What Every Mathematician Should Know about Standards-Based Tests

Alan Tucker

Abstract. This article examines serious psychometric-based problems that caused the New York Math A graduation test introduced in 1999 to produce a 70% failure rate in 2003. In discussing this test, we highlight potential difficulties with all standards-based tests.

1. INTRODUCTION. There have been many criticisms of the extensive use of testing in school classrooms, such as excessive time spent preparing for tests. This article reveals another serious testing problem whose technical nature has shielded it from scrutiny, namely, deficiencies in the way psychometric theory is being applied to design standards-based tests. Given the tradition of social promotion and variable expectations in many schools, serious standards are only credible in the U.S. when validated by high-stakes tests. Accordingly, the No Child Left Behind Act gave great weight to year-to-year improvement in a school's test scores, but psychometric flaws in test design can make year-to-year comparisons meaningless.

The case study used to illustrate this problem is the Math A graduation test that New York State introduced in 1999. In the late 1990s, the NY Board of Regents mandated that all NY high school graduates pass five exams, one being mathematics, and switched from traditional teacher-designed tests to psychometrically designed "standards-based" tests. Simultaneously, a new three-semester Math A course replaced the old two-semester Math I course, whose test was dominated by one-step, "mechanical" problems. The Math A test would have more topics and more multi-step problems. Test questions varied considerably from year to year. This raised the critical issue of how a constant performance standard can be maintained for passing such a test. On the first Math A test in 1999, the psychometrically determined passing score was 43 out of 85 points. It rose to 51 out of 85 by 2003, although the questions were becoming harder. In June 2003, a year after the Math A test become mandatory, this graduation test made headlines with a 70% failure rate.

New York's Commissioner of Education fired the head of testing, set aside the June 2003 Math A scores of juniors and seniors, and appointed a special Math A Panel, which included this writer, to examine what went wrong. The Panel report [**5**] highlighted the challenges that face well-intentioned efforts to bring higher standards to school mathematics. It documented serious shortcomings, including unexplained psychometric problems. Subsequent studies by this writer uncovered many details of the psychometric failures. Because these studies were based on NY State Education Department proprietary field test data provided to the Math A Panel, it seemed prudent to delay publication of these controversial findings. There is now a different curriculum and test, and the State Education administrators who worked with the Math A Panel have left.

Standards-based tests are often designed using Item Response Theory (IRT), a psychometric theory for testing that has methodology for maintaining a constant performance standard over time. This article describes design flaws arising from New York's application of this theory to the Math A test. Some of the flaws have likely occurred on

doi:10.4169/amer.math.monthly.118.05.434

 © THE MATHEMATICAL ASSOCIATION OF AMERICA

other state mathematics graduation tests. Massachusetts and Oregon had unexpectedly high failure rates on their first standards-based mathematics graduation tests.

Policy makers want a way to assess year-to-year progress when new programs are instituted to improve school achievement. Standards-based tests have been presented as a scientifically based assessment method that can measure student performance with scores that are consistent over time, like SAT scores, so as to be the basis of high-stakes decisions, such as to let parents remove their children from a school that is not improving enough. Organizations such as Educational Testing Service (ETS) use IRT to produce highly reliable test scores. State education agencies, however, have limited psychometric expertise and much smaller test development budgets. They rely on educational contractors to design and implement the psychometric components of a standards-based test.

The problems discussed here do not arise in states with tests of highly predictable questions. However, with those tests accurate annual comparisons are distorted by instruction centered around the test questions. It is serious efforts to assess challenging new curricula aiming for higher achievement, as the Race To The Top legislation calls for, that are most vulnerable to these problems.

2. DESIGN FRAMEWORK FOR THE MATH A CURRICULUM AND TEST. While the performance standard of the Math A test is the focus of this article, it is closely linked to the content standards, which had their own problems. The Math A course had 32 content standards that were developed by high school mathematics teachers and State Education Department staff. These 32 standards were broken down into 103 sub-indicators—specific concepts, techniques, or classes of problems. Performance standards for the mastery of the content standards were assessed by the 35-question Math A test. The test was 3 hours long and contained 20 multiple-choice problems, each worth 2 points, and 15 written-response questions, 5 each with 2-point, 3-point, and 4-point values.

The content standards were grouped into the following seven *Key Ideas*:

I. Mathematical Reasoning
II. Number and Numeration
III. Operations
IV. Modeling/Multiple Representations
V. Measurement
VI. Uncertainty
VII. Patterns/Functions.

This classification was a mixture of traditional disciplinary topics, such as Number and Operation, and crosscutting topics, such as Mathematical Reasoning and Modeling/Multiple Representations. Geometry and algebra were spread across several Key Ideas. Even the topic of functions, with its own category, was spread across several other Key Ideas. Many other states have similar classifications of their content standards.

There were approximate percentages for the proportion of test questions coming from each Key Idea. However, the content of the Key Ideas overlapped so much that the Math A panelists frequently had no idea from which Key Idea a question was chosen. There were no requirements about major topics within Key Ideas, although most test questions were based on a small subset of the sub-indicators. However, the June 2003 test had no questions involving trigonometry while all previous tests had had several trig questions.

