Consistency of Standard Setting in an Augmented State Testing System

Robert W. Lissitz and Hua Wei, *University of Maryland*

In this article we address the issue of consistency in standard setting in the context of an augmented state testing program. Information gained from the external NRT scores is used to help make an informed decision on the determination of cut scores on the state test. The consistency of cut scores on the CRT across grades is maintained by forcing a consistency model based on the NRT scores and translating that information back to the CRT scores. The inconsistency of standards and the application of this model are illustrated using data from the Maryland MSA large state testing program involving cut points for basic, proficient and advanced in mathematics and reading across years and across grades. The model is discussed in some detail and shown to be a promising approach, although not without assumptions that must be made and issues that might be raised.

Keywords: standard setting, consistency, augmented testing

The No Child Left Behind (NCLB) Act (2001) requires that all schools and districts be reviewed annually on the basis of the percentage of students who perform at or above the state-defined proficient level and improvement from the previous year in terms of this percentage. Students identified as below proficient are provided with additional assistance or enrolled in remediation programs that are designed to help them reach the required level. Schools that fail to show adequate yearly progress (AYP) for consecutive years will receive serious sanctions. Therefore, the determination of cut scores that leads to student classifications has important implications for school accountability and instructional practice.

Traditionally, standard setting is a grade-by-grade activity. Participants of standard setting are divided into a number of groups, each responsible for determining the cut scores for a particular grade of a content area. Each group is made up of policymakers, content specialists, and, most importantly, teachers who have had considerable experience with that particular grade. Each group sets the cut scores independently, although their decisions may be influenced by the across-group discussions held during the standard setting process. This practice often leads to the result that the cut scores are not set with a consistent level of rigor across grades (Ensign, MacQuarrie, & Beck, 2002), and sometimes the variation among the resulting cut scores across grades may be too big to be considered reasonable.

Inconsistency among the cut scores across grades is often reconciled by a single articulation committee, whose job is to evaluate the grade-by-grade committee recommendations and put forward a coherent and consistent system of cut scores. However, the actual results of articulation are hard to justify due to a lack of guiding policy and accepted methodology. Besides, the across-grade alignment is post hoc by nature and can hardly be generalized beyond the task at hand.

Lissitz and Huynh (2003) introduced the concept of vertical moderation in standard setting to achieve across-grade consistency. It was a combination of professional judgments and statistical adjustments. Professional judgments are based on a common set of definitions for the achievement levels across grades and a "forward-looking" perspective toward proficiency. Statistical adjustments are made either to interpolate (or extrapolate) cut scores for the grades for which standard setting is not actually conducted or to smooth out the peaks and valleys of the achievement trajectories. This approach has been put into operational use and obtained satisfactory results in a number of large-scale statewide assessments (e.g., Huynh, Barton, Meyer, Porchea, & Gallant, 2005; Buckendahl, Huynh, Siskind, & Saunders, 2005).

Consistency of vertically moderated standard setting is achieved by maintaining "a consistent across grade trend line" (Huynh & Schneider, 2005) on the percentages of students assigned to the targeted category across grades. By "consistent" they meant "no change, a moderate level of increase, and a moderate level of decrease" (p. 106). The four models of cut score consistency proposed by Lewis and Haug (2005) reflected the same concept and the four models result in an equal percentage, an approximately equal percentage, a smoothly decreasing percentage, or a smoothly increasing percentage of students categorized as proficient across grades. This concern also impacts the current interest in growth modeling.

Robert W. Lissitz is a Professor, Department of Measurement, Statistics, and Evaluation, College of Education, 1229 Benjamin Building, University of Maryland, College Park, MD 20742-1115; rlissitz@umd.edu. ffua Wei is a Graduate Assistant, College of Behavioral and Social Sciences, 1158 LeFrak Hall, University of Maryland, College Park, MD 20742-8225.

Copyright © 2008 by the National Council on Measurement in Education

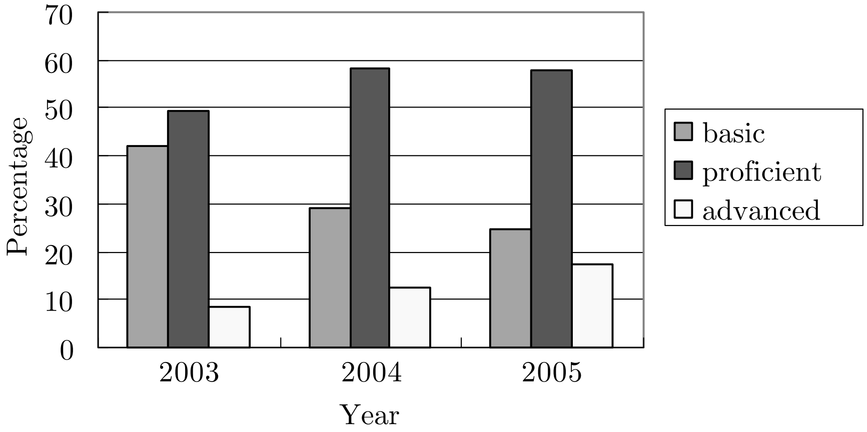

FIGURE 1. Distribution of students across performance categories for Grade 3 in Reading, 2003–2005.

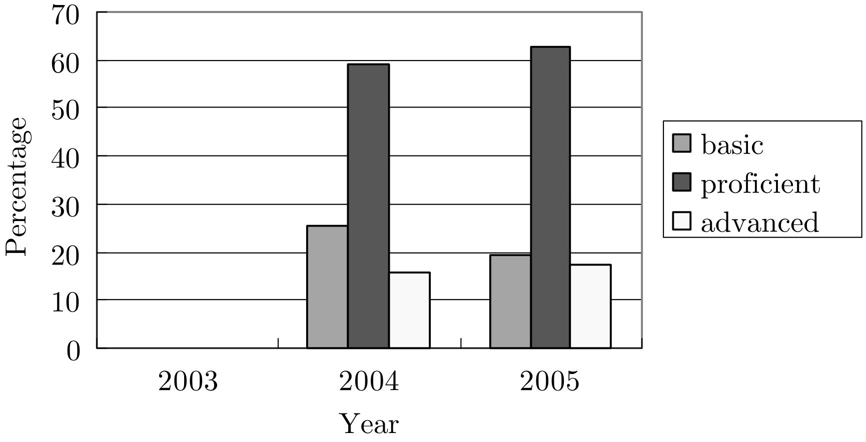

FIGURE 2. Distribution of students across performance categories for Grade 4 in Reading, 2003–2005.

