2-6-2006

Education Policy Analysis Archives 14/04

Arizona State University

University of South Florida

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Sherman Dorn
College of Education
University of South Florida

Volume 14 Number 4 February 6, 2006 ISSN 1068–2341

The Legend of the Large MCAS Gains of 2000–2001

Gregory Camilli
Sadako Vargas
Rutgers, The State University of New Jersey

Citation: Camilli, G. & Vargas, S. (2006). The legend of the large MCAS gains of 2000–2001. *Education Policy Analysis Archives, 14*(4). Retrieved [date] from http://epaa.asu.edu/epaa/v14n4/.

Abstract
Issues related to student, teacher, and school accountability have been at the forefront of current educational policy initiatives. Recently, the state of Massachusetts has become a focal point in debate regarding the efficacy of high-stakes accountability models based on an ostensibly large gain at 10th grade. This paper uses an IRT method for evaluating the validity of 10th grade performance gains from 2000 to 2001 on the Massachusetts Comprehensive Assessment System (MCAS) tests in English Language Arts (ELA) and mathematics. We conclude that a moderate gain was obtained in ELA and a small gain in mathematics.
Keywords: MCAS; Performance Standards; Item Response Theory; Validating Score Gains; Accountability.

Introduction

Apparent achievement gains on high-stakes tests have been received with mixed reviews. Some researchers hold a positive view of the potential of high-stakes graduation tests while others remain unconvinced of the positive impact of high-stakes designations. For example, Mehrens and Cizek (2001) argued that "increases in scores most often represent real improvement with respect to the domain the tests sample" (p. 481). Koretz, Linn, Dunbar & Shepard (1991), on the other hand,

compared test results on an existing third-grade test to a newly introduced high-stakes test. They found that average student scores rose for the ensuing four years on forms of the new test. However, in the sixth year students showed essentially no growth when reassessed with the original test. More recently, a number of reports have provided widely different and inconsistent conclusions regarding student achievement in regard to the No Child Left Behind law (Braun, 2004; Education Trust, 2004; Fuller, 2004; Webb & Kane, 2004).

Resolving inconsistencies in results from high-stakes accountability efforts is a key challenge in educational research because tests have substantial consequences for test-takers, parents, and educators. It is important to know how well the accountability models (and associated high-stakes testing) are working in order to support the claim that they are legitimate tools for driving educational reform. In this paper, we argue that measurement data used to analyze trends in test scores must also be scrutinized because such data lie at the very core of claims of efficacy or invalidity. Increases in test scores may have various causes, and to facilitate analyses for sorting through competing hypotheses, we must be sure that the methods by which tests are created, scored, and maintained do not lead to false impressions of achievement trends.

In 2001, students taking the 10th grade test of the Massachusetts Comprehensive Assessment System (MCAS) obtained a large gain over students of the previous year's administration. This gain was consequently taken as strong evidence by some observers of the efficacy of high-stakes accountability based on test results. In this paper, we examine the validity of the 2000–2001 gain. As we show below, rather than large gains, there was a moderate gain in English Language Arts (ELA) and a small gain in 10th grade mathematics. We conclude the paper with a discussion of why the consequences of the errors in estimating student gains in Massachusetts have resulted in a mixed blessing.

Policy Context

The 2000–2001 MCAS gain in Massachusetts is particularly important because it has been taken as *prima facie* evidence of the efficacy of high stakes consequences. In the spring of 2001, 81% of high school sophomores in Massachusetts passed the ELA subject area of the MCAS examination. Similarly, 75% of high school sophomores passed the Mathematics subject area of the MCAS examination. The ELA and Mathematics pass rates for 10th graders in 2001 suggest sizable performance improvements from the previous year's assessment in which pass rates were 66% and 55%, respectively. Here, the pass rate is the percentage of students in the performance categories *Advanced*, *Proficient*, and *Needs Improvement*, that is, the percentage of students who did not fall in the fourth and final category, which is labeled and *Failing* or *Warning*. A more complete set of these statistics is given in Table 1.

Table 1
Percentage of Tenth Grade Students Passing MCAS ELA and Mathematics Assessments, 1998–2004

Subject	1998	1999	2000	2001	2002	2003	2004
Mathematics	58 (24)	57 (24)	55 (33)	75 (45)	75 (44)	79 (51)	85 (57)
English	72 (38)	68 (34)	66 (36)	81 (50)	86 (59)	89 (61)	90 (62)

Passing means above the *Failing* category. Combined percentage for *Proficient* and *Advanced* is in parentheses for each year and subject.

The corresponding percentages of students achieving at the *Advanced* and *Proficient* levels are shown in parentheses in Table 1. It can be noted that there is relatively no trend for the years 1998–2000 and then a sharp jump in 2001 when passing these two tests *became a requirement for graduation*. Increases were very consistent across Racial/Ethnic groups, and were fairly consistent across high schools. According to Michael Russell and Laura O'Dwyer from Boston College (personal communication), the large increases in both 10th grade ELA and Mathematics scores have been attributed by the Massachusetts Department of Education to an increase in student motivation in preparation and performance as well as to improvements in the quality of instruction.

