}{\text{NR}}},$$

where $\widehat{\theta}$ is the estimated cut score for that method–grade–AL combination, θ is the corresponding estimated true cut score, r refers to the replication, and NR refers to the total number of replications. RMSD was computed for each AL at each grade and for both articulation methods. Additional parametric analyses, including multivariate analysis of variance (MANOVA), were considered to evaluate the absolute bias (i.e., $\sqrt{(\widehat{\theta}_r - \theta)^2}$) computed for each replication. However, the distribution of these variables did not meet the assumptions of normality and homogeneity of variance, and therefore such analyses were not carried out.

Results

The average cut-score estimates along with the true values across the ALs are graphically illustrated for each grade level and each method in Figure 2. As expected, across the ALs and grade levels, the estimated cut scores do not deviate largely

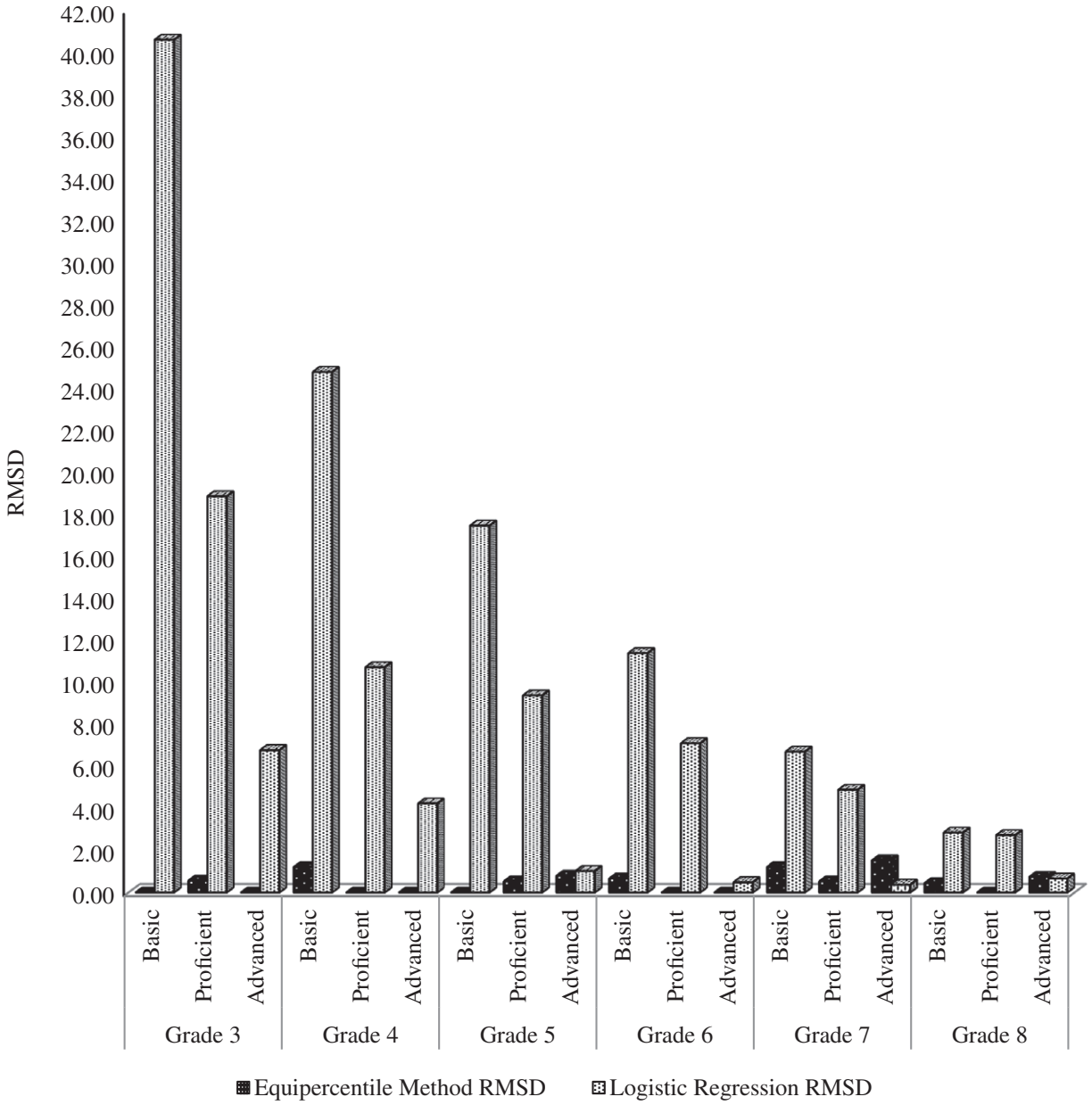

Figure 3 Mean RMSDs in estimating multiple achievement level cut scores using the equipercentile and logistic regression methods.