A standards-based test is supposed to assess a fixed performance standard. Statistical equating methods make test scores comparable on different versions of a test (e.g., year to year) under the assumption that all versions of the test have questions of relatively similar difficulty and content. However, the Math A test was phased in, transitioning to the more demanding questions in the new curriculum. A sample Math A test was released for the first Math A test, but there was no plan for what the steady-state Math A test would be like. Over time, test questions evolved to more difficult versions of previous types of problems and introduced new problem types. For example, two June 2003 problems required construction of a median bisector using ruler and compass. This topic had never appeared on a previous Math A test and many teachers stopped covering it. State Educational staff assumed that psychometric equating could adjust the passing score appropriately as the tests got harder, but the assumptions underlying equating methods were being violated. Even worse, we will show that this misuse of equating methodology appears to be the real reason the test questions were getting harder.

3. SETTING A PERFORMANCE STANDARD.

3.1. The Bookmark and Benchmark. The first step in a standards-based test is setting the performance standard. This standard is an ability level, called the **benchmark** ability value. With psychometric methods, the expected score of a student with the benchmark ability can be computed for each test. That expected score will be the test's passing score.

One assembles a collection of questions that might appear on the test and gives them to a representative sample of students. The questions are ranked by their p-score, the percentage of students who get them right. A group of experts—mathematics teachers and professors—go down the ranking, from easiest to hardest, looking to set a **bookmark**, a question judged to be of a difficulty such that students just meeting the performance standard would get this question right on average $2/3$ of the time. (Multiple bookmarks could have been used, but NY State Ed and its contractors decided to use just one bookmark.) Psychometric methods are used to assign a numerical value to the ability of a student who can solve the bookmark question correctly $2/3$ of the time (this process is described in the next section). This ability number is the benchmark for the test. The Math A Panel learned that the bookmark chosen for the Math A test was a question that 55% of students in the sample answered correctly (the Math A Panel never saw the actual question).

One criterion for setting this bookmark is if the experts think that the next three (harder) questions on the list are unlikely to be correctly solved $2/3$ of the time by a student at the borderline for passing. There are procedures for helping the group of experts reconcile their different initial choices of a bookmark question to collectively agree on a single bookmark. For more about setting performance standards, see Cizek [**3**].

The use of a single question to set a performance standard for assessing a whole year (or more) of study in mathematics is ludicrous to anyone who teaches mathematics. There is obviously a major component of underlying mathematical ability that is reflected in students' performances on a question about any topic; statistical analyses confirm this. But factual knowledge is also important: teachers may cover some topics thoroughly and others superficially or not at all. Given the way the topics on the Math A test varied greatly, this factual knowledge would be a significant factor.

Most importantly, for many students there is a significant difference between their ability to solve problems on which they have been drilled and their ability to solve

simple but novel problems. If the performance standard is to solve a moderately hard, familiar procedural question 2/3 of the time, these students can meet this standard. If the standard is to solve a moderately easy, but unfamiliar, question 2/3 of the time, they will not meet the standard. This is a dangerous intersection of performance and content standards.

Finally, there are many pragmatic details in the standards-setting process. Who are the expert professors and teachers? The most experienced teachers tend to teach in high-performing schools, since low-performing schools have high teacher turnover. There are also issues with the sample of students chosen to work the bookmarking questions. Teachers on the Math A Panel asserted that schools usually offer better students for such activities because they fear the results will be used to judge the schools. Finally, without high stakes, students will not put in the effort in the bookmarking questions they would on a graduation test.

3.2. Subjective Nature of the Judgment. A well-known problem with setting a performance standard is that different groups of experts will come up with quite different standards. One factor in New York was that a new curriculum was being developed with higher standards and so the bookmarking panel was told to set a standard above the current performance of an average student. While the bookmark question was correctly solved by only 55% of students in the sample, the performance standard required this question to be correctly answered with probability 2/3. Thus the new performance standard represented a graduation expectation that was likely to forever be out of reach of below-average students.

3.3. Problems in Ordering the Difficulty of Standard-Setting Questions. Students' performances in questions used to set the standard are dependent on students' familiarity with the types of questions. Mathematics teachers and professors may rank the difficulty of questions in a very different order. When students and teachers have different scales of difficulty, the whole bookmark selection process breaks down.

This mismatch was apparent to the Math A panelists who reviewed the January 2004 Math A test. They complained that many of the early questions were much harder than later questions. However, State Ed staff showed the panelists field test data indicating that the test questions were actually ordered by increasing difficulty for students. One suspects that in another year, when students would have drilled on questions from more recent Math A tests, their performance would rank the questions in a different order, leading to a different bookmark if the performance standard were set again. Some other states have gotten around these problems by having predictable types of questions on their tests every year. Then student performance will be more consistent, but these tests encourage the type of mindless learning that New York had wanted to de-emphasize. (Under pressure to show rising performance under NCLB, New York subsequently turned to highly predictable questions on its annual math tests in grades 3 thru 8, questions for which many students were excessively drilled. In 2010, after criticism, a greater variety of questions was used. Again low passing rates made headlines; for example, the percentage of proficient eighth graders in Buffalo fell from 58% in 2009 to 26% in 2010.)