The 2000–2001 increase has been received with warm enthusiasm by policy makers. Finn (2002) viewed the 2000–2001 gain as a signal that testing, as a component of accountability, functions to increase student learning:

> On Monday, Department of Education officials released the results of the spring 2001 MCAS exams which showed that 82% of 10th graders passed the English test and 75% passed the math test—increases of 16% and 20%, respectively, from the previous year. The results—which would be good news at any time—are all the more pleasing because high school students must now pass these sections of the MCAS to graduate…. *Now that it has teeth* [emphasis added], the MCAS is even better poised to promote reform and boost student achievement.

Likewise, Cizek (2003) touted the MCAS testing program as providing "strong evidence of positive consequences" of high stakes testing (p. 42). Regarding positive consequences of testing in Massachusetts, he argued that the glass is more than half full:

> The combination of real increases in learning, gap decrease, student motivation, and drop-out prevention makes this result the equivalent of the big E on a Snellen chart. It seems possible to discount this as evidence that tangible positive consequences are actually accruing in the context of high-stakes testing only if one is not even looking at the glass. (p. 43)

Others have argued that the test scores must be further examined, but have accepted the basic validity of MCAS gains. Gaudet reported that even with the large overall gain, in 2002 only about 70% of students in the poorest communities passed the MCAS after retesting:

> Looking at the overall numbers, however, gives us no information about the specific challenges that face us. While it is encouraging that the MCAS pass rate increased dramatically between 2000 and 2001, there are still many students who have not yet mastered the basic skills needed to live and work in contemporary Massachusetts. (Gaudet, 2002, p. 2)

It seems clear that the MCAS has become the central and unquestioned means of describing the success of educational practices in Massachusetts, and continues to receive intense media focus.

MCAS results for 2000 and 2001

We used a method based on item response theory (IRT) for evaluating the 2000–2001 change in test performance on the 10th grade MCAS. In particular we examined the large gain based percentage of students who passed the test, i.e., had scores above a particular criterion score. The methods we used directly estimated the distribution of scores assuming normality. Technical details concerning the distributional methods are given in the Appendix. Here we only note that the basic approach has several advantages: it is unaffected by rounding, it accounts for measurement error, it

avoids complexities due to inestimable proficiencies for individual students, and it is broadly applicable to programs using IRT technologies.

Statistical descriptions of MCAS 2000 and 2001[1]

Classical and IRT statistics are given in Tables 2 and 3 for operational 10th grade MCAS items in 2000 and 2001. Also included are the cut scores for the *Warning* (*Failing*) achievement band. The *p*-values indicate success rates on both multiple choice (MC) and open response (OR) items for ELA were higher in 2001. The more interesting statistics concern the relative difficulties of the test items in 2000 and 2001, which can be determined by examining the IRT *b*-values which are estimated in a way that is comparable across years by means of test equating (fixed common item parameter or FCIP equating is used). The average IRT *b*-values indicate that the difficulty of the MC items stayed about the same, and the difficulty of the OR items decreased dramatically. The *b*-values roughly follow the scale of *z* scores (with a mean of 0.0 and standard deviation of 1.0), and lower *b*s indicate *easier* items. We refer to the scale of the *b*s as the "logit" scale—where the term logit is derived from the IRT "logistic" item model. In IRT, estimates of student ability (also termed proficiency), labeled θ, also have this logit scale.

Based on average *b*-values, the test became easier in 2001. But in Table 2, it can be seen that the cut score (39.0) in 2001 was two points *lower* than the cut score for 2000 (41.0). This is an unusual finding, even with one less MC item in 2001. Under normal circumstances, if a test gets easier, the cut point would be expected to go up to remain consistent with the previous year's test. The fact that the cut point dropped is curious.

Table 2
Key Test Characteristics, MCAS English Languages Arts Test, 2000 and 2001

		2000		2001	
Item Type	Items	*p*-value	Average *b*	*p*-value	Average *b*
Multiple choice* (MC)	36	.70	-0.74	.74	-0.74
Open response (OR)	6	.54	-0.40	.61	-1.85
Raw score cut point		41		39	

* One MC item was dropped in 2001.
Source: Massachusetts Department of Education, 2002a, 2002b.

In Table 3, it can be seen that the *p*-values indicate that for Mathematics, success rates on multiple choice (MC), short answer (SA), and open response (OR) items were also slightly higher in 2001 than 2000. The average IRT *b*-values indicate that the average difficulty for the 38 MC and OR items dropped, though the difficulty of 4 SA items increased slightly. Thus, the Mathematics test also became substantially easier, yet it can be seen in Table 3 that the cut score (20.0) in 2001 was one point lower than the cut score for 2000 (21.0). This finding is even more curious than that for ELA.

[1] More information for examining the 2001 MCAS phenomenon can be found in the MCAS technical reports available from the Massachusetts Department of Education website.

Table 3
Key Test Characteristics, MCAS English Languages Arts Test, 2000 and 2001

		2000		2001	
Item Type	Items	*p*-value	Average *b*	*p*-value	Average *b*
Multiple choice* (MC)	32	.49	0.51	.55	0.43
Short answer (SA)	4	.43	0.27	.44	0.36
Open response (OR)	6	.35	0.63	.52	0.29
Raw score cut point		21		20	

* One MC item was dropped in 2001.
Source: Massachusetts Department of Education, 2002a, 2002b.

For both Math and ELA, OR items contribute about one-half of the total possible points on the assessment. And because the OR items will become an important part of the investigation reported in this paper, we will examine changes in scoring procedures for the OR items in the next section before returning to an analysis of the apparently problematic "Failing" cut score.