from the true cut scores for the equipercentile method. This confirms that the equipercentile method is not sensitive to the underlying correlations between grade-level assessments. However, the cut-score estimates for the LR method increasingly deviate from the true values as the correlation between the EOC (criterion) and the respective grade level becomes smaller (see Table 1 for the correlation matrix used to simulate data). This discrepancy between estimated and true values is largest for the lowest AL (Cut 1 or *basic*) and smallest for the highest AL (Cut 3 or *advanced*); in addition, this discrepancy is largest for Grade 3 (smallest simulated correlation with the criterion) and smallest for Grade 8 (largest simulated correlation with the criterion).

The RMSD in estimating the cut scores using each method for each grade and AL are graphically represented in Figure 3. Across the ALs and grade levels, the bias in recovering the true cut score for the equipercentile method is zero or close to zero. The one exception to this is the *advanced* cut score for Grade 7, for which the RMSD for the equipercentile method is slightly larger, and in fact larger than the LR method (equipercentile RMSD = 1.54;

Table 2 Average Estimated Scale Score for Each Grade in One Study Replication

Grade	Average scale score	*SD*	Sample size	Cut-score		
				Basic	Proficient	Advanced
3	238.31	18.26	20,000	220	240	260
4	239.89	18.18	20,000	220	240	260
5	248.79	18.71	20,000	220	240	260
6	250.94	19.16	20,000	220	240	260
7	254.62	18.57	20,000	220	240	260
8	253.83	19.25	20,000	220	240	260
10	257.93	19.59	20,000	220	240	260

Note. The cut scores for basic, proficient, and advanced levels are for the MA data set and are not the estimated cut scores based on the different articulation methods; these values are provided in Figure 2.

LR RMSD = 0.37). However, overall for the LR method, the observed absolute bias (RMSD) gets larger as the correlation between the EOC (criterion) and the respective grade level becomes smaller (i.e., Grade 3 was simulated to have the lowest correlation with EOC; see Table 1); moreover, this observed bias is largest for the lowest AL (i.e., *basic*).

These findings confirm what would be expected based on previous research (Ho, 2013). Ho (2013) pointed out that the articulated cut scores will be biased in the direction of the future (criterion) cut score and that the bias will be more pronounced when the criterion cut score is above or below the mean. However, because the focus of this study was on the articulation of multiple ALs, the results from this study specifically demonstrate that the estimated cut scores for the lowest AL (i.e., *basic*), which are likely to be significantly below the EOC means, become more and more lenient at the lower grade levels, as the correlation with the criterion becomes smaller. Conversely, the cut score of 260 for the highest AL (i.e., *advanced*) was likely closer to the mean of the Grade 10 distribution (see Table 2). Therefore, despite the variability in the underlying grade-level correlations, the predicted cut scores for this AL remain largely unbiased in using an LR method.

Study 2

The purpose of this simulation study was to evaluate the bias of several LR linking models in predicting articulated cut scores. Despite the increase in popularity of LR methods for articulating cut scores, results from Study 1 indicate that cut scores articulated using the LR method have large bias, especially at lower grades, where the correlation between predictor and criterion is likely to be smaller. However, all grades were regressed on the criterion (Grade 10 or EOC) in the first study, and therefore the impact of the underlying correlations on the predicted cut scores was large in this study. Moreover, predictive methods are subject to regression effects (Ho, 2013), and therefore the number of linkages when regression-based methods are used to articulate multiple cut scores across grade levels should also be considered. To evaluate if alternative linking models might help reduce the degree of bias observed in Study 1, different LR models, each using different criteria and different numbers of links, were used to predict cut scores at lower grades in Study 2.