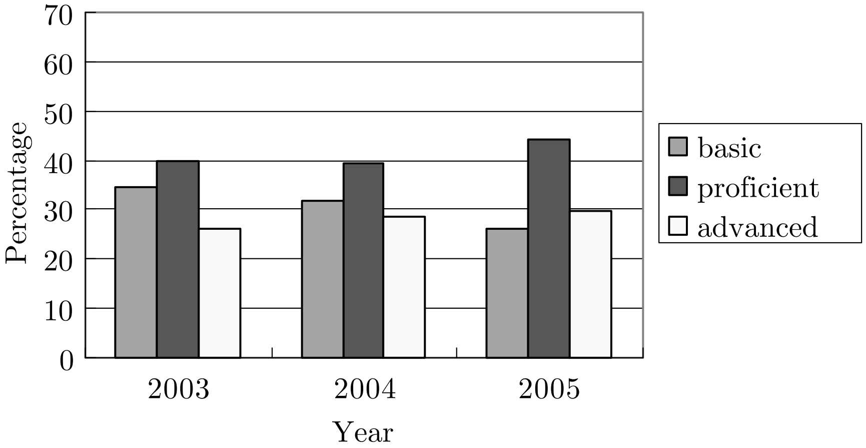

FIGURE 3. Distribution of students across performance categories for Grade 5 in Reading, 2003–2005.

Looking at changes in student performance in terms of performance level descriptors requires some sense of comparability across years and grades. In other words, if standards are not set at comparable levels across years or grades, it is unfair to draw conclusions regarding observed changes in the percentage of students meeting a standard.

Augmented state testing programs, which include a custom-made criterion-referenced test (CRT) and a commercial norm-referenced test (NRT), arose in response to the call for the use of multiple measures for making decisions about students, as specified in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The CRT assesses the state's content standards and the scores indicate how well the students have learned in the content areas of interest. The NRT covers a broader range of content and generates normative scores that show how well the students perform in comparison with other students across the nation. Suggestions have been made that scores from both tests should be used jointly, following certain predetermined rules, to improve the reliability and validity of the decisions made about individual students (Henderson-Montero, Julian, & Yen, 2003; Chester, 2003).

More importantly, a NRT is administered as part of a state testing program to gauge the validity of the state CRT. Information obtained from the NRT in the form of scale scores, percentile ranks, grade equivalents, and normal curve equivalents can be used to corroborate results from the state test. As a matter of fact, many research studies (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000; Grissmer, Flanagan, Kawata, & Williamson, 2000) have been conducted to examine the trustworthiness of state tests by comparing them with nationally acknowledged NRT and answer questions like whether scores on the high-stakes state tests accurately reflect student achievement.

Besides, the performance standards of a NRT can be used as base line to compare with those of the state test. They are considered as benchmarks because they are endorsed by a national panel of experts, unaffected by the consequences attached to the results, and

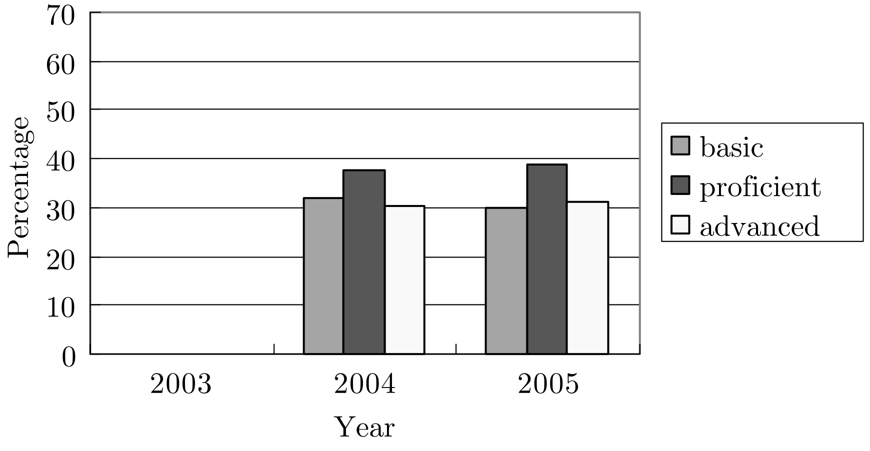

FIGURE 4. Distribution of students across performance categories for Grade 6 in Reading, 2003–2005.

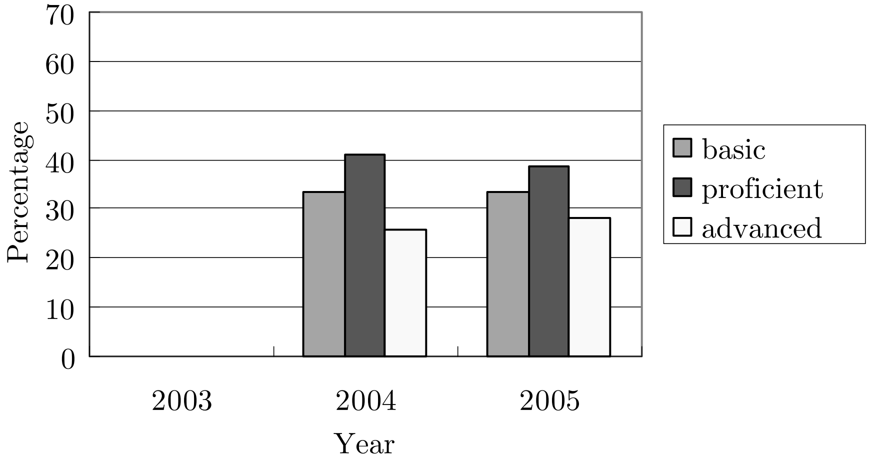

FIGURE 5. Distribution of students across performance categories for Grade 7 in Reading, 2003–2005.