3.4. Out-of-Date Questions in Standard Setting. The performance standard was set before the Math A course existed. The Math A Panel was told that the bookmarking questions were based on the old Math I, II courses, and the students who worked these questions were taking the old Math I, II courses. The new standard was an extrapolation from the old math skills in Math I, II to higher future expectations in the problem-

solving Math A course. Using out-of-date questions to set a new higher standard to be met with a new curriculum is a serious misuse of the standards-setting process. Any state moving to more demanding mathematics curricula, as encouraged by the Race To The Top legislation, will face a similar problem in a standards-based graduation test.

3.5. Two Roles for a Performance Standard. There is a major conflict in the two different ways that a performance standard can be used. The first, most common use is as a proficiency standard that is set at a comparatively demanding level. Initially many students, in some schools a majority, may not meet this standard, but over time most students should meet it. Such a standard can be helpful in documenting differences in achievement by subgroups of the population. The second use is as a graduation requirement. Such a standard needs to be lower. If expectations for graduation are rising, this graduation standard would move up with these rising expectations, but still allow most students to graduate each year. Unlike the first type, students would be allowed to take it several times with remedial assistance.

The performance standard for the Math A test should have been the second type. However, it was set at a level significantly above the ability of the average 1998 student and likely out of reach of below-average students (as discussed in Section 3.2).

4. MAINTAINING A CONSTANT PERFORMANCE STANDARD.

4.1. Overview of the Math A Test. A performance standard is independent of a particular year's test. To implement this, the raw scores (between 1 and 85) on each administration of a Math A test are mapped onto scaled scores (between 1 and 100), with the mapping adjusted for each test by a process that is meant to equate comparable performances over time.

In New York, a scaled score of 65 corresponds to the performance standard, i.e., a scaled score of 65 is passing. A performance standard is also established for a High Pass, and a scaled score of 85 corresponds to the high passing raw score. Suppose a is the passing raw score and b is the high passing raw score. Then the mapping function takes a to 65 and b to 85, as well as 0 to 0 and 85 (the maximum raw score) to 100.

Regents graduation tests are given annually in June, August, and January. The proposed tests for year N are field tested in year $N-1$. These tentative tests are constructed from questions that were created in year $N-3$ and performed well on pretests in year $N-2$. A set of "anchor questions" whose difficulty was established at the time of the first Math A field test is included in the field tests every year. Differences in the abilities of the students taking field tests are determined by comparing their performance on the anchor questions against the anchor performance of the original group of students on the first field test. Once the ability levels of the current test field takers and the difficulty levels of the test questions are known, psychometric methods determine the raw score that meets the performance standard; that score is mapped to a scaled score of 65. New York is a Truth-in-Testing state where test questions are released two days after a test and so field tests are the only way to collect data on test questions.

4.2. Introduction to Item Response Theory. Here is a summary of the relevant parts of Item Response Theory (IRT) (Baker [**1, 2**], Lord [**4**], Van der Linden [**6**]). The version of IRT used for the Math A test has a quantitative model for assessing student performance based on the following three premises.

A. *Student proficiency.* IRT assumes that each individual's level of mathematical proficiency can be accurately represented by a single number, called a θ-value.

B. *Question difficulty.* IRT assumes that each test question's difficulty can be described by a single number called the question's b-value. There is also a scale parameter α associated with a question's response curve.

C. *Probability of a student's success on a question.* IRT assumes that an item response curve of the form $p_{b,\alpha}(x) = 1/\{1 + e^{-\alpha(x-b)}\}$ describes the probability of a correct answer by a student of ability x on a question of difficulty b.

There is a fourth technical premise in IRT involving the stochastic independence of responses to questions by students with a given θ.

Observe that the b's and θ's are on the same scale. The b-value of a question is defined by the location of the midpoint (50 percent point) on the question's response curve. That is, a question is assigned a b-value of θ if a student of ability θ has a .50 probability of correctly answering the question. The common b/θ scale is typically centered so that the average θ-value of students' proficiencies is 0, and units are measured in standard deviations.

Item response curves used in this version of IRT are logistic functions. These curves are similar in form to the cumulative distribution function of a normal distribution. The underlying model is that for students of very low or very high ability, the chances of success on a question are essentially 0% or 100%, respectively, while the chances of success increase significantly with increasing ability for students' whose ability level is close to the b-value of the question. Figure 1 shows the response curve with $b = 0$, $\alpha = 1$. According to the curve in Figure 1, among a group of students with proficiency $\theta = +.7$, on average $2/3$ would correctly answer a question with difficulty $b = 0$, while among students with proficiency $\theta = -.7$, on average $1/3$ would correctly answer this question.