The following LR linking models were evaluated: (a) Grade 10 (or EOC) as the criterion for all linkages; (b) adjacent higher grade as criterion for each linkage; and (c) two-step linkages—from Grade 10 to Grade 7, and a second linkage from Grade 7 to the specific lower grade (either Grade 3 or Grade 4). It was expected that the two-step model might not only alleviate the impact of low correlation between predictor and criterion (a condition known to increase bias in cut-score prediction from the results observed in Study 1) but also help reduce cumulative measurement error by reducing the total number of linkages made for all predictions. This is because only two linkages were made to predict the cut scores at the lower grade levels (Grades 3 and 4) in this model, and the correlations between predictor and criterion were controlled at .5 or greater for all linkages. Finally, given the degree of bias resulting from the EOC as the criterion model observed in Study 1, these alternative linking models (i.e., adjacent grades and the two-step linkage model) were evaluated. It is conceivable that the method with the most linkages (i.e., the adjacent grade as criterion) could result in an accumulation of systematic variability between the true and estimated cut scores. Likewise, with only two LR linkages in the two-step procedure, less bias may be introduced, thus the true and estimated cut scores may be more similar to

one another than when compared to the adjacent-grades method. Similarly, with slightly higher correlations compared to the EOC as criterion model, the two-step linkage model could conceivably result in reduced bias. Therefore the following specific research question was investigated in this study: To what extent do different LR linking models alleviate the bias when articulating cut scores using LR observed in Study 1?

Method

The data sources for this study were identical to those used in Study 1. Therefore, in this section, we fully describe the procedures and the methods of data analysis used specifically in Study 2.

Simulation Procedure

All data were generated using SAS 9.3; 1,000 replications were performed using each of the three LR models evaluated in this study. Steps 1, 2, and 3 of the simulation procedure were identical to those used in Study 1 (see Figure 1); the only difference was in Step 4, as follows.

Logistic Regression Models

Once data were generated for each grade (as described in Study 1) and a scale score was computed for each student, in Step 4, the LR models were applied; an LR analysis was conducted for each cut score using a continuous scale score to predict a dichotomous AL membership. The logistic function was then used to find the scale score at which the probability of AL membership was equal to .5. That scale score value was used as the cut point between adjacent ALs.

End-of-course as criterion. Using matched data sets, each cut score for each lower grade (Grades 3 – 8) was estimated with an LR based on the known AL membership in Grade 10. This procedure resulted in only one linkage to obtain the on-track cut score for each lower grade level; however, the predictor – criterion correlation underlying each of these predictive relationships was different: as high as .8 for the Grade 8 linkage to as low as .3 for the Grade 3 linkage.

Adjacent grade as criterion. Using matched data sets, each cut score for each lower grade (Grades 3 – 8) was estimated with an LR based on the known AL membership in the adjacent higher grade. In other words, Grade 8 cut scores were first predicted based on known Grade 10 AL memberships. Once Grade 8 students were placed into the appropriate ALs based on their scale scores, the cut scores for Grade 7 were predicted; this was repeated for Grade 6 from Grade 7, and so on. This procedure resulted in six linkages to obtain the final articulated on-track cut score at Grade 3, where within each linking, an LR analysis was run for each AL cut score.

Two-step linkage: End-of-course criterion to intermediate grade to lower grade. This model is similar to and inspired by the one used by Michigan (see MDE, 2011); the mechanics of this model are the same as for the other two models, except that we used two linkages to predict cut scores for the lowest two grades (i.e., Grades 3 and 4, known to be the most biased from Study 1 results). For each of these lower grades, two linkages were made: one from Grade 10 to Grade 7, and others from Grade 7 to Grade 4 and from Grade 7 to Grade 3. For matched data sets, Grade 7 AL membership is predicted based on known Grade 10 membership, and Grade 4 (or 3) AL membership is predicted based on the predicted Grade 7 results. For each linking, an LR analysis was run for each AL cut score.

Data Analysis

Similar to Study 1, in Step 5 (see Figure 1), RMSD was computed to evaluate the cut-score recovery using each of the three LR models. The steps used in calculating the bias are the same as those used in Study 1 and are only briefly described here.