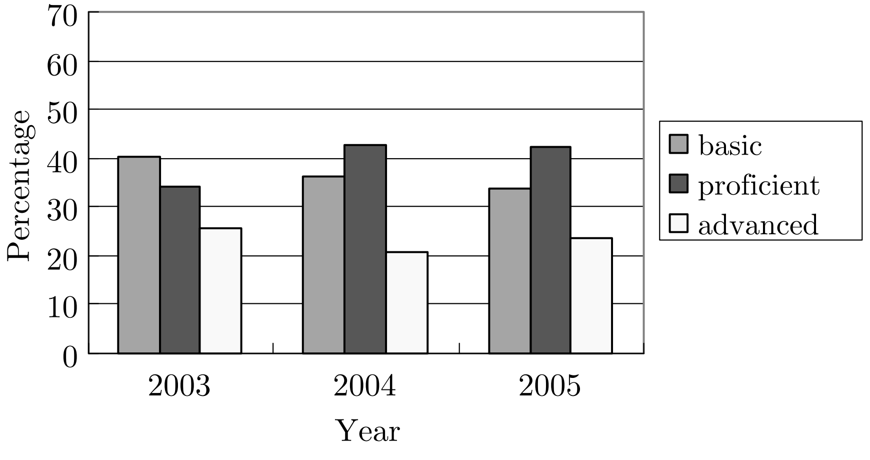

FIGURE 6. Distribution of students across performance categories for Grade 8 in Reading, 2003–2005.

based on more solid impact data that come from a more representative and stable national sample. A recent study by Schafer, Liu, and Wang (2007) compared the state-defined standards of 43 states with each other and with the standards of the National Assessment of Educational Progress (NAEP) in mathematics and reading, and discovered patterns of consistencies and inconsistencies. Although the results of the study could not be used directly to declare the validity or invalidity of the state standards, they pointed out a trend, based on which the states could make their own interpretations and take actions accordingly. This study is just one example where a national test is used to provide a base line for a state exam.

In this article, we address the issue of consistency in standard setting in the context of an augmented state testing program. Information gained from the external NRT scores is used to help make an informed decision on the determination of cut scores on the state test. Specifically, the consistency of cut scores on the CRT across grades is maintained by forcing a consistency model on the results of the NRT and then translating them back to the corresponding CRT scores. The research questions answered in this study are:

(1) Is there a pattern among the cut scores across years and grades?
(2) Is there any evidence that the cut scores for categorizing students as proficient are consistent (or inconsistent) across grades?
(3) How can the scores or associated indices from the NRT be used to help obtain consistent standards on the state test?

Data

The data examined in the study are Maryland cross-sectional student performance assessment data for Grades 3 through 8 for three consecutive years from 2003 through 2005. Two tests, a CRT and a NRT, were administered to each grade in the content areas of mathematics and reading. In 2003, mathematics and reading assessments were administered in Grades 3, 5, and 8. In 2004 and 2005, all the grades from Grade 3 through 8 were assessed on both subject areas. CTB/McGraw-Hill and Harcourt Assessment were responsible for the mathematics and reading assessments, respectively. The NRT scores were computed using TerraNova or Stanford 10 items, and the

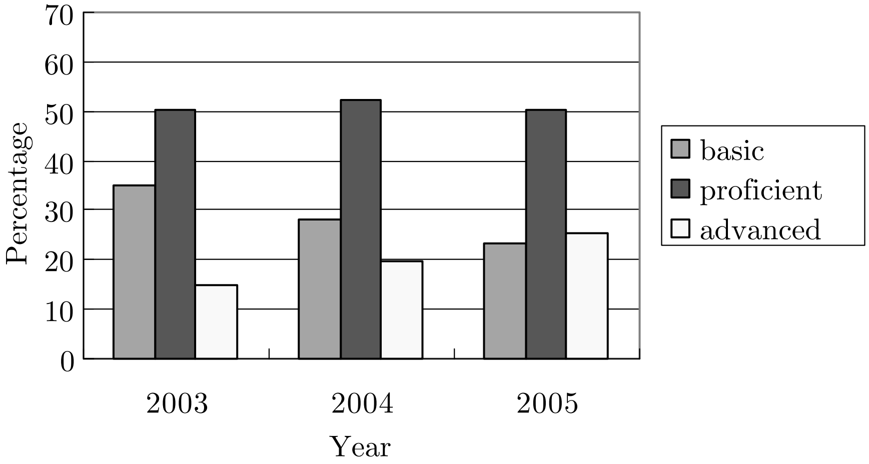

FIGURE 7. Distribution of students across performance categories for Grade 3 in Math, 2003–2005.

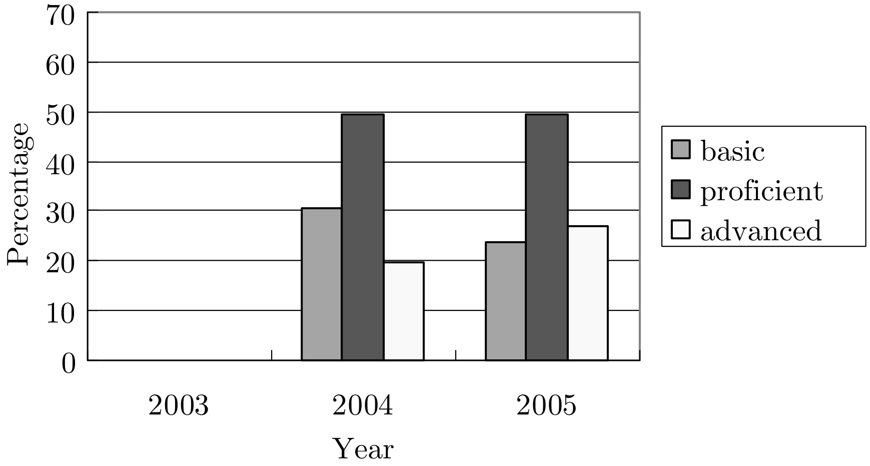

FIGURE 8. Distribution of students across performance categories for Grade 4 in Math, 2003–2005.

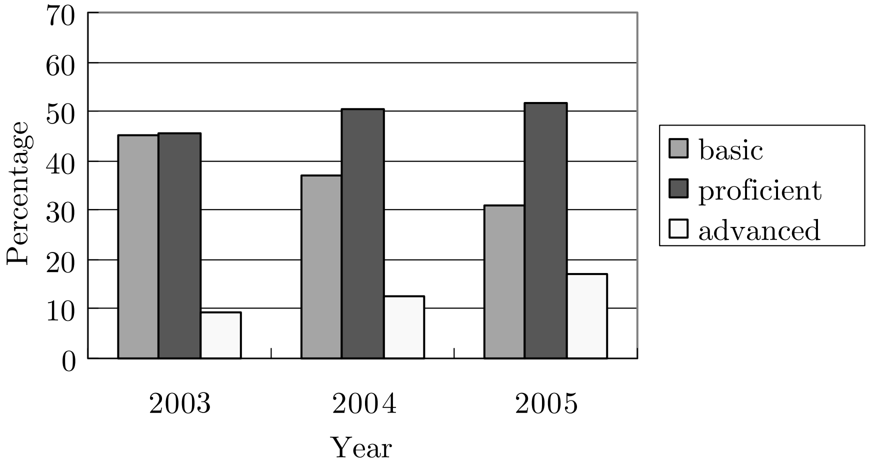

FIGURE 9. Distribution of students across performance categories for Grade 5 in Math, 2003–2005.