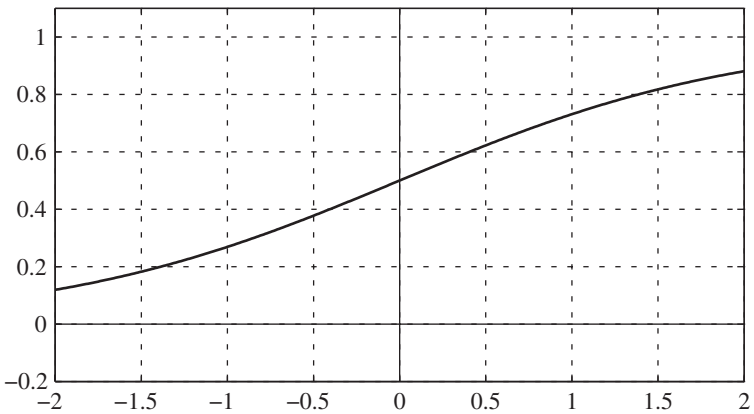

Figure 1. Graph of $p_{0,1}(x) = 1/(1 + e^{-x})$.

4.3. Determining Model Parameters. After a group of students takes a field test, maximum likelihood estimation (MLE) is applied to the results to determine an ability rating θ for each student and the parameters b and α in the logistic response curve for each question. In the simplified situation where the b-values and α-values of the questions have been determined in advance, MLE works in the following way to determine the θ-value of each student. Let b_i and α_i be the b-value and α-value of the ith question. The θ-value of student S is the value of x that maximizes the expression

$$\prod_i p_{b_i,\alpha_i}(x) \prod_j (1 - p_{b_j,\alpha_j}(x))$$

where i ranges over the problems that the student got right and j ranges over the problems that the student got wrong. This maximum is found by taking the derivative and using a root-finding method such as Newton-Raphson to find the zeros of this expression. (Technical notes: (i) a variation of MLE called marginal MLE is often used because it has better statistical properties; (ii) for written-response questions, a more complicated version of MLE must be used.)

The b-values and α-values of questions are determined in a similar fashion if the θ-values of students are known. The process typically starts with a sample of students whose proficiencies were previously determined. Then test questions are field tested with this sample to determine question parameters. It was never clear how field test data were used to determine parameters for Math A test questions.

Because the sample sizes in the New York field tests were not large enough, the Rasch IRT model was used in which the scale parameter α is dropped. The scale parameter indicates the rate at which the probability of a right answer increases as a student's ability increases. In the Rasch model, a common scale parameter is used to fit the item response curves of all questions. This scale parameter is incorporated into the θ-scale. That is, the size of one unit in the θ-scale is determined by this scale parameter.

By the design of the Rasch model, the b-values of questions should have an exact functional relationship with the p-scores (probability of a correct answer in field tests). Figure 2 shows a plot of b-values versus p-scores for the questions on the initial June 1999 Math A test (based on 1998 field test data). The line that best fits the data in Figure 2 is:

$$(p\text{-score}) = .55 - .19(b\text{-value}) \qquad (*)$$

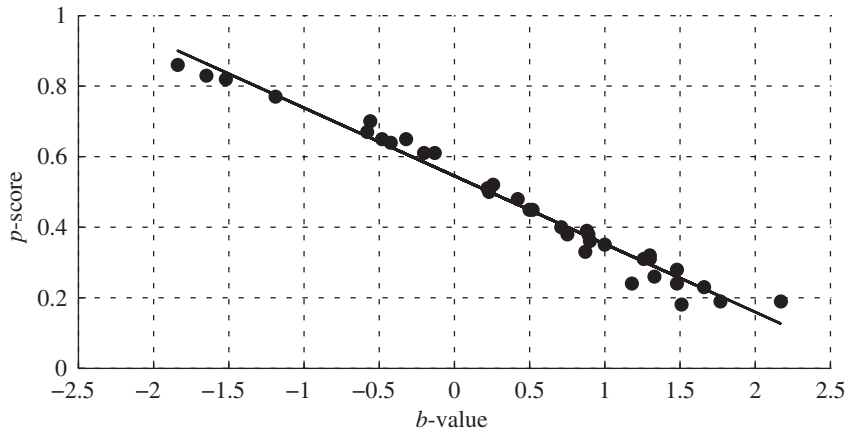

Figure 2. June 1999 test b-values versus p-scores, based on 1998 field tests.

Recall that the bookmark question had a p-score of .55. Let b_{bookmark} denote the b-value of this bookmark question. One sees from the regression line $(*)$ that the bookmark p-score of .55 corresponds to the b-value of 0; that is, the θ/b scale was centered so that $0 = b_{\text{bookmark}}$. Normally, the θ/b scale is set so that the average θ-value of student proficiencies is 0. This non-traditional centering would later cause a big problem when a new psychometric contractor was hired who assumed the θ/b scale was based on a different centering method. Recall that the ability benchmark, $\theta_{\text{benchmark}}$, is the ability level of a student who just meets the Math A performance standard. Such a student

correctly answers the bookmark question of difficulty $b_{\text{bookmark}} = 0$ with probability $2/3$. Figure 1 shows that $p_{0,1}(.7) = 2/3$. Thus $\theta_{\text{benchmark}} = .7$. Note how important the response curve is for setting the benchmark. If the slope of this curve is too high or too low (compared to the actual student performance in the field tests), the flawed curve will lead to an incorrect value for $\theta_{\text{benchmark}}$, which is used to set the passing score. If the response curve for the bookmark question should have had a steeper slope, say $\alpha = 2$, then we find that $p_{0,2}(.35) = 2/3$ and $\theta_{\text{benchmark}} = .35$. This decrease of .35 in $\theta_{\text{benchmark}}$ would lower the passing score by approximately 5 points, a major change in the performance standard.

