

## The New York Regents Math Test Problems, by Alan Tucker, SUNY-Stony Brook

### 1. Introduction

Current efforts to set higher standards for the mathematical skills of all K-12 students face a gauntlet of undesirable side effects that undermine their intended benefits. 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. There is a danger that the response to such tests may be instruction focused on ‘teaching to the test’—excessive time spent drilling on practice tests; or instruction that treats the test questions as an upper bound on learning instead of the intended lower bound. A more subtle problem is that it is very difficult to set, and maintain over time, meaningful performance standards in school mathematics. This problem is the focus of this article.

New York State has a long tradition of high-stakes Regents tests that have been generally well received. For decades, college bound students have qualified for state funded college scholarships by earning a test-based Regents diploma. In the late 1990s, the Board of Regents mandated that *all* NY high school graduates had to pass five Regents exams, in English, math, U.S. history, world history, and science. As the initiative was being phased in, initial results were encouraging with high passing rates on all five graduation tests. Also, presaging the No Child Left Behind Act (NCLB), 4<sup>th</sup> and 8<sup>th</sup> grade assessments of student learning in mathematics and English were instituted to monitor the performance of individual schools and measure progress towards meeting the graduation requirements.

A new harder three-semester\* Math A course and test was simultaneously being developed in the 1990’s to replace the old two-semester Math I course and test, which was dominated by one-step, ‘mechanical’ problems. The new test covered more topics, had more multi-step problems, and couched the majority of problems in ‘real world’ contexts. (No significant changes were made in the other graduation tests.) The New York Math A test would be the hardest mathematics graduation test in the country. To phase it in, the Math A test was offered from 1999 to 2002 as an alternative to the Math I test with the passing score set at 55 instead of 65.

While mathematicians have criticized the precision or appropriateness of some questions on this test, most Math A test questions are reasonable for a graduation test for all students. Interested readers can download the past tests from [www.nysedregents.org/testing/hsregents.html](http://www.nysedregents.org/testing/hsregents.html). A number of the questions are quite challenging for such tests. If one compares, say, the January 2003 and June 2003 tests, one sees that the questions vary considerably from test to test. This raises the critical issue of how does one maintain a constant performance standard for passing on a series of tests whose questions vary from year to year.

In June 2003, a year after the old Math I test was eliminated, the Math A test made headlines with its high failure rate, estimated at 70%. The New York Commissioner of Education set aside the Math A test scores of juniors and seniors and appointed a special Math A Panel to examine the cause of the poor test results and make recommendations to avoid future problems. This writer was a member of that panel. The Panel’s findings [Math A Panel] are an instructive lesson in the challenges that face efforts to bring higher standards to school mathematics learning. The Panel’s investigations also illustrate the range of ancillary issues that confront NCLB-mandated, standards-based mathematics tests.

---

\* In reality, weaker students are given Math A over four semesters, with further Math A instruction if they fail the Math A test.

## 2. Design of Math A

The Math A course had 32 content standards that were developed by high school mathematics teachers and State Education Department staff. They were quite similar to content standards in other states' mathematics graduation requirements. Math A, and a sequel Math B course for college-bound students, build on precursor content standards for grades K-8.

These 32 standards were further broken down into 103 sub-indicators-- specific concepts, techniques or types of problems (see [www.emsc.nysed.gov/ciai/mst/pub/matha&b.pdf](http://www.emsc.nysed.gov/ciai/mst/pub/matha&b.pdf)). Performance standards for the mastery of the content standards are measured by the Math A test. The test is 3 hours long and initially contained 20 multiple choice problems, each worth 2 points, and 15 free response questions (which require written answers), five worth 2 points each, five worth 3 points each, and five worth 4 points each.

A Mathematics Resource Guide for teachers was created that gives a sample question, and in some cases a sample classroom exercise, for the 32 content standards (see previously cited URL). A sample complete test was also released, along with examples of student work earning different amounts of partial credit on the test's free response questions. With this information, teachers at individual schools were supposed to create their own curricula for realizing these content and performance standards. Experienced teachers were enthusiastic about the chance to build their own curricula. The test, in theory, would just confirm that these standards were met by most students. Most of the sample problems in the Resource Guide and sample test did not look too different from the questions associated with the first half of the previous Math I, II, III course sequence.