CRT scores were calculated using the customized items written to the Maryland content standards (Maryland Voluntary State Curriculum) plus a subset of TerraNova or Stanford 10 items that were aligned with the state content standards.

Standard setting was conducted for Grades 3, 5, and 8 in 2003, and for Grades 4, 6, and 7 in 2004. The Bookmark method was applied in both years. The cut scores obtained from the two sessions of standard setting were used to assign students to three performance levels: basic, proficient, and advanced. Information about the standard setting procedure and the resulting cut scores can be found in the Maryland Standard Setting Technical Report (Maryland State Department of Education, 2003a).

Methodology

In the Maryland School Assessment (MSA) system, the CRT and the NRT are designed to measure the same general constructs but intended for different purposes. Take MSA-Math as an example. Despite the fact that the CRT items are selected to assess learning on the Maryland content standards and the NRT items are written to measure concepts, processes, and skills taught throughout the nation, the two tests have substantial content overlap. Naturally, this was intentional and planned into the MSA testing. Moreover, more than one-third of the items in the NRT contribute to the CRT scores and those NRT items that are included in the CRT are aligned with the state content standards (Maryland State Department of Education, 2003a, 2003b). As a matter of fact, in 2004 the correlations between the NRT and CRT scores ranged from .80 to .85 in mathematics, and from .89 to .90 in reading (Maryland State Department of Education, 2004), and they are of the same magnitude across years. Therefore, accuracy and appropriateness of the high-stakes decisions on performance standards and cut scores of the CRT can be enhanced by taking advantage of its relationship with the NRT.

In this study, degree of consistency in the CRT cut scores for a particular performance category across grades is evaluated by "translating" the cut score for each grade into a NRT national percentile rank and then comparing the percentile ranks across grades. "Translating" is done by selecting those

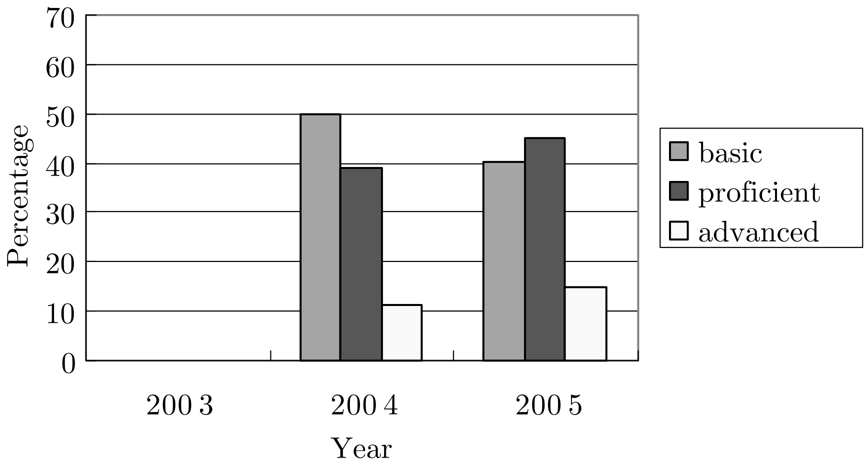

FIGURE 10. Distribution of students across performance categories for Grade 6 in Math, 2003–2005.

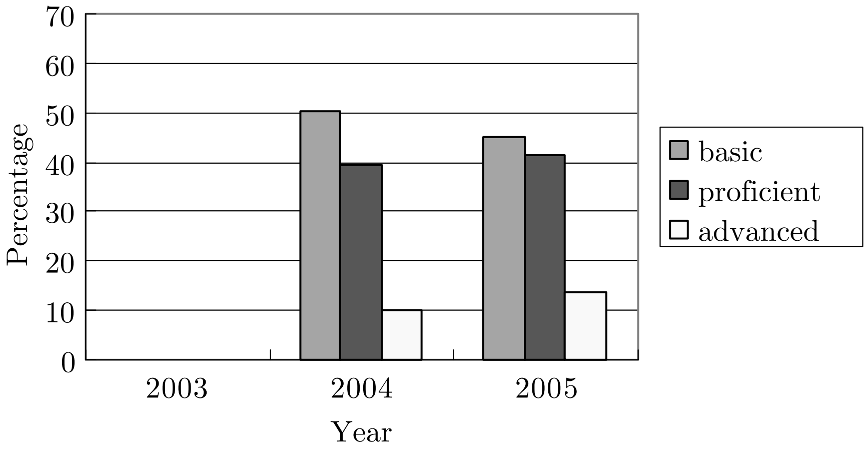

FIGURE 11. Distribution of students across performance categories for Grade 7 in Math, 2003–2005.

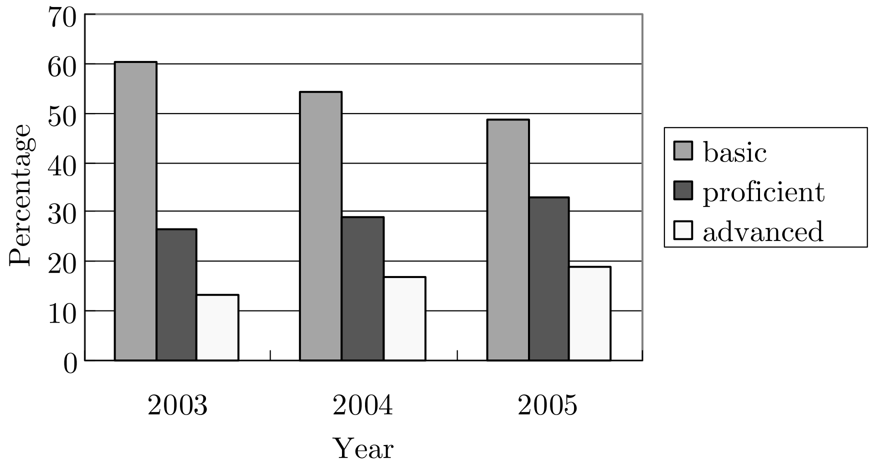

FIGURE 12. Distribution of students across performance categories for Grade 8 in Math, 2003–2005.

students whose CRT scores are equal (or close, in some cases) to the targeted cut score, looking at their NRT scores and associated national percentile ranks (NPRs), and finding out the median NPR. If the NPRs associated with the cut scores are following the same pattern (i.e., smoothly increasing, smoothly decreasing, or remaining constant across grades), we have some evidence to say that the cut scores are consistent across grades. Following the same line of thought, a consistent system of cut scores can be established for consecutive grades if we set the cut scores by forcing their corresponding NRT NPRs to be consistent. The meaning for "consistent" can be specified by any of the consistency models discussed above or could even be uniquely defined by a particular state. We can have the same fixed percentile rank for each grade or we can have the percentile ranks increasing or decreasing continuously for the consecutive grades to reflect state policies regarding standards.