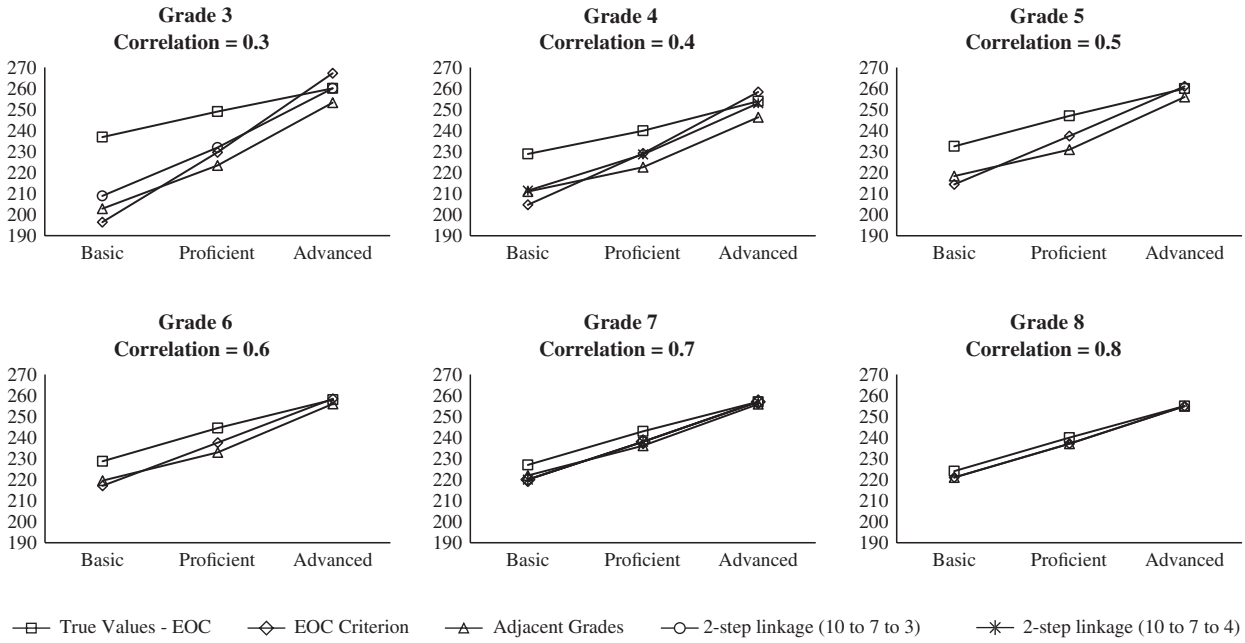

Figure 4 Achievement level cut scores across grades predicted using the three logistic regression models and true values obtained under a condition of perfect correlation.

Root mean square deviation. Similar to Study 1, true cut scores were computed when the correlations among all Grades 3–8 and 10 were fixed at 1.0 and where all grades were predicted from Grade 10 (using EOC as criterion). RMSD was calculated for each AL cut score at each grade, for each replication, and for each of the three LR models. Also as in Study 1, additional parametric analyses, including MANOVA, were considered to evaluate the absolute bias (i.e., $\sqrt{(\hat{\theta}_r - \theta)^2}$) computed for each replication. However, the distribution of these variables did not meet the assumptions of normality and homogeneity of variance, and therefore such analyses were not carried out.

Results

The mean cut-score estimates for each grade, averaged across the replications, are presented in Figure 4 for each of the LR models. Across all models, the estimated cut scores are almost always biased and lower than the true cut scores. The *advanced* cut score is less biased across all three LR models compared to the other ALs, and this is more evident for the lower grades (i.e., where the correlation with the criterion was lowest). In particular, for the two-step linking models, the estimated *advanced* cut scores are almost identical to the true scores, even at the lowest grade levels, which exhibited the most bias in Study 1 (in which the EOC as criterion model was used). Whereas the *advanced* cut-score estimates from the EOC as criterion model are almost identical to the true scores for Grades 8, 7, 6, and 5, the estimated *advanced* cut scores are biased and noticeably higher than the true cut scores for the lowest two grades—a pattern also observed in Study 1. Finally, the *advanced* cut scores estimated using the adjacent-grades model are biased and noticeably lower than the true cut scores for the lowest two grades (i.e., for which the correlation with the criterion was lowest).

Across the board, for the other two ALs (i.e., *proficient* and *basic*), as the correlation between the grades decreases (i.e., at lower grade levels), the predicted cut scores are increasingly more lenient than the true cut scores, and this bias is most pronounced for the lowest, *basic* AL. It should be noted that the estimated *basic* cut score for the EOC as criterion model at Grade 3 was lower than the range of estimable scale scores (the lowest observable scale score was 200).

To evaluate the recovery of the true cut scores, the RMSD was calculated for each LR model and is presented graphically in Figure 5. Identical to the results observed previously, across all grades and for all three LR models, the values of RMSD and bias are always smallest for the *advanced* cut scores. Moreover, across models, RMSD is smallest in Grades 7 and 8 (i.e., grades that are more highly correlated with EOC), with RMSD becoming increasingly larger through