Possible MCAS 2000–2001 Scoring Changes

The 6 OR items from the MCAS Mathematics and their scoring rubrics from 2000 and 2001 are given for public examination on the Massachusetts DOE website. Each OR item was scored on a scale 0–4 and the abbreviated descriptions of the rubric description for scores of 4 and 2 are given in Tables 4–5. In Table 4, the left-hand column contains 2 numbers that the item identifiers for the years 2000 and 2001, respectively. Looking down the second column, it is evident that to obtain a score of "4" in 2000, the adjectives "correct" and "accurate" appear for each of the 6 items, while the word "correct" appears just twice in 2001. Moreover, the 2000 rubrics appear semantically denser than those in 2001, and a similar pattern exists for the score of "2." These observations are quite consistent with the 2000–2001 rubrics for ELA items scored on the 0–4 scale.

Table 4
MCAS ELA "4" Rubric Descriptors, 2000 and 2001 Tests

Released Item (2000, 2001)	2000 descriptors	2001 descriptors
13, 13	accurately using; communicate correct strategy	facility with
16, 16	accurately creating and interpreting	correctly analyzing; using & explaining correct
21, 21	correct procedures; consistent accuracy	
22, 22	determining correct; accurately solving	
41, 40	accurately applying correct	communicates
42, 41	correctly solving; developing	describing & analyzing

Table 5
Analysis of rubric descriptor differences required for a score of 2 on the scale 0–4.

Released Item (2000, 2001)	2000 descriptors	2001 descriptors
13, 13	basic understanding; accurately applying; communicating correct	partial understanding
16, 16	general understanding; applying … with some accuracy; creating an accurate	basic understanding; analyze
21, 21	some understanding; implementing correct procedures at least once	basic understanding
22, 22	basic understanding; correct strategy in solving; making one or more correct	partial understanding
41, 40	general understanding; making some accurate …; applying some correct	basic understanding; communicates
42, 41	basic understanding; using appropriate strategies to solve problems; developing	basic understanding; describing

Although one might be tempted to attribute the 2000–2001 increases in student performance to the OR items, it is also the case that the equated IRT *b* parameters, which describe item difficulty, should take this fact into account.[2] All things being equal, a simple change in item difficulty will not affect the percentage of students that reach a particular achievement level on the MCAS. As we shall show below, however, it does appear plausible that some of the change in difficulty of OR items may not be accounted for by the IRT scaling and equating process.

Methods

Data and Sample Description

A copy of the MCAS 2000 and 2001 10th grade data were obtained without student, school, or district identifiers.[3] Students who were classified as LEP (n = 846) or who had raw scores of zero were not included. This resulted in population sizes of n = 57,542 for ELA 2000, n = 61,968 for

[2] "The item calibration for the 2001 and 2000 groups was performed separately using the combined IRT models (three parameter logistic [3PL] for multiple choice items, two parameter logistic [2PL] for short answer items, and the graded response model [GRM] for open-response items). Calibration of parameter estimates in 2001 placed items on the same scale as in the 2000 calibration by fixing the parameters for the anchoring items to 2000 calibration values. It is noteworthy that at least 25 percent of the 2001–2000 equating items were also used for 2000–1999 equating, so that their parameters were actually fixed to 1999 calibration values" (MDOE, 2002a, p. 48).

[3] Data were provided by researchers at the Center for the Study of Testing, Evaluation and Educational Testing at Boston College. The data obtained were strictly anonymous. No information was present in the data file regarding student, teacher, school, or district identity.

ELA 2001, n = 59,946 for Mathematics 2000, and n = 62,900 for Mathematics 2001. These values are relatively close to the sample sizes n = 57,681, 62,620, 59,978, and 62,921 given in the 2000 and 2001 technical reports (Massachusetts Department of Education [MDOE], 2002a & 2002b) for the classical reliability statistics. Data consisted of responses for each student to the 42 ELA and Mathematics operational items in 2000 and the 41 ELA and Mathematics operational items in 2001.

Cut Score Drift

The primary source of bias investigated in this paper concerns changes in IRT scaling procedures. A series of complex changes occurred from 2000 to 2001 that impacted both scale and cut scores. In particular, the reporting scale scores, which had been linked to the original raw score metric from 1998 to 2000, were modified in 2001.[4] For quality control purposes, the contractor conducted an investigation in which the implemented cut score was examined for potential drift. In particular, the cut score was examined in terms of its IRT θ equivalent by year. For 10th grade, the relevant graph in the 2001 MCAS technical report (MDOE, 2002a) is given in Figure 1. It can be seen that the cut score in the logit metric appeared to rise in 1999 and 2000 and then drop again in 2001.