The Maryland standard-setting committees had an interest in gradually increasing the expectation for achieving the minimal proficient level of performance across grades when they set the standards in 2003 (Personal communication with staff at Maryland State Department of Education). The committees, as far as we have been able to determine, did not use any systematic approach to achieve this outcome. The method proposed in this study serves the function of providing a mechanism to steer the committee decisions in the direction that they have declared an interest in going. In other words, in addition to state impact data, the standard committee could be given consistency data defined in the way we suggest in this article.

Results and Discussion

Distributions of Performance Categories Across Years

An analysis of the CRT results yields basic statistics about the distribution of performance categories for each grade across the 3 years. Two sets of graphs, from Figures 1–12 as shown, display the percentage of students in each performance category for each grade across years in the two content areas, and they provide an answer to the first research question with regard to the pattern of the cut scores across years and grades. A pattern that is observed in

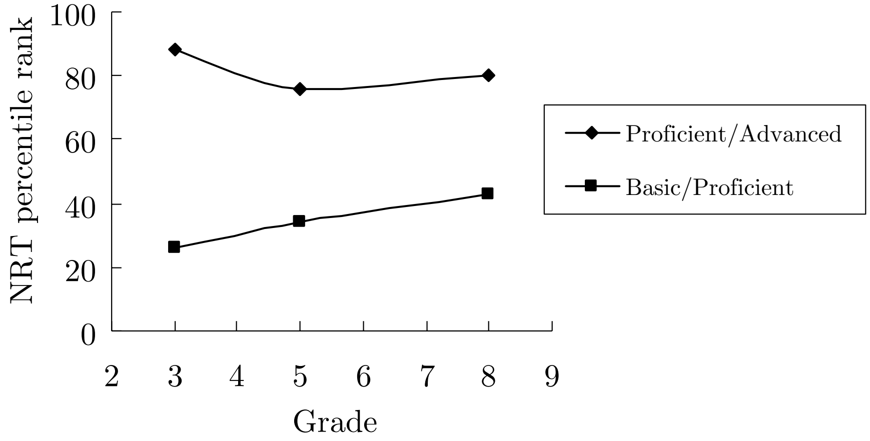

FIGURE 13. NRT NPRs associated with CRT cut scores for Reading in 2003.

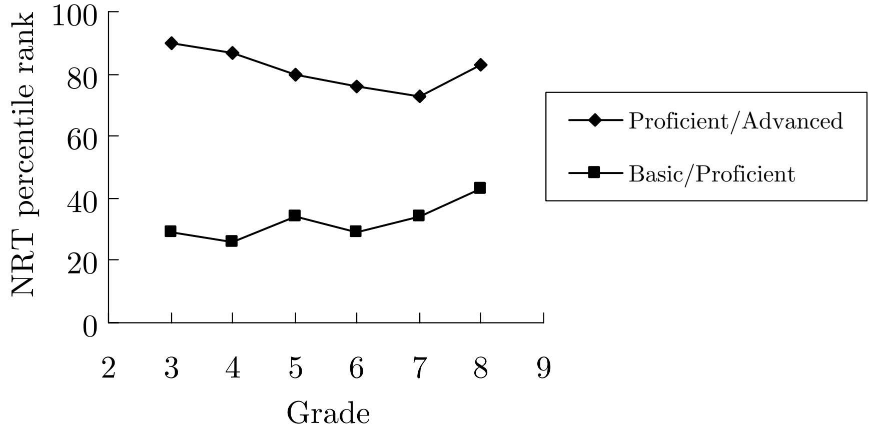

FIGURE 14. NRT NPRs associated with CRT cut scores for Reading in 2004.

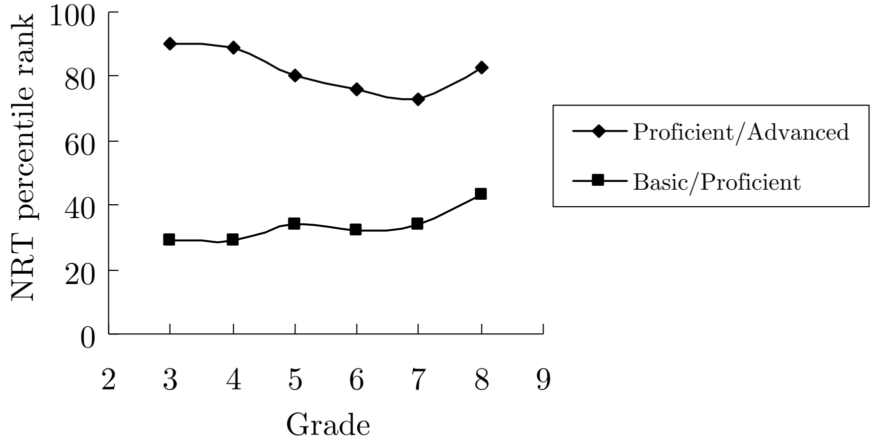

FIGURE 15. NRT NPRs associated with CRT cut scores for Reading in 2005.

every graph is that the percentage of students in the basic category was decreasing across years as they, presumably, moved to the proficient category, and the percentage of students in the higher two categories (proficient and advanced) was increasing. This finding is encouraging. It is our understanding that the CRT scores for each grade have been equated across years (i.e., horizontal equating), and in the 2004 MSA Technical Report for Math (Maryland State Department of Education, 2004), there was an explicit indication of the equating procedure. Therefore, assuming that the CRT scores for the same grade are comparable across years, this finding implies that new students perform better than their predecessors. This is exactly what a school system and the state would like to see.