## 3. Problems with Content Standards

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, Operations, and Patterns/Functions— and non-traditional, cross-cutting topics, such as Mathematical Reasoning and Modeling/Multiple Representations. While the seven key ideas worked fairly well at lower grades, important topics at the high school level did not fit neatly into the seven key ideas. 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. For example, problems about functions could arise in modeling word problems (key idea Modeling), in measuring geometric figures with variable sides (key idea Measurement), and in operations on polynomials (key idea Operations), as well as under its own key idea Patterns/Functions.

A compounding problem was that many of the content standards covered considerable ground. For example, standard 6C under key idea Uncertainty was:

**6C. Use the concept of random variable in computing probabilities.**

This standard included the following four sub-indicators:

- *Mutually exclusive and independent events*
- *Counting principle*
- *Probability distributions*
- *Probability of the complement of an event*

While the first and last of these four sub-indicators were relatively specific, each of the middle two sub-indicators could occupy weeks of class time, if done in moderate depth. Recall that there are about 16 weeks of instruction in a semester and so over three semesters each of the 32 content standards of Math A would have received on average 1 1/2 weeks of instruction.

Some content standards cut across a wide swath of the curriculum, overlapping other standards. Standard 4A in key idea Modeling/Multiple Representations was:

**4A. Represent problem situations symbolically by using algebraic expressions, sequences, tree diagrams, geometric figures and graphs.**

The first three sub-indicators of this standard were:

- *Use of variables/algebraic representations*
- *Inequalities*
- *Formulas and literal equations*

The first sub-indicator subsumed a large segment of Algebra I. The next nine sub-indicators for standard 4A ranged across some of the basic topics in a geometry course, such as:

- *Parallel and intersecting lines and perpendicular lines*

. The final sub-indicator was:

- *Sample spaces: list of ordered pairs or n-tuples, tree diagrams.*

The Mathematics Resource Guide had one sample question for standard 4A, which involved two figures of a ladder 10 feet long leaning against a wall. In the first figure, the ladder's bottom was 8 feet from the wall, and in a second figure the ladder was moved so that its bottom is 4 feet from the wall. Students were asked to determine how much farther up the wall the top of the ladder moved. Clearly, the problem was indicative of only a tiny fraction of the types of Math A test questions associated with content standard 4A and gave teachers little guidance of what to teach students under standard 4A.

**4. Problems with the Performance Standards.**

Difficulties caused by the wide range of topics associated with individual content standards were compounded by the increasing difficulty of test questions over time. A standards-based test is supposed to stay the same, but the free response Math A questions became noticeably harder over time. Standard 5A was a good illustration. It stated:

**5A. Apply formulas to find measures such as length, area, volume, weight, time, and angle in real-world situations.**

The four performance sub-indicators for 5A were:

- *Perimeter of polygons and circumference of circles.*
- *Area of polygons and circles.*
- *Volume of solids.*
- *Pythagorean theorem.*

The range of test questions associated with standard 5A on recent Math A exams is shown in the following table.

**Table of Questions Associated with Math A Content Standard 5A**

Mathematics Resource Guide's Example of Standard 5A	June 2002 Math A Question based on Standard 5A	Aug. 2002 Math A Questions based on Standard 5A	January 2003 Math A Question based on Standard 5A	June 2003 Math A Questions based on Standard 5A
<p>Ms. Brown plans to carpet part of her living room. The living room floor is a square 20 feet by 20 feet. She wants to carpet a quarter-circle of radius 20 ft. as shown. Find to the nearest square foot the amount of the floor that will be uncovered.</p>	<p>31. As shown, a person can travel from New York City to Buffalo by going north 170 miles to Albany and then west 280 miles to Buffalo.  <i>a</i> If an engineer wants to design a highway to connect New York City directly to Buffalo, at what angle, <math>x</math>, would she need to build the highway? Find the angle to the <i>nearest degree</i>.  <i>b</i> To the <i>nearest mile</i>, how many miles would be saved by traveling directly from New York City to Buffalo rather than by traveling through Albany?</p>	<p>31. In the accompanying diagram, <math>x</math> is the length of a ladder that is leaning against a wall of a building, and <math>y</math> is the distance from the foot of the ladder to the base of the wall. The ladder makes a <math>60^\circ</math> angle with the ground and reaches a point on the wall 17 feet high. Find the number of feet in <math>x</math> and <math>y</math>.</p>	<p>30. A rectangular garden is going to be planted in a person's rectangular backyard, as shown in the diagram [on the next page]. Some dimensions of the backyard and the width of the garden are given. Find the area of the garden to the <i>nearest square foot</i>.</p>	<p>30. To get from his high school to his home, Jamal travels 5.0 miles east and then 4.0 miles north. When Sheila goes to her home from the same high school, she travels 8.0 miles east and 2.0 miles south. What is the measure of the shortest distance, to the nearest tenth of a mile, between Jamal's home and Sheila's home? [The use of the accompanying grid is optional.]</p>
		<p>35. Determine the distance between point <math>A(-1,-3)</math> and point <math>B(5,5)</math>. Write an equation of the perpendicular bisector of <math>AB</math>.</p>		<p>34. A straw is placed into a rectangular box that is 3 inches by 4 inches by 8 inches- see diagram. If the straw fits exactly into the box diagonally from the bottom left front corner to the top right back corner, how long is the straw, to the nearest <math>1/10</math> of an inch?</p>
				<p>2. (Multiple Choice) The accompanying diagram shows a square with side <math>y</math> located inside a square with side <math>x</math>. Which expression represents the area inside the larger square but not containing the smaller square?            (1) <math>x^2</math>      (3) <math>y^2 - x^2</math>            (2) <math>y^2</math>      (4) <math>x^2 - y^2</math></p>