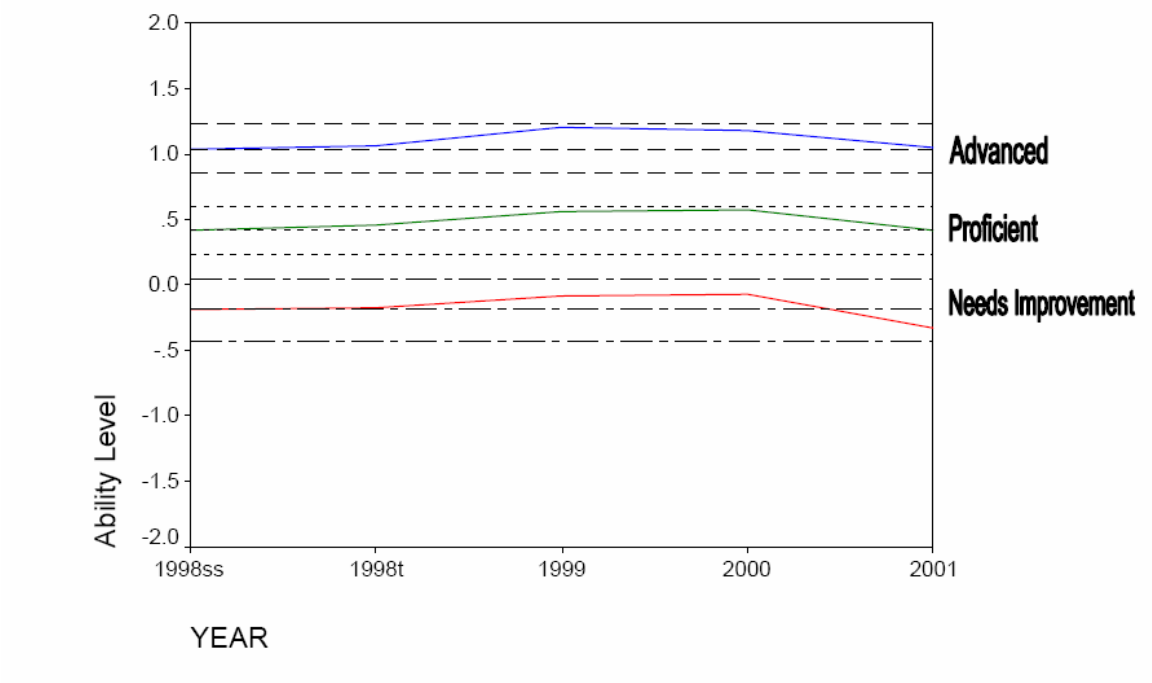

Figure 1
Tenth grade cut scores mapped against θ by year, excerpted from the 2001 technical report (MDOE, 2001a, p. 81).

[4] The new scale was based on two considerations. First, it was to be based directly on the underlying IRT logit (θ) metric. Second, it was adjusted in a way that, on the surface, resulted in more sensitivity to changes at the lower end of the proficiency continuum. The actual procedure is quite complex and is likely to be understood only by psychometricians. However, the results of the rescaling can be analyzed independently of the procedure itself.

The legend from the 2001 technical report (MDOE, 2002a) for Figure 1 provides some helpful elaboration:

> Theta scale (ability or measured construct level) established by calibration of the reference forms in 1998. Dashed lines represent cut points … obtained by standard setting in 1998…. (p. 81)

In 1998 passing scores were set for 10th grade MCAS mathematics and ELA. As new forms of the test are given, these two cut points should remain the same in the logit metric (labeled "Ability Level" in Figure 1) because each new form is equated to the 1998 base year with the FCIP method. The logic is that a cut point is like a hurdle, and the bar must remain at the same height to accurately gauge passing performance. Thus, the cut point in the logit metric should be a flat line in Figure 1, but it is not. Rather, it can be seen that the cut point for the "Failing" level drifted upward from its original value of -0.19 in 1998 to -0.06 in 2000. It then sharply decreased from -0.06 to -0.39 in 2001.[5] In other words, the cut point dropped about one third of a standard deviation assuming the logit scale was established as z-score scale in 1998 (a very common IRT practice). The effect of this downward change (i.e., lowering the bar) for Mathematics, as we shall see in the next section, is much greater than the 2001 technical report estimate of about 2% (MDOE, 2002a, p. 26). For ELA, the cut point for the "Failing" level drifted down slightly from its original value of -0.41 in 1998 to about -0.42 in 2000. It then decreased from -0.42 to -0.59 in 2001. Note that pushing the cut point (i.e., the bar) *down* is equivalent to pushing the entire proficiency distribution *up*.

Because the original cut points should not in principle change across administrations, they provide highly useful numerical values for evaluating pass rates for subsequent years in the IRT logit metric. We conduct a formal IRT analysis in the next section. However, the effect of the change in scaling is striking when the raw score distribution for mathematics given in Figure 2 is compared to its corresponding MCAS scale score[6] distribution given in Figure 3. It can be seen that while raw scores roughly follow a bell-shaped distribution, the scale scores spike at 220 and 260, the cuts for the passing and *Proficient* levels, respectively (i.e., the lowest scores in the passing and *Proficient* categories). Visually, it appears that the 2001 rescaling has pushed a number of scores up to the next level. We will show below that this impression is correct.

---

[5] The -0.39 value appears in Table 5.3.2.2 on p. 25 of the 2001 technical report (MDOE, 2002a). The value -0.06 was approximated from Figure 1.

[6] Currently, student scores are first defined in the IRT logit metric, which requires decimal values, and then transformed for practical purposes to a reporting scale of whole numbers that runs from 200 to 280.