Comparison of Cut Scores in Terms of NRT NPRs

Figures 13–18 display the relationship between the CRT cut scores and the NRT NPRs in each year across grades in reading and mathematics, and they serve to answer the second research question as to whether the cut scores are consistent across grades. Students who were at the boundaries of two performance levels, that is, students whose CRT scale scores were equal to the cut score were selected and the median of their NRT NPRs was identified. For some particular year and grade combinations, cut scores did not actually occur among students. When this was the case, students who scored closest to the cut score were selected. (Please note that the data in each of the figures are cross-sectional, but we have connected the dots with lines to make them easier to understand. The lines do not imply longitudinal data.)

A pattern that is common in all the graphs is that the cut score that distinguished the basic from the proficient category was associated with a general increasing trend in terms of the NRT NPR with increasing grades for each year. This is what was desired by the standard setting committees, according to MSDE staff. It is noticeable that the cut scores for Grades 3, 5, and 8 were dissimilar from those for Grades 4, 6, and 7, probably because they were set at two time points (in 2003 and 2004, respectively), and by different committees. The cut scores set in the same year were increasingly demanding

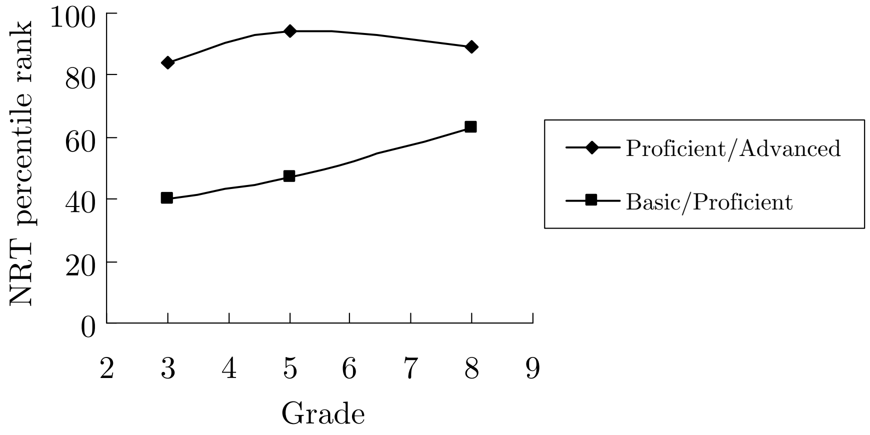

FIGURE 16. NRT NPRs associated with CRT cut scores for Math in 2003.

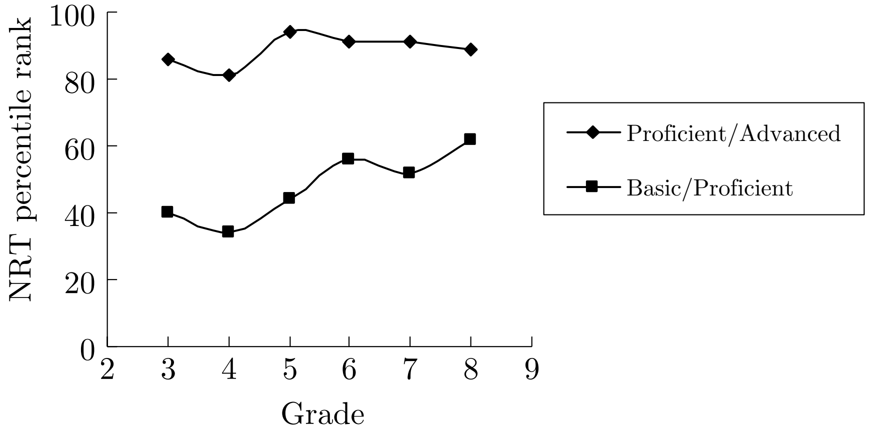

FIGURE 17. NRT NPRs associated with CRT cut scores for Math in 2004.

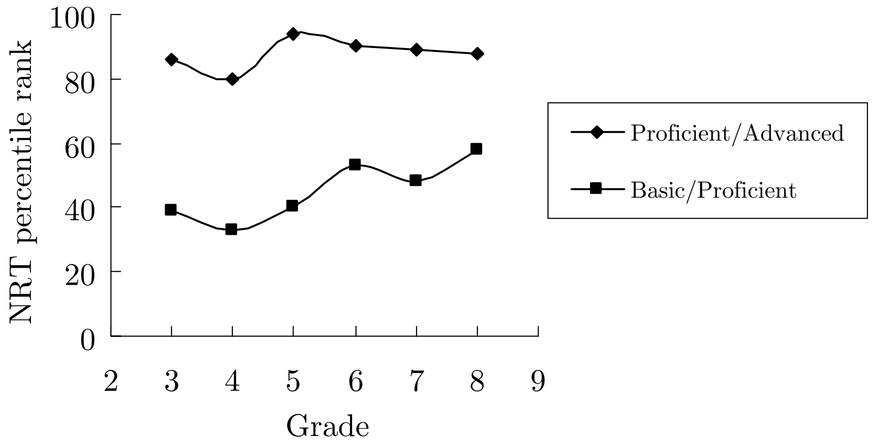

FIGURE 18. NRT NPRs associated with CRT cut scores for Math in 2005.

as the grade level increased, although the trend was less obvious over the entire span of grades.

In contrast, the NPRs corresponding to the cut scores that distinguished the advanced from the proficient category bounced around with no clear trend. A conclusion can be drawn from this finding that the cut score for the proficient category was generally more demanding for higher grades than for lower grades. In other words, students had to become better, at least compared to the national norming group, in order to remain in the proficient category as they progressed through school. If they were not, they would drop down into the basic category in subsequent years.

Adjustment of Cut Scores

Tables 1 and 2 show the original cut scores that were determined for Grades 3, 5, and 8 in 2003, and the median NPRs associated with the students who scored at or near the cut scores. Looking at the NPRs across grades in each year, we see evidence that the cut scores are increasingly demanding as the grade level increases, as we discussed above.