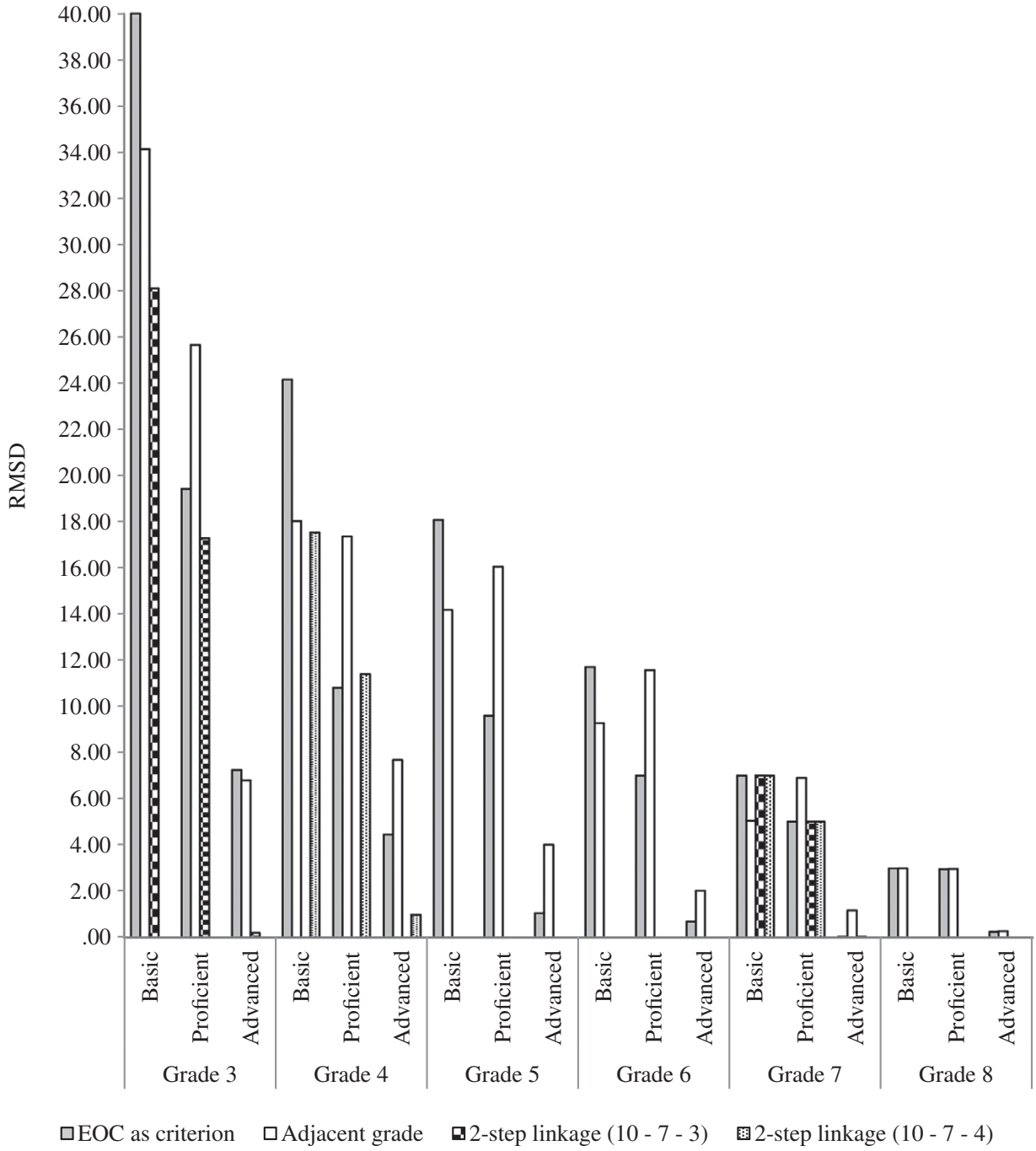

Figure 5 Mean root mean square deviations in estimating multiple achievement level cut scores using three logistic regression models.

Grade 3 (i.e., the grade simulated to have the lowest correlation with EOC). The RMSD is smallest for the *advanced* cut score in Grade 7, ranging from 0 for the two-step models to 1.16 for the adjacent-grades model. Even at Grade 7, the RMSD values are noticeably larger for the *proficient* and *basic* cut scores, ranging from about 5 to 7 for each of the three models.

Across all grades, RMSD values were the lowest for the two-step model, with comparatively larger RMSD values observed for the other two models. However, it is interesting to note that, across the grades, the EOC as criterion model was the most biased in estimating the *basic* cut score, whereas the adjacent-grades model was the most biased in estimating the *proficient* and *advanced* cut scores. Overall, across the 18 total cut points associated with estimates from both the

adjacent-grades model and the EOC as criterion model, the adjacent-grades model is associated with higher values of RMSD for 12 of those cut scores. From Figure 5, for Grade 3, where the correlation among predictor and criterion was set lowest, the adjacent-grades model yields comparatively much larger RMSD values than the two-step model. However, even the two-step model yields high RMSD values, ranging from 17.29 to 28.10 at Grade 3 for the *basic* and *proficient* ALs.