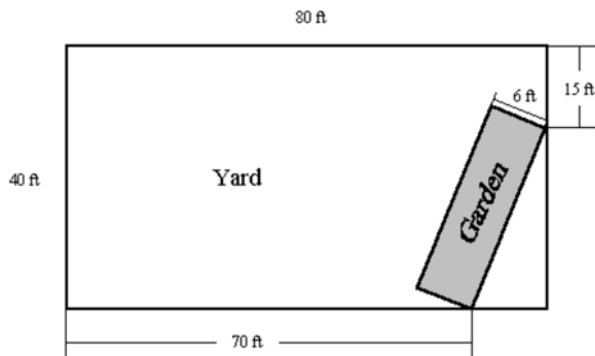


Diagram for Question #30 on June 2002 Math A Test

Compare the routine question in the Mathematics Resource Guide for 5A with question #31 in June 2002 (involving the arctan) and question #34 in June 2003 (involving a 3-dimensional distance; to be solved by two iterations of the 2-dim. distance formula, which is all that students know). This table illustrates how hard it was for teachers to plan what difficulty of problem-solving and what breadth of instruction to give to their students for a content standard.

Note that the earlier Mathematics Resource Guide sample question for standard 4A about the ladder is an application on the Pythagorean Theorem, and so would more properly be classified under standard 5A.

Along with the increasing difficulty of questions, teachers also had to contend with a lack of consistency in the coverage of important topics. The only constraint on the test construction was a specification giving the approximate percentage of questions that came from each of the seven key ideas. However, because of the breadth of key ideas cited above, the number of questions devoted to basic topics such as trigonometry was quite variable. The June 2003 test had no questions involving trigonometry while all previous tests had had several trig questions. At the other extreme, a 35-question test could have two or more questions associated with the same particular sub-indicator.

Summarizing the situation for the content standards and performance standards, there came to be a huge variety of challenging questions possible within the 103 content sub-indicators, with little predictability. Teachers faced a daunting challenge preparing students in the lower half of their classes to pass a Math A test. These uncertainties also made the test a stimulating, fair assessment for more capable mathematics students. A “high pass”-- over 85— marked a student highly proficient in the first half of the high school mathematics curriculum.

To correct this problem, the Math A Panel recommended that a well-defined curriculum be developed for the Math A course on which the Math A test would be based, as opposed to the original situation in which the course was based on the test. Further, the Panel recommended that the Math A course be reduced to a year’s length, since there was too much material in the original Math A course to be covered in a 3-hour test. This recommendation affects the Math B course, too. The two courses would be split into three one-year courses. A committee is now developing a new high school mathematics curriculum for New York.

## 5. Setting a Performance Standard.

The rest of this article examines the psychometric theory for setting and maintaining a performance standard and what went wrong with the implementation of this theory for the Math A test. Some of the problems were specific to New York's Math A test, but could occur on other state mathematics tests. Some of the problems appear to be inherent weaknesses in the theory of standards-based testing.

Many of the problems in standards-based testing presented here do not appear to have been known to policy makers at the state and national level when standards-based tests were proposed to monitor the performance of America's schools and their students. Standards-based state tests are fairly new, and state education departments have limited psychometric expertise.