Different approaches could be adopted to adjust the CRT cut scores across grades by controlling the NRT NPRs associated with the cut scores. The results of the approaches that we recommend in response to the third research question are summarized in Tables 3 through 6. In the first approach, we looked at the original CRT cut score for the proficient category in Grade 3, selected among the third graders in 2003 those who achieved that score, and identified the median of their NRT NPRs. The median NPR thus obtained was used as a baseline and translated back to a cut score for each of the other two grades (Grades 5 and 8) assessed in the year of 2003. Specifically, for each of the two grades, we selected those students who ranked at the baseline NPR on the NRT and found out their median CRT score. That score became the new cut score. The newly obtained cut scores for the three grades were then used on students of the same grades in 2004 and 2005, and the median NPR associated with the cut score for each grade in each year was identified. In this way, we allow differences in student performance in years 2004 and 2005 for all three grades to show increases or

Table 1. Median NRT Percentile Ranks Obtained by Students who Scored at the CRT Cut Scores for Grades 3, 5, and 8 in Reading

		Median NPR		
Grade	Cut Score	2003	2004	2005
3	388	26	29	29
		(*n* = 675)	(*n* = 2,103)	(*n* = 1,739)
5	384	34	34	34
		(*n* = 644)	(*n* = 2,238)	(*n* = 1,161)
8	391	43	43	43
		(*n* = 375)	(*n* = 1,511)	(*n* = 1,405)

Table 2. Median NRT Percentile Ranks Obtained by Students who Scored at the CRT Cut Scores for Grades 3, 5, and 8 in Math

		Median NPR		
Grade	Cut Score	2003	2004	2005
3	379	40	40	39
		(*n* = 540)	(*n* = 511)	(*n* = 424)
5	392	47	44	40
		(*n* = 611)	(*n* = 536)	(*n* = 531)
8	407	63	62	58
		(*n* = 622)	(*n* = 621)	(*n* = 654)

Table 3. Results of Statistical Adjustment in Reading—Approach One

		Median NPR		
Grade	Cut Score	2003	2004	2005
3	388	26	29	29
		(*n* = 1,454)	(*n* = 2,103)	(*n* = 1,739)
5	378	26	31	29
		(*n* = 1,877)	(*n* = 2,188)	(*n* = 1,029)
8	374	26	28	28
		(*n* = 1,712)	(*n* = 931)	(*n* = 1,770)

Table 4. Results of Statistical Adjustment in Math—Approach One

		Median NPR		
Grade	Cut Score	2003	2004	2005
3	384	40	45	45
		(*n* = 632)	(*n* = 489)	(*n* = 441)
5	386	40	41	35
		(*n* = 520)	(*n* = 505)	(*n* = 485)
8	378	40	43	35
		(*n* = 586)	(*n* = 499)	(*n* = 471)

decreases relative to the baseline NPR of Grade 3 in 2003.

Tables 3 and 4 report the new cut scores and their corresponding NPRs in each year in the two content areas. It is noticeable that the resulting median NPRs associated with students who performed at or near the cut scores for the three grades in 2004 and 2005 were almost the same in reading, but they bounced around in mathematics, especially in the year of 2005.

Alternatively, in our second approach, we used the average empirical NPR of Grades 3, 5, and 8 in 2003 as the baseline and translated it into different scale scores for the three grades. Specifically, for each of the three grades assessed in 2003, we selected those students who were right at the cut point, and retained their median NPR on the NRT. The average of the three NPRs for the three grades was calculated and used as the baseline. Clearly, we could have used the median of the three NPRs, but in our example the mean and the median were identical. The average NPR thus obtained was translated back to a CRT score for each of the three grades in 2003. These three scores became the new cut scores for the three grades. They were used on students of the same grades in 2004 and 2005, and the median NPR associated with the cut score for each grade in each year was identified. Again, this allows differences in performance to emerge in 2004 and 2005 relative to the average of the performance fixed in 2003 as the baseline. The new cut scores and their corresponding NPRs are summarized in Tables 5 and 6. In the content area of reading, the resulting median NPRs bounced around in 2004, but remained almost constant in 2005. In mathematics, the resulting median NPRs showed more differences across grades in 2005 than in 2004.

The results of the adjustments are satisfactory in the subject area of reading. The median NPRs associated with students at or near the adjusted cut scores are similar across grades and across years. The fluctuation in the NPRs is within reasonable limits and can be partially accounted for by the error of estimation associated with the test. Consider the fact that for some grade and year combinations, the cut score was not obtained by any of the students. In that case, we chose a score closest to the cut score and found the median NPR associated with students having that score. Therefore, the

Table 5. Results of Statistical Adjustment in Reading—Approach Two

		Median NPR		
Grade	Cut Score	2003	2004	2005
3	398	34 (*n* = 1,477)	34 (*n* = 2,435)	37 (*n* = 1,997)
5	386	34 (*n* = 1,997)	38 (*n* = 1,233)	34 (*n* = 1,187)
8	381	34 (*n* = 278)	33 (*n* = 1,125)	33 (*n* = 2,127)

Table 6. Results of Statistical Adjustment in Math—Approach Two

		Median NPR		
Grade	Cut Score	2003	2004	2005
3	392	50 (*n* = 690)	51 (*n* = 558)	49 (*n* = 482)
5	396	50 (*n* = 564)	51 (*n* = 593)	44 (*n* = 529)
8	389	50 (*n* = 685)	48 (*n* = 527)	44 (*n* = 606)

median NPR we obtained was not exactly corresponding to the cut score, but to a score one or two points above or below the cut score.

The fluctuation of the NPRs is larger in mathematics, especially in the year of 2005. The adjusted cut scores are more demanding for lower grades than for higher grades, because the associated NPRs go down consistently as the grade level increases. It reflects the fact that students categorized as proficient must maintain a higher ranking compared with the national norming group as they progress through school. This is consistent with the results we have presented before. The difference in this analysis is that we have adjusted the cut scores in the later years to be consistent relative to the performance expectation obtained at the base year of 2003. Differences in performance that are due to changes in the standard have been eliminated or lessened at least.

Examining the impact data based on the CRT scores could help supplement the information based on the NRT scores when considering the revision of the cut scores to obtain consistency. Patterns residing in the empirical data indicate levels of demand of the cut score for different grades and could inform another round of revision.

Conclusions and Recommendations

How to examine and set consistent cut scores across grades is the theme of this study. We have identified two approaches to setting consistent cut scores. One is to influence the judgments that the standard setting committees render using empirical data related to the performance levels. The other is to statistically adjust the results of independently working standard setting committees after they have completed their tasks and data have been obtained from the testing. In the case of augmented testing, the first approach might be accomplished by presenting the NRT percentile ranks at the cut point, as we illustrated above. In the typical standard setting approach, the committee is presented with information on the percentage of students in the school system (usually their specific state) that would be declared proficient or below proficient based on the selection of the cut point. How this use of a national perspective would influence the work of the state committee and how to control that influence is unknown at this time and is believed to be a worthwhile direction for research in the study of standard setting.