Discussion

In the two studies presented here, we used different articulation methods and different linking models to evaluate the impact of varying the underlying grade-level correlations on the resultant bias when cut scores are articulated for multiple ALs. In Study 1, we evaluated the impact of varying the underlying correlations on multiple AL cut scores articulated using two methods. In Study 2, we evaluated alternate LR linking models that may help alleviate the bias in articulated cut-scores observed in Study 1. Both studies used some form of an LR method to estimate cut scores for lower grades based on known information from Grade 10, and the results involving the LR models across the studies are similar in terms of pattern and magnitude. The combined results from the studies indicate that some AL cut-scores articulated using the LR method became increasingly biased as the correlation between predictor and criterion became smaller (i.e., at the lower grade levels).

In addition, several specific patterns of results emerged from these studies, which are summarized here. First, even though the cut scores articulated using the LR method were found to be biased across all grade levels (levels of correlation), this bias was somewhat negligible for correlations of .7 or higher (note that minimal bias is observed for Grades 7 and 8 in both studies). Second, for all LR models, whereas large amounts of bias were observed for the *basic* and *proficient* cut scores, the bias for the *advanced* cut score was minimal; this pattern was observed for all grades (and levels of correlation). Third, the two-step linking model produced relatively smaller bias when compared to the other two models; the bias produced by this model for the *advanced* cut score was almost negligible.

Practical Implications

Overall, while results from these studies suggest that LR-based methods must be used with caution in articulating cut scores, such caution is particularly necessary when either the criterion cut score (e.g., CCR cut score at the high school level) is not near the criterion mean or when the correlations between the two grades used to make cut-score predictions is .6 or lower. In essence, a very interesting pattern of results emerged from our two studies; that is, despite the variability in the underlying grade-level correlations, the predicted cut scores for the *advanced* AL in the Massachusetts data set remained largely unbiased (particularly for the two-step linking model in Study 2). This was the case likely because the criterion cut scores (i.e., Grade 10 cut scores) for the *advanced* AL were closer to the mean of the Grade 10 distribution in our simulated data set (see Table 2).

The pattern of results described herein has some important practical implications for using the LR method in cut-score articulation. It should be noted that certain ALs are more important for practical decision-making purposes. For example, the *proficient* cut score has been the focus of accountability decision making in the post-NCLB era. Similarly, the CCR or on-track to CCR cut scores are likely to be the focus in the next-generation assessments. Even though results from our studies point to the unbiased prediction of the *advanced* cut score, these results should not be interpreted to indicate that the LR method can produce relatively unbiased cut scores for the highest or *advanced* proficiency level (or AL)—the label (or level) of the predicted cut score is immaterial. These results simply mean that, if the LR method is used to predict cut scores at lower grades, there would be minimal prediction errors and biases if the criterion cut score for this critical AL (be it *proficient* or CCR) is close to the mean of the criterion distribution. Therefore, in practical applications, if the LR method is used to predict a single on-track cut score, and if the criterion CCR cut score at the high school level is set close to the mean of the EOC distribution, then, irrespective of the underlying grade-level correlations, the articulated cut scores at the lower grade levels might be relatively unbiased and closer to their true values. Moreover, this could be more accurately accomplished for the lower grades by employing a two-step linking model, because the predicted cut scores for the *advanced* AL were the least biased for this model in our studies.

However, when the criterion cut score deviates from the mean of the distribution, the articulated cut scores at the lower grade levels (i.e., grades with lower correlation with the criterion) are likely to be stringent or lenient to the corresponding

degree. Even at Grade 6, with sufficiently high correlations with EOC (∼.6), the bias in estimation for the *basic* and *proficient* cut scores in the Massachusetts data set is arguably nonnegligible. Therefore, for tests that aid in making high-stakes decisions, for ALs where the criterion cut score deviates from the mean of the criterion distribution, it would be advisable to use predictive articulation methods with caution and perhaps supplemented by other methods. This may be practically applied by developing a briefing book (Haertel et al., 2012) in which articulated cut scores using LR are provided as one piece of evidence along with articulated cut-score data resulting from other methods (e.g., impact percentage smoothing, statistical interpolation). This briefing book may then be provided to a cross-grade panel of subject matter experts who might consider this information along with other pieces of evidence and use an informed judgment process in articulating cut scores across grades.