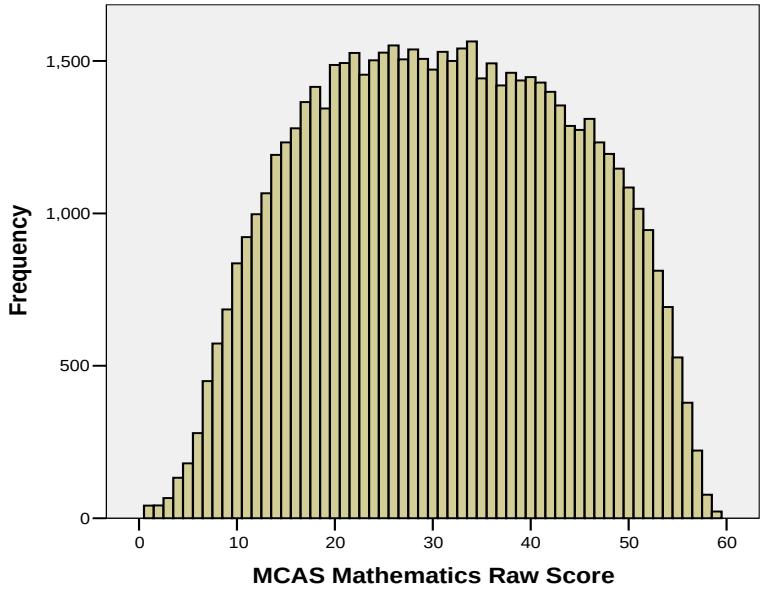

Figure 2
Frequency distribution of 10th grade 2001 MCAS mathematics raw scores, with scores of zero omitted.

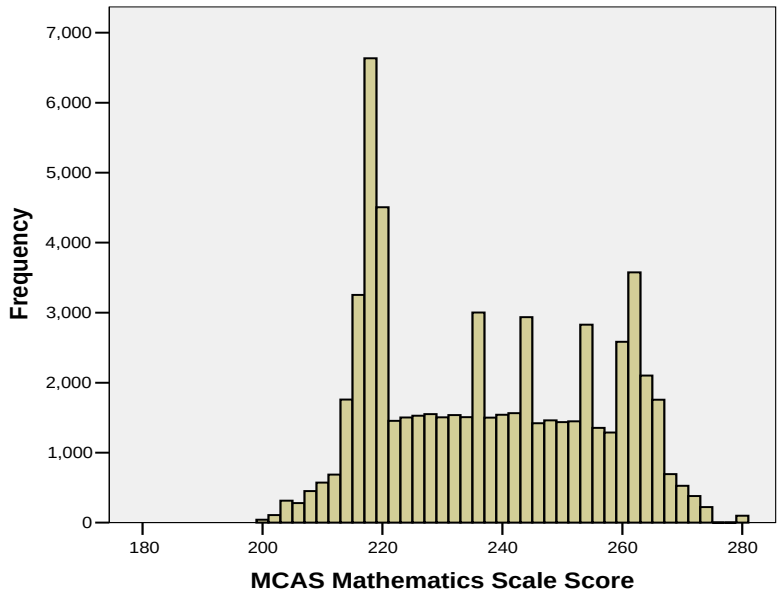

Figure 3
Frequency distribution of 10th grade 2001 MCAS mathematics scale scores (with corresponding raw scores of zero omitted).

Preliminary IRT Analyses[7]

As a first step, we attempted with IRT analyses of the 2000 and 2001 examinee data for both Mathematics and ELA to reproduce the operational *b* parameter estimates (or IRT item difficulties) as reported in the MCAS technical manuals. For the 2000 ELA and Mathematics examinations, we were able to reproduce the item difficulties for all items. However, when the same steps were taken with the 2001 test data, we observed large discrepancies between our OR *b*-values and those reported in the MCAS technical manuals.

Recall that as a *b*-value moves in the negative direction, a test item becomes easier. In 2000, our OR *b*-values were lower than the reported MCAS values by approximately .10 logits for the ELA examination and approximately 0.03 logits for the Mathematics examination. In contrast, our *b*-values for OR items in 2001 were lower by approximately 0.20 logits for the ELA examination and 0.40 logits for the Mathematics examination. (Note that these logit differences can be interpreted accurately as effect sizes). Thus, in 2001, it appears that the OR items were easier than expected relative to the other test items.[8]

Main Analysis

Given this discrepancy, we used the two alternative IRT methods (given in the Appendix) to estimate pass rates. The first used OR *b*-values from the MCAS technical report, and the second adjusted the *b*-values for the decrease in difficulty from 2000 to 2001. In both methods, the original 1998 cut scores in the logit metric, as shown in Figure 1, were employed. Additional description is given in the Appendix.

Results

The results of these analyses are given in Table 6, which presents the reported total percentages of students who passed each examination in each year as well as the percentage of students falling in the *Proficient* category or higher.

An "adjusted reported" category is indicates the percentage based solely on students who were present for the examination in each category. When reporting results, those students who were not present for an examination (but were eligible to participate) were considered failing. There is approximately a 2% difference between the reported pass rates and the adjusted pass rates in 2000, while less than 1% in 2001. For simplicity, we shall consider the adjusted MCAS pass rates when drawing our conclusions since our samples did not include non-present student (i.e., those students considered failing, not present).

---

[7] We hope that psychometricians will understand that we have access to the operational MCAS *b*-value estimates of 2000 and 2001. A previous reviewer of this work failed to grasp this elementary point in contending that nothing could be established without access to the linking items' *b*-values.