4.4. Determining the Passing Score. After the b-values of a test's questions have been determined from field test data and psychometrically put on the same scale as used to set the performance benchmark, the passing score is determined. It is the expected score of a student of ability $\theta_{\text{benchmark}}$, the θ-value of the performance standard. Recall that $\theta_{\text{benchmark}} = .7$.

The expected score of a .7-ability student on a test is found by summing up this student's probability of success on each question times the point value for that question. On a question with difficulty b, a .7-ability student's probability of a correct answer is $p_{b,1}(.7) = 1/\{1 + e^{-(.7-b)}\}$. A more complicated procedure is used for estimating expected scores on written-response questions.

4.5. Maintaining the Scale over Time. In the field testing during year $N - 1$, data are collected for the three tests to be given in year N, plus a fourth back-up test. A copy of the anchor questions is paired with each test. Recall that the anchor questions were selected at the time the bookmark was set. There are many ways used to maintain a constant b/θ-scale. Based on the field test data he saw, the author believes that the following method was used to maintain a constant b/θ-scale.

> *Method of Maintaining the Scale:* First, one determines the b-values of test questions (how this was done is unknown) without considering the anchor questions at all. Second, one determines how much the average of the p-scores of the anchor questions has changed on the current field test from this average on the original standard-setting field test. Use this p-score change to compute a b-value change (e.g., using the line relating b-values and p-scores; see Figure 2). Shift all the test questions' b-values by an amount equal to the change in the average anchor b-value.

5. WHAT WENT WRONG WITH MAINTAINING THE PERFORMANCE STANDARD. Section 4.5 presented the way that Item Response Theory attempts to maintain a constant performance standard over time. As the Math A tests got harder, the equating mechanism to maintain a constant scale should have lowered the passing score. However, the passing score actually rose, from 43 in 1999 to 51 in 2003. So, what went wrong? The answer is almost everything. The analysis is broken into six parts: setting the bookmark, performance of anchor questions, test construction, field test design, field test psychometrics, and equating calculations.

State Education departments do not have the staff or expertise to develop standards-based tests. At the time of the Math A Panel, New York had about 60 staff in its test development division who were responsible for about 60 different tests (different subjects, different grade levels). The mathematical expertise of the staff consisted of two former high school teachers. There was one psychometrician in charge of the testing program, who was too busy with administration and discussions with test construction

contractors to look for the problems discussed here. The staff were totally dependent on psychometric contractors to develop the tests and do the IRT analysis.

When faced with the massive failure rate on the June 2003 Math A test, New York State Education officials blamed an "avalanche" of weak juniors and seniors who had put off taking the new Math A test, after having failed the old, easier Math I test in their sophomore year. This defense collapsed when the Math A Panel found that Honors freshman (taking the test a year early) had over double the failure rate of the previous year.

5.1. Setting the Performance Standard. As discussed in Section 3.4, the old Math I, II curriculum was the source of questions for setting the performance standard and these questions were worked by students who learned from that curriculum. Thus the Math A performance standard for the Math A test was at best a forward-looking extrapolation, set *before* teachers started to develop a syllabus for the Math A course. The bookmark-setting process should have used as input field test data showing how students taking the Math A curriculum performed on Math A problems. This was a classic chicken-and-egg problem: students need some sort of instruction on the types of problems on which they will be tested to provide input to setting the performance standard, but their instruction cannot be planned until the performance standard is set.

The standard-setting process relied in large part on the p-scores (percentage of students getting the right answer) of the questions used to set the bookmark. These p-scores came from the 1998 field test. The Math A Panel saw data indicating that the p-scores on the 1998 field test of questions on the June 1999 Math A test were very close to the p-scores on the actual June 1999 test. It was well known among teachers that only better students took the first (optional) Math A test in June 1999. So, the students used in the standard setting were clearly above average. Recall that the performance standard required a $2/3$ probability of success on the bookmark question, yet only 55% of those 1998 field test students got the question correct.

5.2. Anchor Questions. The Math A Panel was told that a subset of questions in the initial field test, used to set the bookmark, were chosen to be the anchor questions. The Math A Panel saw the anchor questions and confirmed that they were easier and more routine (procedural)—based on the previous Math I and II courses—than typical questions on subsequent Math A tests. This was a critical flaw that prevented the anchor questions from detecting the improving ability of students on Math A questions (documented below). Students did not perform better on the anchor questions, presumably because some anchor questions assessed Math I, II skills that were de-emphasized in the Math A course.

A further problem was that in 2001, half of the original 35 anchor questions were dropped. While State Education staff did not know the reason for this action, it is assumed that the decision was made because these questions were performing poorly by psychometric measures. This change in itself is troublesome, but more significantly, most of the discarded anchor questions were written-response questions. There were only three written-response questions among the remaining 18 anchor questions. On the one hand, because of the value judgments by graders in interpreting the scoring rubrics for written-response questions, they are inherently less reliable than multiple-choice questions. On the other hand, the written-response questions constitute the majority of points on the Math A test. Thus, the anchor items became even more disconnected from the types of questions on the Math A test.