The second approach of statistically adjusting standard setting to ensure various options for consistency is also an interesting problem and worth further exploration. We have been trying to create a system to accomplish this. Different approaches could be adopted to adjust the CRT cut scores, as we have shown above. The method we are currently favoring is to use the NRT percentile ranks in an augmented testing environment to provide a mechanism to statistically adjust the standard setting decision in the direction that the state chooses. Basically, the following sequence, as illustrated above, could be utilized to accomplish this task:

1. Decide on policy regarding the type of consistency we want to achieve in terms of a normative standard (NRT national percentile rank), that is, do we want the standards to be constant, smoothly increasing, or smoothly decreasing across grades? If we want higher standards for higher grades, what is the level of growth we expect from one grade to the next? If we want lower standards for higher grades, what is the level of decrease that we expect from one grade to the next? Do we want mathematics to have the same standard that is held for reading?
2. Meet with the standard setting committee to explain the policy that was determined in Step 1. Before they set the standards, the committee needs to know that we may adjust their results. One option is to use the standard for a specific grade (perhaps Grade 3) as a base year standard and have consistency defined relative to that grade. If this were done, the Grade 3 (or other) standard setting committee could be the only committee that sets standards. Standards for the other grades would be the result of statistical analysis operationalizing a policy decision regarding consistency.

3a. If the policy from Step 1 says that the standards should be constant across grades, convert the standard from Step 2 to a scale score for each grade. For example, if the standard we set in Step 2 has a NPR of 25 and we fix it for each grade, what we need to do is:

(1) For each grade, identify all the students whose percentile ranks are equal to 25 on the NRT.

(2) Compute the mean/median CRT score for those students. That is

the cut score for that particular grade associated with the chosen NPR.

It is very likely that the same NRT normative standard is associated with a different CRT scale score in each grade, unless considerable care was given to equating and standardizing the scales across grades with national data. In most cases this will not have been done, and the relationship between the NRT and CRT will vary to some extent from grade to grade.

3b. If the policy from Step 1 says that the standards should be smoothly increasing or decreasing, adjust the standard set in Step 2 for each of the other grades to reflect that policy.

For example, if the standard we set in Step 2 is an NPR of 25 and we want the standards to be increasing by an increment of 3 percentiles, then:

(1) For Grade 3, identify all the students whose NPR is equal to 25 on the NRT.

(2) Compute the mean/median CRT score for those students. That is the cut score for Grade 3.

(3) For Grade 4, identify all the students whose NPR is equal to 28 on the NRT.

(4) Compute the mean/median CRT score for those students. That is the cut score for Grade 4. The same thing can be done for the other grades.

4. Apply the cut scores determined in Step 3 to other subject areas to achieve consistency across subject areas.

It should be noted that statistically adjusting the CRT cut scores on the basis of the NRT scores and associated measures requires several assumptions, of course, including the idea that the norming groups for different grades are equivalent or adjusted to be so and that the norming was done well enough to serve as a source for a standard. The most important assumptions have to do with the vertical equating of the tests across years and the specification of the scale for each grade, year and subject matter and their comparability. We have assumed that the tests are measuring comparable qualities across grades and doing so using a scale score system that is directly comparable. This assumption suggests that the procedure is likely to be limited to the typical NCLB 3rd to 8th grade testing. Testing across diverse high school subject matter would not make sense, for example, ensuring consistency of performance from biology to social studies would not satisfy these assumptions. The robustness of the adjustment procedure to these assumptions could be investigated in subsequent research.

In an augmented testing program, measures provided by the national NRT can be used to inform the standard setting for the state or local CRT, assuming that the two tests are assessing the same general construct. In this article, a standard setting design is outlined that uses the NRT results as impact data to influence the judgments of the standard setting committee. Statistically adjusting the cut scores by using information from the NRT is another approach to fostering consistency across grades, and it is illustrated with real data.

In a coherent educational assessment system, one can argue that not only should the standards across grades of a content area be consistent, but the standards across content areas should also be consistent. The approaches discussed in this article can be applied to obtain consistent standards across subject areas. Additional assumptions need to be satisfied that the content standards are of equivalent difficulty across subjects and that the educational resources are equally allocated so that it is reasonable to maintain consistent standards across grades.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful advice. A portion of this article was presented at the International Conference on Educational Evaluation, November 11–12, 2006, National Taiwan Normal University, Taipei. We would like to thank the Maryland State Department of Education for their financial support of this project and Mark Moody and Marty Kehe for their advice. An earlier version of this paper was presented at the annual meeting of the National Council on Measurement in Education, April 10, 2007, Chicago.

References

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.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice*, *22*, 32–41.

Ensign, G., MacQuarrie, D., & Beck, M. (2002, June). *Validation of state achievement performance standards*. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Palm Desert, CA.

Grissmer, D. W., Flanagan, A., Kawata, J. H., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND, MR-924-EDU.

Henderson-Montero, D., Julian, M. W., & Yen, W. M. (2003). Multiple measures: Alternative design and analysis models. *Educational Measurement: Issues and Practice*, *22*, 7–12.

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

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.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, *8*(49). Retrieved August 8, 2007 from http://epaa.asu.edu/epaa/v8n49/.

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.

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). Retrieved August 8, 2007 from http://pareonline.net/getvn.asp?v=8&n=10.

Maryland State Department of Education (2003a). *Maryland Standard Setting Technical Report*. Retrieved August 8, 2007 from http://www.marylandpublicschools.org/MSDE/divisions/planningresultstest/Maryland±Standard±Setting±Technical±Reports.htm.

Maryland State Department of Education (2003b). *2003 MSA Technical Report: Math Grades 3 Through 8 and Reading Grade 10*. Retrieved August 8, 2007 from http://www.marylandpublicschools.org/NR/

rdonlyres/97925205-7AC6-452C-BE37-BFD495C2264F/3040/MSA_Tech_CTB.pdf.

Maryland State Department of Education (2004). *2004 MSA Technical Report: Math Grades 3 Through 8 and Reading Grade 10*. Retrieved August 8, 2007 from http://www.marylandpublicschools.org/NR/rdonlyres/97925205-7AC6-452C-BE37-BFD495C2264F/6502/FinalTechnicalReportMSASpring200461505.pdf.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425.

Schafer, W., Liu, M., & Wang, H. (2007). Content and grade trends in state assessments and NAEP. *Practical Assessment, Research & Evaluation*, *12*(9). Retrieved April 4, 2008 from http://pareonline.net/pdf/vl2n9.pdf.

Copyright of Educational Measurement: Issues & Practice is the property of Blackwell Publishing Limited and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.