[8] There are several reasons why this might be the case including pre-equating or samples used for item calibration. We can not determine the exact reason from publicly available information.

Table 6
Reported and Estimated 2001 Percentages of Passing and Proficient Students

	English Language Arts			Mathematics		
Parameter	2000	2001	Gain	2000	2001	Gain
Reported passing rates						
Pass	66.0%	82.0%	16.0%	55.0%	75.0%	20.0%
Proficient	36.0%	50.0%	14.0%	33.0%	45.0%	12.0%
Reported rates, adjusted for non-participation						
Pass	68.0%	82.8%	14.8%	56.7%	75.8%	19.1%
Proficient	37.1%	50.5%	13.4%	34.0%	45.5%	11.4%
Official *b*-values for all items (Method 1)						
Pass	67.3%	77.8%	10.5%	57.4%	69.3%	11.9%
Proficient	37.5%	48.2%	10.6%	34.8%	41.4%	6.6%
Official *b*-values with estimated OR (Method 2)						
Pass	65.9%	75.7%	9.8%	56.8%	60.9%	4.1%
Proficient	35.9%	44.5%	8.7%	34.4%	36.3%	1.9%

For the 2000 data, we were nearly able to reproduce (about a 1% discrepancy) the pass rates[9] for both the ELA and Mathematics examinations using the MCAS reported item parameters. The results for 2001 differed by slightly more with this method at approximately 5% and 6% for the ELA and Mathematics examinations, respectively. However, when we used the 2001 implemented cut scores as shown in Figure 1, we were able to estimate pass rates in 2001 that were within 1–2% of adjusted pass rates. The latter finding suggests that potential sample differences did not unduly affect the results.

In Table 6, it can be seen that the unadjusted pass rate increased from 68% to 82.8% (+14.8%) for 10th grade ELA. Taking into account the drift in the cut score and changes in OR item behavior, we estimated that the adjusted pass rate increased from 65.9% to 75.7% (+9.8%). Thus, the 2000–2001 gain in ELA is likely to be over-estimated by about 5% (14.8% - 9.9%). Similarly for Mathematics, we estimated that the adjusted pass rate increased from 56.8% in 2000 to 60.9% in 2001. Thus, the 2000–2001 gain in Mathematics is likely too high by about 15%.

The change in scaling affected all of the MCAS tests to some degree, which can be seen upon inspecting the figures in Appendix I of the 2001 technical report (MDOE, 2002a, "MCAS Performance Levels Mapped to Theta Scale"). We have only analyzed the results for 10th grade ELA and Mathematics in this report; the 2000–2001 gains for all 2001 MCAS tests may contain varying degrees of statistical bias.

Discussion

Our goal is to show how the psychometric aspects of tests (i.e., scaling and equating procedures) can adversely affect reported student pass rates.[10] States have made large investments in

[9] "Pass rate" refers to the area of the theta distribution above the reported MCAS theta cut-point for a particular examination.

[10] Scaling changes made in 2001 are reported in the Massachusetts DOE on pages 20–26 in the 2001 technical manual.

assessment programs, and it is critically important that scaling issues do not distort the interpretation of achievement gains. It is important to note at the same time that the present study supports the quality of the MCAS and its technical documentation. Though a flaw was found with the manner in which cut scores were determined, this should not be taken as evidence against the validity or reliability of the test per se.

As Cizek (2001) and Popham (2003) have noted, the educational measurement community should be involved in debates regarding the efficacy of testing. However, while Cizek (2001) complained that measurement experts have been silent on the benefits of high-stakes testing, Popham (2003) argued that most students are receiving educations of decisively lower quality as a result of high-stakes testing (in stark contrast to his positions of 20 years ago[11]). Popham further elaborated that in contrast to sins of *commission*, which are easily spotted,

> Sins of *omission* can also have serious consequences. And those are the kinds of sins that the educational measurement crowd has been committing during recent decades. We have been silent while certain sorts of assessment tools, the very assessment tools that we know the most about, have been misused in ways that harm children. (p. 46)

What sense can be made of these competing claims of "silence"? We would argue first and foremost that the conceptual territory here is not black and white. There are many purposes for testing, and these purposes are ranked differently depending on one's values, philosophy, and political ideology. The Massachusetts experience has provided a sort of ink blot into which various beliefs about causality have been projected. Measurement specialists are not immune to such influences, and there is no reason to believe that they will forge a greater consensus on the issue of high-stakes testing than exists in the prevailing social context. However, one thing that all stakeholders can agree upon is the need for accurate estimation of program effects. Yes, there is much explaining to do once the effects are measured, but the difficulty of accurate measurement, especially with regard to annual progress, should not be underestimated.

In Massachusetts it was possible to refute the grander claims regarding the effects of high-stakes testing in 2001 because the State Department of Education scrupulously detailed the psychometric characteristics of the MCAS assessment. The kind of analyses undertaken in this paper could not be carried out in most states using publicly available information. This is one reason that we do not currently have a very good notion of how broadly (across states) assessment errors have affected educational policies. Though some errors have been reported, it is likely the case that many others have passed silently (Rhoades & Madaus, 2003). Many identified problems have concerned inconsistent or incorrectly scored items. Such cases are relatively easier to detect than those involving scaling, scoring and equating. Yet the latter are more likely to confuse state educational policies as well as to muddy the debate on the merits of high-stakes testing.