Limitations and Future Directions

In the absence of a publicly available correlation matrix, we assumed that the correlation matrix we used reasonably resembles the true correlations among grades "in the field." Furthermore, we assumed that our definition of true cut scores (i.e., all grades predicted from EOC-CCR) is reasonable. However, this might not reflect how on-track prediction is operationally implemented and interpreted in several states. Several states have, in fact, made the assumption that, at each grade, it is most important to predict if the student is on-track for the subsequent grade (see Kannan, 2014) and have used the adjacent higher grade cut scores as the true criterion. For such applications, we acknowledge that the correlations between two adjacent grades are likely to be much larger than those assumed for certain grades in this study, and LR-based prediction models may yield unbiased predicted cut scores for all ALs.

Moreover, both our studies used a single data source (i.e., the Massachusetts 2012 publicly available parameter estimates) and simulated data based on that single source. Therefore results for both studies were influenced by the same presumed regression effects whereby the estimated cut scores were appreciably more lenient when compared to the true values, and especially so for the lowest, *basic* AL. This was likely the case, because the criterion (i.e., Grade 10) cut scores for the *basic* AL in the Massachusetts data sets were likely significantly below the mean of the Grade 10 distribution, while the criterion (Grade 10) cut scores for the *advanced* or highest AL in the Massachusetts data sets were likely closer to the criterion mean. Future studies should aim to explore data sets from different states (or completely simulated data sets). This would help (a) evaluate whether and to what extent the specific data source has resulted in the pattern of results observed in this study and (b) cross-validate if the articulated cut scores (for the AL with criterion cut scores closer to the mean) are in fact less biased across all levels of correlation. This would also serve to reemphasize the point that the level (or label) of the specific AL is immaterial—what matters is if the AL that is used to predict regression-based cut scores has a mean close to the mean of the criterion distribution.

In addition, we used a logistic function to find the scale score at which the probability of AL membership was equal to .5. We used a .5 probability to predict LR-based cut scores because this reflects common practice to determine AL membership in state applications (including the Michigan and Texas examples reviewed in this report). It is possible that using a different probability to predict AL membership would have resulted in a pattern of results different from what was observed in this study. Future studies should evaluate other probability values, such as .67, in predicting AL membership to determine the impact of varying the underlying correlations on the bias in predicted cut scores.

Finally, even though the results for the equipercentile method from Study 1 exhibit low bias across all grades and all cut scores, these results should not be taken to indicate that this method was superior to the LR method. Unlike regression-based methods, the equipercentile method is not impacted by the underlying correlations—and this was essentially the reason that this method was used on a comparative basis in Study 1, so that we could demonstrate the degree of bias that would be seen in using a regression-based method. Moreover, we also assumed that similar percentages of students will be classified at the various ALs across grades. However, in practice, there have been observed patterns in state testing (see Kannan, 2014) in which proficiency rates become notably lower or higher for higher grades, or in other instances, they might demonstrate increasing rates in elementary school, plateau in middle school, and decelerate at high school. In such instances, an equipercentile smoothing method would not be advisable, and it would be very important to evaluate historical performance data before this method is adopted. Therefore, though the equipercentile method is very popular and predominantly used in a number of states (see Kannan, 2014), it is not necessarily free of flaws (see Kannan, 2016). Nevertheless, if a vertical scale is used along with expert-based mediation using a vertical articulation

or smoothing committee, a number of the concerns regarding assumptions made by the equipercentile method may be mitigated.

Therefore, prior to selecting any method for articulating cut scores across grades, a first step is to verify that the lower grade test is indeed related to the outcome of interest, for example, CCR. However, to do this, longitudinal data are required. Most states rolling out a next-generation CCR assessment will not have data linking Grade 3 to CCR, which limits the application of several articulation methods that require longitudinal data for validation or historical data to assume appropriate trajectories for smoothing. In addition, the nuances of each articulation method, such as evaluating the means of the distributions, underlying grade-level correlations (particularly when considering regression-based methods), the normality or regularity of the score distribution, reasonableness of assumptions for articulating the cut scores based on historical data, and so on, must all be considered before a decision to adopt a method (or combination of methods) is made. For example, if adjacent grades are being used to make cut-score predictions, it is likely that the LR method is reasonable to use. Alternatively, when historical data on student performance are available, it would be reasonable to use an impact percentage smoothing method selected to reflect what the historical data suggest. Future simulation studies should evaluate various combinations of articulation methods under different assumption violations to be able to guide practitioners in choosing the best course of action.