As a result of the Math A Panel report, new anchor questions were chosen that are better aligned with current Math A test questions. However, the problem still remains

that written-response anchor questions have questionable reliability because of their subjective grading.

5.3. Construction of Tests. The Math A Panel was given no information about the pre-testing of potential test questions or how the tests were assembled from the pre-test results, beyond the requirement for given percentages of questions from each of the seven Key Idea areas. However, in looking at the p-scores of test questions from field tests, it is clear that there was an effort to keep a constant average p-score of the questions on a test. The average p-score on each test was about .47. Quite possibly, the p-scores ran a little higher on the pre-tests, so that the average p-score of a test's questions was aimed to be about .50, based on pre-test data. However, the harder written-response questions carried more weight, and so the weighted average p-score was more like .45 or lower, meaning an average score of 38 or lower out of 85. For most students to be able to pass the test, the passing score would have to be under 30. Actually, if a constant performance standard had been maintained from June 1999 to June 2003, the passing score in June 2003 probably would have been under 30.

How does one reconcile constant average p-scores for tests over time with the claim that the tests were getting harder? The Math A Panel members judged that the multiple-choice questions had gotten a little harder over time, but their average p-score in field tests had increased from .52 in 1998 to .59 in 2002. Accounting for the greater difficulty, the p-score improvement on multiple-choice questions was more like .10, a very significant improvement. Keeping the average p-score on the tests constant meant that much harder written-response questions were needed to balance the higher p-scores of multiple-choice questions.

5.4. Field Tests. The State Education Department depended on the goodwill of schools and teachers to run field tests. Many schools did not want to spend the class time or effort to participate. The result was that the sample sizes were sometimes too small for good statistical validity. While the group of high schools asked to participate in each field test was a demographically representative sample of NY high schools, there was limited data on the schools and no data on the students that did participate. As with the choice of students chosen for the standards setting, State Education staff worried that schools gave the tests to honors students, fearing that the test results would be used to evaluate the schools.

To fit into 45-minute class periods, a three-hour, 35-question test was broken into three 16-question subtests for field testing (the tests typically take most students well under 3 hours to complete and so 45 minutes is adequate time to complete 16 questions). In total, a Math A field test involved 16 subtests: for each of the four complete tests for the coming year, there were three subtests, and also there were four copies of the 18-question anchor test, one to be matched with each of the four Math A tests. Each subtest was taken by 250 to 600 students. Breaking each Math A test into three (overlapping) subtests is probably why there was not an exact functional relationship between the p-scores and b-values in Figure 2.

Breaking a 35-question test into three 16-question subtests involves troublesomely thin overlap for aligning the subtest b-scales into the b-scale for a whole test (especially if several of the common questions are hard written-response questions on which most students have very low scores). Another big problem was that anchor items were not mixed in with the test questions. The anchor questions have better psychometric consistency when mixed in with the test questions, but there was not enough time to do this. The separate anchor subtest appears to be the reason why the equating method described in Subsection 4.4 was used.

State Ed staff knew that the field tests were poorly designed. Field test performance was historically weaker than actual exam performance when the test's high stakes made students prepare better and try harder. For example, up to 15% of the easiest problems on field tests were left blank.

Because of the flawed structure of field tests, State Ed staff claimed that they saw no cause for alarm in the low scores on the June 2003 Math A test during 2002 field test (the average score on the field test was 36, while passing was 51). Yet they believed that field test data were an appropriate basis for psychometric equating and the calculation of questions' *b*-values.

5.5. Psychometric Problems in Field Tests. The Math A Panel was shown *p*-scores, *b*-values, and other psychometric measures for each question on every field test. There were many technical inconsistencies. In the 2000 field tests, the average *p*-score on the 18 anchor questions was .76, while in every other year the average was between .62 and .67. Within a particular field test, the *p*-scores of some anchor questions varied considerably among the four copies of the anchor test. Similarly, for a particular test, questions that were common to all subtests had substantially varying *p*-scores among the subtests. In 2002, there were large variations in the average unadjusted *b*-values among the 16 subtests that made no sense at all. The psychometric expert on the Math A Panel warned that such anomalies were likely to be signaling very serious problems with the field test data and the validity of the underlying psychometric curve-fitting. One way to show the defective calculation of *b*-values is to look at the breakdown in the functional relation between *p*-scores and *b*-values for questions; see Figure 3.