It was hard to anticipate problems with a new mode of testing. Only after disasters like the high failure rate on the June 2003 New York Math A would education officials start to examine carefully how standards-based tests work in practice (earlier, Massachusetts had a similar disaster with the first administration of its mathematics graduation tests; in spring 2004, Oregon had a high failure rate on its initial mathematics graduation test). What is special about the New York Math A test is the unprecedented access to confidential data and the full cooperation that the New York State Education Department gave to the Math A Panel.

**5.1. Setting the Performance Standard.** One gives a collection of questions to a sample of students. The questions are ranked by their p-score, the percentage of students who get them right. A group of mathematics teachers and professors go down the ranking, from easiest to hardest, looking to set a 'bookmark' at a question judged to be of a difficulty that someone meeting the desired performance standard would get right, say, 2/3rds of the time. 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 2/3rds of the time by a student at the borderline for passing. For more about setting performance standards, see [Cizek].

The use of a single question to set a performance standard for assessing a whole year (or more) of study in mathematics is offensive to most mathematicians. An assessment of written English may involve skills that are fairly universal across a range of writing samples (although this writer is not qualified to make such a claim). There is obviously a major component of underlying mathematical ability that is reflected in students' performances on any question, but specific knowledge is also important. Teachers may cover some topics thoroughly and others too superficially for any student to learn key ideas. A minimum score on a representative collection of questions would be a far better basis for a performance standard.

**5.2. Subjective Nature of the Judgment.** A well known problem with setting a performance standard is that different groups of experts have quite different standards (see the article in [New York Times] on the variability in state mathematics standards). The Math A panel reviewed the performance standards for mathematics graduation tests in a number of states and found great variation. New York's was judged to be clearly more demanding than other states'. Massachusetts is one of the few states with above average performance on the NAEP 8<sup>th</sup> mathematics test (New York is rated average). What is the basis for New York having a higher performance standard than Massachusetts in mathematics?

There are many factors, involving the choice of experts and the choice of questions used in the standards setting process and other inputs, that greatly affect the difficulty of the performance standard set by the committee of experts. These factors are probably the real reason that New

York has the highest mathematics graduation standard in the country. The Commissioner and the Regents, in their oversight responsibility for the SED operations, control these factors. Thus, they have a greater influence in how the performance standard is set than the actual standard setting committee of experts. However, the Regents and Commissioner do not appear to appreciate this control that SED and they have.

**5.3. Problem in Ordering the Difficulty of Standard-Setting Questions.** There is a major problem in standard setting that may be fairly unique to mathematics. Students' performances in questions used to set the standard are dependent on which questions the students are familiar with (i.e., were drilled on). Mathematics teachers and professors will frequently rank the difficulty of questions in a very different order. When students and teachers have different standards of difficulty, the whole bookmark process breaks down.

This mismatch was apparent to the Math A panelists who reviewed the January 2004 Math A test. We complained that many of the early questions were much harder than later questions. However, SED staff showed us field test data indicating that the test questions were actually ordered by increasing difficulty for students. One suspects that in another year, when students will 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 tend to promote the type of mindless learning that New York wants to de-emphasize.

**5.4. Out-of-Date Questions in Standard Setting.** The performance standard for Math A was set before the Math A course existed. It turns that the questions from which the bookmarked question was chosen were based on the previous Math I, II courses, and the students who worked these questions were taking the old, procedurally oriented Math I, II courses. The standard was actually an extrapolation from the old math skills in Math I, II to higher future expectations in the problem-solving Math A course. Other states moving to more demanding mathematics curricula will face a similar hurdle in trying to set a performance standard before new curricula are in use. Furthermore, a constant performance standard is inherently unsound: a standard which under 50% of the students can pass is probably reasonably connected to instructional goals, but if students improve over time and subsequently 70+% of students can pass the standard, then the standard has become too low and it would force a teacher to focus on instruction aimed at just the weakest students.

**5.5. Confusion about the Two Roles for a Performance Standard.** There is a potential problem with the two different ways that a performance standard can be used. The first, most common use is as an absolute proficiency standard that is fixed at a demanding level. Initially many students, perhaps a majority, may not meet this standard but in time almost all students should meet it. The second use is as a graduation requirement. If expectations for graduation are rising, a graduation performance standard needs to move up with these rising expectations, but it would allow most students each year to graduate. The performance standard for Math A test should have been the second type of standard but in reality was a hybrid. It was set in 1999 as an expectation for future performance with the new curriculum and would not be fully implemented until 2003. However, it was set at a level significantly above the ability of the average 1999 student.