Conclusions

Overall, results from this study indicate that the LR method must be used with caution in predicting or articulating cut scores across the grades. When states are considering a predictive method to articulate cut scores, a first step would be to verify that the lower grade test, say, Grade 3, can indeed predict the outcome of interest, say, CCR, at the higher grade. In addition, it might help to temper results based on predictive relationships using supplementary methods (e.g., an impact percentage smoothing method guided by historical trend data)—this aligns with recommendations suggested by other researchers (e.g., Haertel et al., 2012; McClarty et al., 2013). Research (see Kannan, 2014) has shown that several states have already been doing this to some extent and have tried to use a combination of methods to articulate cut scores across the grades. Results from simulation studies such as ours point out the limitations of particular methods and help practitioners make an informed decision in choosing appropriate cut-score articulation methods for different circumstances that would support clear and accurate interpretations of student ability for the stakeholders. Nevertheless, such simulation studies are only a first step and should be supplemented by resampling studies that use operational state assessment data to replicate these results—the reasonableness of using various articulation methods should be clearly evaluated by testing the assumptions, and the impact of various cut-score articulation methods (or a combination of methods) should be weighed by evaluating the resultant bias from employing these methods.

References

Allen, J. (2013). *Updating the ACT college readiness benchmarks* (ACT Research Report Series No. 2013 [6]). Iowa City, IA: ACT. Retrieved from ERIC database. (ED546851)

Barton, P. E. (2009). *National education standards, getting beneath the surface* (ETS policy information perspective). Retrieved from https://www.ets.org/Media/Research/pdf/PICNATEDSTAND.pdf

Bay, L., Dunn, J., Kim, W., McGuire, L., & Sukin, T. (2012). *Developing achievement levels on the 2011 National Assessment of Educational Progress in Grades 8 and 12 writing* (Technical report). Retrieved from http://www.nagb.org/content/nagb/assets/documents/publications/achievement/developing-achievement-levels-2011-naep-grade8-grade12-writing-technical-report.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 schools* (pp. 1–30). Maple Grove, MN: JAM Press.

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.

Boyd, A., Davis, L. L., Powers, S., Schwartz, R., & Phan, H. (2014, April). *Evidence based standard setting: Vertically aligning Grades 3–8 assessments.* Paper presented at the meeting of the National Council for Measurement in Education, Philadelphia, PA.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Council of Chief State School Officers & National Governors Association Center. (2010). Common Core State Standards initiative. Retrieved from http://www.corestandards.org/

Every Student Succeeds Act (ESSA) of 2015. Pub. L. No. 114–95, § 115, S. 1177 (2015).

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.

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.

Hao, S., & Wyse, A. E. (2015, April). *Developing college and career readiness cut scores in one state.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

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.

Hussar, W. J., & Bailey, T. M. (2014). *Projections of education statistics to 2022: Forty first edition* (Report No. 2014-051). Washington, DC: National Center for Education Statistics, Institute of Education Sciences. Retrieved from http://nces.ed.gov/pubs2014/2014051.pdf

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.

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* (Research Memorandum No. RM-14-09). Princeton, NJ: Educational Testing Service.

Kannan, P. (2016). *Vertical articulation of cut-scores across the 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. https://doi.org/10.1002/ets2.12115

Kobrin, J. (2007). *Determining SAT benchmarks for college readiness* (College Board Research Note No. RN-30). New York, NY: The College Board. Retrieved from ERIC database. (ED562605)

Kolen, M. J. (2011). *Issues associated with vertical scaling for PARCC assessments* (White paper). Retrieved from http://www.parcconline.org/files/40/Technical%20Advisory%20Committee/43/Vertical-Scales-Kolen.pdf

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Massachusetts Department of Elementary and Secondary Education. (2012). 2012 MCAS and MCAS-Alt technical report. Retrieved from http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2012_Tech/2011-12%20MCAS%20Tech%20Rep.pdf

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*, 78–88.

Michigan Department of Education. (2011). Appendix E: New developed cut scores. In *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.* Retrieved from http://www.michigan.gov/documents/mde/Appendix_E_-_Independent_Quality_Assurance_Review_394630_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.

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

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. (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.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

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 Perspective, 4*, 1–98.

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. http://doi.org/10.1002/j.2333-8504.2013.tb02316.x

Suggested citation:

Kannan, P., & Sgammato, A. (2017). *Articulation of cut-scores in the context of the next-generation assessments* (Research Report No. RR-17-34). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12162

Action Editor: James Carlson

Reviewers: Richard Tannenbaum and Caroline Wylie