Several conclusions and recommendations can be drawn regarding the change in MCAS scores form 2000 to 2001 considered in this paper. First, we did estimate gains from 2000 to 2001 in both English Language Arts and Mathematics, but the gains were much smaller than those in official reports. The ELA 10th grade gain was moderate resulting in a reported pass rate of 81% in 2001; but in 1998, the 8th grade NAEP rate for *Basic and Above* (the lowest category being *Below Basic*) in reading was 79%. The MCAS gain for Mathematics was relatively small resulting in a reported pass rate (partially proficient) of 75% in 2001; but in 1996, the 8th grade NAEP rate for *Basic and Above* was 68%. The *de facto* achievement levels implemented in 2001 thus appear more consistent with *Basic* achievement levels set in the National Assessment of Educational Progress (NAEP) in 8th grade, while the original achievement levels set in 1998 were more severe. Such discrepancies should be understood as general and common problem of setting standards. Standard setting procedures

[11] See Camilli, Cizek and Lugg (2001) for a short history of Popham's views.

are designed to be internally consistent but require essentially establishing arbitrary cut points on a continuum of test scores (Camilli, Cizek & Lugg, 2001). Different procedures and contexts produce different results, and this topic has remained controversial. In 2005, the Education Department, in a shift from previous practices, presented state results with "charts showing state-by-state trends focused on results for just the basic level, which denotes what NAGB regards as 'partial mastery' of the skills students should acquire at particular grade levels" (Viadero & Olson, 2005, p. 14). Critics such as Diane Ravitch have decried this phenomenon as a lowering of standards, while state policy makers tend to view the *Basic* level as a plausible criterion for student proficiency. With NAEP, there has also been controversy regarding how achievement levels should be established and interpreted (Pellegrino, Jones & Mitchell, 1999; Hambleton et al., 2000). Nonetheless, the consistency of 2001 10th grade MCAS results with extrapolations from earlier NAEP 8th grade performance might be taken as a positive unintended outcome.

Second, the evidence from Massachusetts supporting the efficacy of high-stakes accountability is mixed. A more compelling explanation is that mathematics scores had been rising all along—and the upward trend existed prior to the implementation of new graduation requirements. However, there is evidence from NAEP that reading scores have been stagnant nationally in all grade levels as well as in Massachusetts at 4th and 8th grade. Rising ELA scores in 10th grade thus signify some proficiency not reflected in earlier grades by NAEP. Another hypothesis is that score cut points have continued to drift downward (in the θ metric) because the method of producing scale scores may be overly sensitive to slight changes in testing procedures.

Finally, unprecedented gains, such as those which occurred in 2001 MCAS proficiency at the 10th grade, should be recognized by scholars as prime candidates for further study. Indeed, if an increase in student proficiency seems almost too good to be true, then some degree of skepticism is both appropriate and healthy. Grissmer, Flanagan, Kawata, and Williamson (2000) showed that the annual gains of any state on NAEP average about 0.03σ (but can be as high as 0.06σ) per year. When a gain nearly an order of magnitude larger than this is observed, as it was in 10th grade MCAS Mathematics, it should receive additional scrutiny. This is not simply a technical issue. Failing to obtain accurate estimates of achievement gains can result in false perceptions that lead both educators and pundits astray.

References

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics*, *13*, 227–241.

Camilli, G., Cizek, G.J., & Lugg, C.A. (2001). Psychometric theory and the validation of performance standards: History and future perspectives in G.J. Cizek (ed.) *Setting performance standards* (pp. 445–476), Mahwah, NJ: Lawrence Erlbaum.

Education Trust (2004). Measured progress: States are moving in the right direction in narrowing achievement gaps and raising achievement for all students, but not fast enough. Washington, D.C.: Author.

Finn, C.E., Jr. (2001, October 18). Vindication for the MCAS: Dramatic improvement in student scores in MA. *The Education Gadfly*, *1*(22). Retrieved February 1, 2006, from http://www.edexcellence.net/foundation/gadfly/issue.cfm?edition=&id=86#1296.

Fuller, B. (2004, October 13). Are test scores really rising? School reform and campaign rhetoric. *Education Week, 24*(7), 40, 52.

Gaudet, R. D. (2002, June). *Student achievement in Massachusetts: The lessons of nine years of education reform*. Education Benchmarks Report #1–02. University of Massachusetts, Amherst, MA: Donahue Institute.

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

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Linn, R.L. (1998). Assessments and accountability. *Educational Researcher, 29*(2), 4–14.

Massachusetts Department of Education. (2001a). *Guide to interpreting the spring 2001 reports for schools and districts*. Malden, MA: Author. Retrieved February 1, 2006, from http://www.doe.mass.edu/mcas/2001/interpretive_guides/fullguide.pdf.

Massachusetts Department of Education. (2001b, November). *Spring 2001 MCAS tests: Report of 2000–2001 school results*. Malden, MA: Author. Retrieved February 1, 2006, from http://www.doe.mass.edu/mcas/2001/results/school/g10s_0001res.pdf.

Massachusetts Department of Education. (2002a). *2001 MCAS technical report*. Malden, MA: Author. Retrieved February 1, 2006, from http://www.doe.mass.edu/mcas/2002/news/01techrpt.pdf.