## 6. Maintaining a Constant Performance Standard

**6.1 Overview of The Math A Performance Standard.** The original Math A performance standard was set in 1999. The process of setting a performance standard and inherent problems in this process were described in the previous section. The original Math A performance bookmark was set at the level of a question with p-score (fraction of students correctly answering the question) of .55 on the 1998 field test. The performance standard required a student to correctly answer a question of this difficulty with probability  $\frac{2}{3}$  or better. (The Math A Panel was never shown the wording of this bookmark question.)

Psychometric equating techniques are meant to assure that this performance standard is independent of a particular year's version of the test. To do 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. A scaled score of 65 is supposed to correspond to the performance standard.

Regents 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 pre-tests in year N-2. A set of 'anchor questions' whose difficulty was established at the time of the first Math A field test are included in the field tests every year. Changes in the abilities of the students taking field tests in a given year 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 create a function that maps raw scores into scaled scores for each of the next year's tests so that a scaled score of 65 on a test would match the original bookmark performance for passing.

When the Math A test was first offered in June 1999, as an optional alternative to the Math I test, most test-takers did quite well. On the other hand, in the first three years that the test was offered, only a small fraction of high school students took it and they were believed to be among the best students. The old Math I test was no longer given in June 2002. As the number of test-takers grew in 2002 and January 2003, passing rates declined but were still reasonably high. During the phasing-in period, passing was a scaled score of 55 out of 100. In June 2003, when the passing score moved up to 65, disaster struck. About 70% of students failed at the 65 rate. Even at a scaled score of 55, half the students failed. State Education staff suspected that the high failure rate was due to an avalanche of weak juniors and seniors who had put off as long as they could taking the new Math A test, after having failed the old, easier Math I test in their freshman and sophomores years.

As noted at the outset, the NY Commissioner of Education established an independent Math A Panel to review the June 2003 test. He set aside the June 2003 Math A scores of seniors and juniors. The Panel confirmed that the test questions had become progressively harder over time and were substantially harder than the sample questions in the Mathematics Resource Guide for teachers. There were two aspects of this increased difficulty. One involved test questions. The tests were having more challenging problems (see the table above for standard 5A). In addition, certain categories of word problems appeared on earlier Math A tests in one common format, but in recent tests were presented in an altered format. Less able students misread seemingly easy problems, trying to cast them into the format they knew. The second problem was that the

passing raw score (that mapped to the passing scaled score of 65) was increasing, instead of decreasing, as the tests got harder.

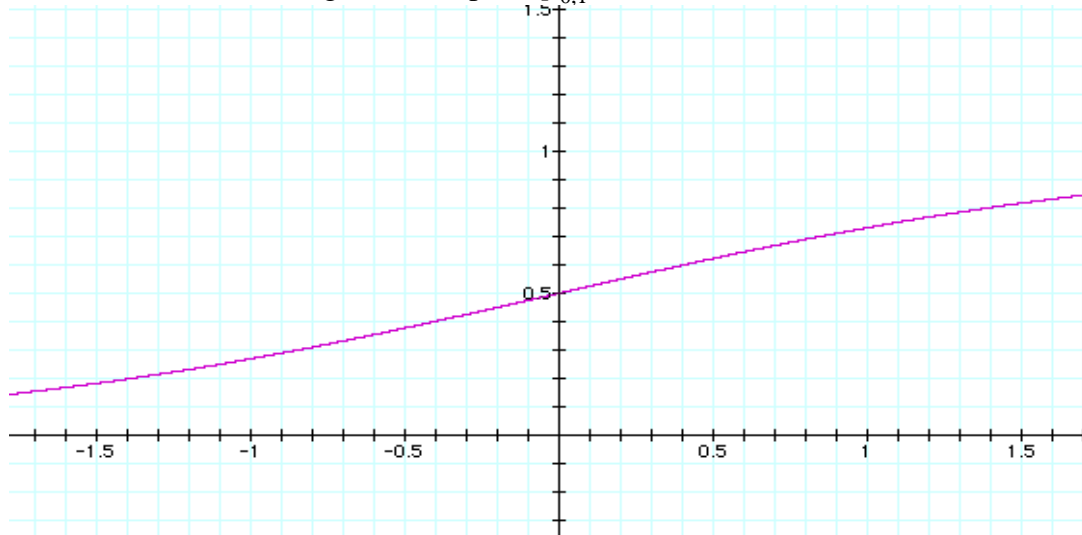
State Education staff confirmed that the increasing difficulty of tests resulted from a conscious effort by the question writers and test assembling committees to increase the difficulty and lack of predictability of test questions over time, as the new Math A course became more established. The high passing rates on the initial Math A tests encouraged this strategy. Many of the State Education staff as well as the teachers who created test questions and assembled tests did not fully appreciate the difficulty of maintaining a constant performance standard on a test whose questions were becoming more difficult.