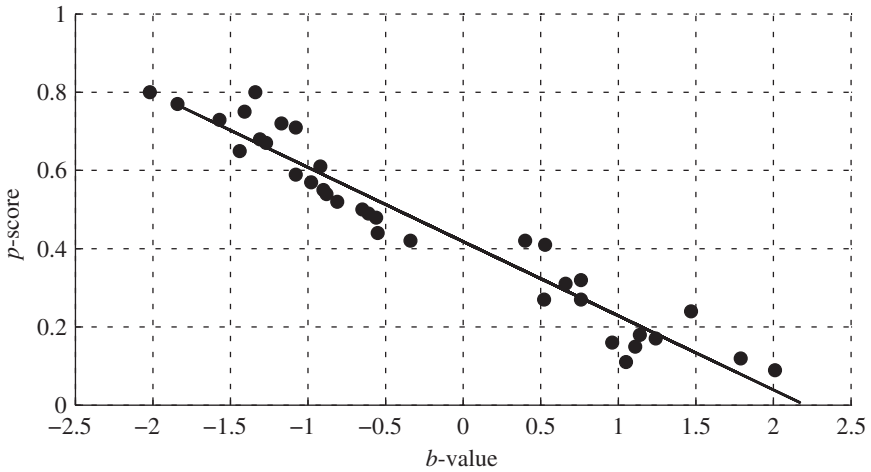

Figure 3. June 2003 test *b*-values versus *p*-scores, based on 2002 field tests.

These field test anomalies were unrelated to the structural problems noted above with the performance standard and anchor items. However, the field test data were so flawed that the *b*-values of questions were totally unsuitable for a high-stakes test.

Following recommendations from the Math A Panel, New York now requires a demographically balanced, statistically valid sample of schools to participate in field tests. The field tests now involve complete 35-question tests administered a year and a half earlier in January during the time slot when the real January Math A test is given. The field tests use students who take the Math A course during their first three

semesters of high school but delay taking the Math A test until June (such a delay is common; these students take the first semester of Math B in the interim). Teachers are encouraged to use the field tests as part of the course grade. In another change, the final determination of the score to pass a Math A test is now made after the actual test is given and is based on a statistically valid sample of students' performance on each question of the real test.

5.6. Equating Calculations. Here is where the most significant problems with the Math A tests arose. Based on the two flaws discussed in this subsection, by June 2003 the passing raw score should have dropped by at least 9 points from its June 1999 value of 43. Instead it rose by 8 points—a stunning 17-point error. Other factors discussed shortly suggest that the passing score should have been even lower, probably below 30.

5.6.A. Shift in b-values. The most blatant error in equating involved a change in the scale for the b-values, which occurred when psychometric contractors were switched in 2000. The new contractor assumed the scale for the b/θ-values had been chosen so that the average of questions' b-values was 0. As noted above, on the initial field test, the scale was chosen so that $0 = b_{\text{bookmark}}$, the b-value of the bookmark question. With the original choice of 0 for the b-scale, the average b-value of questions turned out to be .46 on the first Math A test. When the new psychometric contractor used a b-scale where 0 was the average b-value, this change lowered b-values of all future test questions by .46. In a major mistake, the bookmark b_{bookmark} and the associated benchmark $\theta_{\text{benchmark}}$ were not similarly lowered from 0 to $-.46$ and from .7 to .24, respectively. This error had the effect of raising the passing score by about 7 points (this number is the sum of the differences in the expected number of points on each question when the benchmark was not changed). Because of the other problems with the field test data and anchor questions, the passing score increased by only 4 points from June 2001 to June 2002. However, the "high passing" score, which had been fairly constant around 64, jumped by 8 points to 72 (and stayed at that level for subsequent tests). Somehow, no one noticed the sudden jump. The error and the scale shift that caused it were only uncovered years later by this writer.

5.6.B. Errors in b-values. As noted above, the multiple-choice problems were getting slightly harder. However, a combination of misperforming anchor questions (which included out-of-date topics no longer emphasized), inconsistent field test data, and possibly other psychometric problems caused the average b-value of multiple-choice questions to drop instead of increase. The drop was dramatic: from $-.26$ in June 1999 to $-.88$ in June 2003 (the 1999 value is adjusted for the .46 change in b-scale so that both numbers are on the same scale). Given that the multiple-choice problems were judged to be getting harder, their average b-value should have increased, as an estimate, say, to $-.15$. So the b-scale was too low by approximately .70. Using a linear extrapolation based on the fact that the .46 scale shift (discussed in the previous paragraph) made the passing score 7 points higher than it should have been, this new shift of the b-scale by .70 made the passing score another 10 points higher than it should have been. Combining the two b-scale errors yields an estimated upward shift of 17 points in the passing score.

5.7. The Evolving Bi-modal Distribution of *b*-values. Over time the p-scores of the 3-point and 4-point written-response questions dropped (the questions got considerably harder) so as to keep the average p-score constant as the p-scores of multiple-

choice questions rose. Correspondingly on the b-scale, where the overall average b-value had to be 0, the average b-value of (2-point) multiple-choice questions decreased from $-.26$ to $-.88$ from June 1999 to June 2003, while the average b-value of the 3- and 4-point written-response questions increased correspondingly from $+.5$ to $+1.1$. The breakdown of average b-values by the four sections of the test was as follows.

Test	Aver. Mult. Choice	Aver. 2-pt Written	Aver. 3-pt Written	Aver. 4-pt Written
June 1999	$-.26$	.14	.14	.76
June 2003	$-.88$	$-.30$	.74	1.43

The original June 1999 b-values have been decreased by .46 in this table to compensate for the 2000 shift in the b-scale.