Massachusetts Department of Education. (2002b, May). *2000 MCAS technical report*. Malden, MA: Author. Retrieved February 1, 2006, from http://www.doe.mass.edu/mcas/2002/news/00techrpt.pdf.

Mehrens, W.A., and Cizek, G.J., (2001). Standard setting and the public good. In G.J. Cizek (Ed.) *Setting performance standards* (pp. 477–485), Mahwah, NJ: Erlbaum.

Muraki, E., and Bock, R. D. (2003). *PARSCALE (Version 4.1): Analysis of graded responses and ratings* [Computer program]. Chicago, IL: Scientific Software International, Inc.

Popham, W.J. (2003). Seeking redemption for our psychometric sins. *Educational Measurement*: *Issues and Practices, 22*(1), 45–48.

Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem*. Boston, MA: National Board on Educational Testing and Public Policy, Boston College.

Viadero, D., and Olson, L. (2005, November 2). Focus on 'Basic' achievement level on NAEP stirs concern. *Education Week, 25*(10), 14.

Appendix

Item Calibration. All calibrations were carried out using the IRT software program PARSCALE in which MC, SA, and OR items (Muraki and Bock, 2003) can be jointly scaled. A three parameter logistic model was used for MC items; this model was also used for SA items with the guessing parameter set to $c = 0$. Samejima's graded response model was used for OR items.

Estimating Pass Rates. As a standard feature or PARSCALE, the posterior distribution of examinee proficiency is output. Also referred to as the latent population distribution (LPD), it has been shown (see Camilli, 1988, for a more extensive discussion) that this distribution is far more accurate than the distribution of estimated abilities—especially with respect to measurement error and unestimable examinee proficiencies. Because the LPD is scaled in terms of IRT item parameters (*a*, *b*, and/or *c*), it is fixed on the basis of the FCIP (fixed common item parameters) equating method used with the MCAS. Both individual examinee item response patterns and item parameter estimates are required to obtain the LPD. A normal prior distribution was used, and Figure 2 suggests this assumption is highly plausible.

To use the LPD to estimate pass rates, and original criterion or cut score must be translated into the IRT metric in the year the cut score was determined. For 10^{th} grade, these scores were set in 1998 for both ELA and Mathematics. Once a cut score is obtained in the IRT metric, the percentage of the LPD above (or below) the cut is used to estimate the percent above criterion (PAC). For the lowest cut score on the MCAS the PAC is described as the pass rate whereas the percent below is described as the "Failing" rate. The LPD is obtained as a series of theta scores (quadrature points) with associated probabilities or "weights," that is, as a discrete density function. Percentiles are obtained by summing weights below the cut score, possibly with some interpolation. We found there was little difference between 50 and 100 quadrature points; all the results below are based on the latter number.

Method 1 (fixed MCAS item parameters). Using the item parameters reported in the MCAS technical manuals, we estimated posterior theta distributions for both examinations in both years using all items (MC, SA, and OR). This method was considered a referent analysis. Though there were minor discrepancies between our samples and the samples on which the MCAS reports were based, such discrepancies most likely had a negligible effect.

Our analyses of the classical item statistics and the IRT item parameters reported in the MCAS manuals for the OR items suggested that some error or combination of errors led to systematic misestimation of these items' parameters. The results of our Method 1 analyses further supported this belief due to the disproportionately large chi-square misfit statistics observed for these items. More specifically, review of the model-fit outputs for 2001 ELA and Mathematics examinations revealed that OR items statistics (or rather "misfit" statistics) were on average twice as large, relative to degrees of freedom, than those observed for the non-OR items. Method 2 discussed below addresses this issue.

Method 2 (fixed MCAS MC, estimated OR parameters). Using the procedure discussed above of fixing item parameters, we fixed the MC values on the ELA examination, both the MC and SA values on the Mathematics examination to the reported MCAS values. However, we allowed the OR items to be estimated. Then the posterior theta distributions were again generated for both examinations in both years. The success rates for *Pass* and *Proficient* obtained from these analyses (labeled Method 2 in Table 6) were slightly lower than Method 1 ELA pass rates. For Mathematics the differences was larger. We estimated success rates that 7.8% and 4.7%, respectively, less than the Method 1 success rates. This suggests that OR items were more problematic on the Mathematics examination.

About the Authors

Gregory Camilli
Rutgers, The State University of New Jersey

Sadako Vargas
Rutgers, The State University of New Jersey

Email: Camilli@rci.rutgers.edu

Gregory Camilli is Professor in the Rutgers Graduate School of Education. His interests include measurement, program evaluation, and policy issues regarding student assessment. Dr. Camilli teaches courses in statistics and psychometrics, structural equation modeling, and meta-analysis. His current research interests include school factors in mathematics achievement, test fairness, technical and validity issues in high-stakes assessment, and the use of evidence in determining instructional policies.

As Research Associate at Rutgers Graduate School of Education, and Adjunct Professor at Touro College and Seton Hall University, **Sadako Vargas** has taught in the areas of research methods and occupational therapy. Her interests lie in the use of meta-analysis for investigating intervention effects in the area of rehabilitation and education specifically related to pediatrics and occupational therapy intervention.

EDUCATION POLICY ANALYSIS ARCHIVES http://epaa.asu.edu

Archivos Analíticos de Políticas Educativas