Figure 3 shows the plot of b-values versus p-scores for the June 2003 test. In the June 2003 plot, there are only two questions with b-values between $-.5$ and $+.5$, while there are 10 questions with b-values in the corresponding unit interval in June 1999 (see Figure 2). Common sense tells one that a bi-modal distribution like this is inherently undesirable in any test. In this case, the gap around 0 has a further flaw. The original performance standard was based on a bookmark question with difficulty $b = 0$. (Despite the mistaken shift of scale in 2000, the psychometric contractor still believed the bookmark's b-value was 0). *It is nonsensical to estimate a student's probability of success of a bookmark question by asking questions that are much easier or much harder*. Further, the computation of the passing score, that is, the hypothetical score of a benchmark-ability student, becomes excessively dependent on the assumed form of the item response curves.

Finally, observe that if the June 2003 test questions had been used to set a new bookmark, it would have been impossible to select a bookmark of difficulty close to $b = 0$, because there were no questions with those b-values. The reality is not much different. When a bookmark was set in December 2003 for a new New York graduation test using a sample of recent test questions, the bookmark question chosen had a b-value that was .25 less than the next harder question (source: the writer's brother was on the new standard-setting committee). If the next harder question were chosen for the bookmark, the bookmark would have increased by .25, which translates into an increase of 3 or 4 points in the passing score on future Math A tests, and this change in turn translates into a likely decrease of 8–10% in the percentage of students who would pass the test. Given the imprecision of setting a bookmark discussed above, a much better choice of questions should have been available near the desired performance standard.

6. CONCLUDING REMARKS. When asked to be a member of the Math A Panel, this reviewer assumed that the task would consist of making subjective judgments about the difficulty of problems on the June 2003 Math A test in comparison to earlier Math A tests. He also hoped, like other mathematicians who have looked at state mathematics tests, to have a chance to critique the level of mathematical precision of some test questions. In reality, the analysis of the high failure rate on the June 2003 Math A test turned out to involve a vast array of broader issues that were never anticipated, and mathematical precision moved down the list of priorities.

The Math A Panel was given piles of information to sort through, but much critical information was missing or was withheld, and had to be inferred indirectly. Missing

information included how the b-values of test questions were determined and how year-to-year equating was performed. In the end, the only useable information was the p-scores and b-values of test questions and anchor questions from the field tests, along with the value of b_{bookmark}. This information supplied a huge pile of "dots." Figuring out which dots to use and where to draw the "lines" connecting the dots to form the "picture" took this writer much of a year of intense effort.

The ideal of establishing and consistently maintaining an appropriate performance standard in mathematics at different grade levels may not be attainable if more demanding curricula are being introduced. The Math A experience certainly gives one pause. Nonetheless, standards-based tests are now federally mandated. Thus, it is important for mathematicians to question publicly all claims of scientific precision in state standards-based tests and to make sure that colleagues who are chosen to serve on state mathematics testing committees are aware of both the basic psychometric theory of standards-based tests and huge errors that poor applications of this theory can produce.

ACKNOWLEDGMENTS. The author would like to express his appreciation to Charles Lewis, the paper's psychometric reviewer, for his many helpful corrections and suggestions, to fellow Math A Panelists Gregory Cizek, Franco DiPasqua, Andrew Giordano, Lidia Gonzalez, Robert Gyles, Daniel Jaye, Sophia Maggelakis, Theresa McSweeney, Alfred Posamentier, and Katherine Staltare, and most of all to the Panel's outstanding chair, William Brosnan. The panelists struggled together. Special thanks are due to the helpful staff at the New York State Education Department who assisted the Math A Panel: Jim Kadamus, Anne Schiano, Jerry DeMauro, Jackie Marcano, Gretchen Maresco, Terry Calabrese-Gray, and most of all, Tom Sheldon.

REFERENCES

1. F. Baker, *The Basics of Item Response Theory*, ERIC Clearinghouse for Assessment and Evaluation, University of Maryland, College Park, 2001; also available at http://eric.ed.gov/PDFS/ED458219.pdf.
2. F. Baker and S.-H. Kim, eds., *Item Response Theory: Parameter Estimation Techniques*, Marcel Dekker, New York, 1992.
3. G. Cizek, ed., *Setting Performance Standards*, Lawrence Erlbaum, Mahwah, NJ, 2001.
4. F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Mahwah, NJ, 1980.
5. Math A Panel Report to the New York Board of Regents, October 2003, available at http://www.regents.nysed.gov/meetings/2003Meetings/October2003/1003brd3.htm
6. W. J. Van der Linden and R. K. Hambleton, eds., *Handbook of Modern Item Response Theory*, Springer, New York, 1997.

ALAN TUCKER received his Ph.D. from Stanford in 1969 and since 1970 has been at SUNY-Stony Brook. His family has a long involvement in the MAA and mathematics education: his father A. W. Tucker and grandfather D. R. Curtiss were MAA Presidents; Alan and his brother Tom have been MAA First Vice-Presidents.
Department of Applied Mathematics and Statistics, SUNY-Stony Brook, Stony Brook, NY 11794-3600
atucker@notes.stonybrook.edu