State Education staff had no explanation for why the passing raw score was rising. This score was obtained from highly technical psychometric calculations performed by an outside vendor. It is these psychometric calculations, whose validity policy makers unquestioningly accepted, that are the big unseen problem with standards-based tests.

**6.2 Introduction to Item Response Theory.** Here is a summary of the relevant psychometric methods that are based on Item Response Theory (IRT) [Baker, Holland]. IRT has a quantitative model for assessing student performance based on the following three components.

- A. *Student proficiency.* IRT assumes that each individual's level of mathematical proficiency can be accurately represented by a single number, called a  $\beta$ -value.
- B. *Question difficulty.* IRT assumes that each test question's difficulty can be described by a single number called the question's  $\theta$ -value.
- C. *Probability of A Student's Success on a Question.* IRT assumes that an item response curve of the form  $p_{\theta,\alpha}(x) = 1/\{1+e^{-\alpha(x-\theta)}\}$  describes the probability of a correct answer by a student of ability  $x$  on a question with difficulty  $\theta$ .

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



Observe that the  $\theta$ 's and  $\beta$ 's are on the same scale. The  $\theta$ -value of a question is defined by the location of the midpoint (50<sup>th</sup> percentile) on the question's response curve. That is, a question is assigned a  $\theta$ -value of  $\beta$  if a student of ability  $\beta$  has a .50 probability of correctly answering the question. The  $\beta$  (and  $\theta$ ) scale is typically centered so that the average  $\theta$ -value of test questions is 0, and the  $\beta$ -value units are measured in standard deviations.

Figure 1 shows the response curve with  $\theta = 0$ ,  $\alpha = 1$ . According to the curve in Figure 1, a student of ability  $\beta = .7$  would have a probability of  $2/3$  of answering a question with difficulty  $\theta = 0$ , and a student of ability  $\beta = -.7$  would have a probability of  $1/3$ . There are other versions of the item response curves, using one or three parameters.

**6.3. Determining Model Parameters.** After a group of students takes a test, maximum likelihood estimation, a form of least squares fitting, is applied to the results to determine an ability rating  $\beta$  for each student and the parameters  $\theta$  and  $\alpha$  in the logistic response curve for each question. In the simplified situation where the  $\theta$ -values and  $\alpha$ -values of the questions have been determined in advance, maximum likelihood estimation works in the following way to determine the  $\beta$ -value of each student. Let  $\theta_i$  and  $\alpha_i$  be the  $\theta$ -value and  $\alpha$ -value of the  $i$ -th question and let  $s_i$  be the score of student  $S$  on the  $i$ -th question;  $s_i$  is 0 (wrong) or 1 (correct) on a multiple choice question and equals the fraction of points earned on a free-response question. The  $\beta$ -value of student  $S$  is the value of  $x$  that minimizes the sum of the squares of the differences between the true score  $s_i$  and the predicted probability of success  $p_{\theta_i, \alpha_i}(x)$ , summed over all questions.

The  $\theta$ -values and  $\alpha$ -values of questions are determined in a similar fashion if the  $\beta$ -values of students are known. Suppose  $\beta_j$  is the  $\beta$ -value of the  $j$ -th student and  $s_j$  is that student's score on question  $Q$ . Then the  $\theta$ -value and  $\alpha$ -value of question  $Q$  are the values of  $x$  and  $y$  that minimize the sum of the squares of the differences between the true score  $s_j$  of the  $j$ -th student on question  $Q$  and the predicted probability of success  $p_{x,y}(\beta_j)$ , summed over all students. When one initially does not know the parameters of either the students or questions, more sophisticated generalizations of maximum likelihood estimation are used to simultaneously determine both student  $\beta$ -values and item response curve parameters ([Holland]).

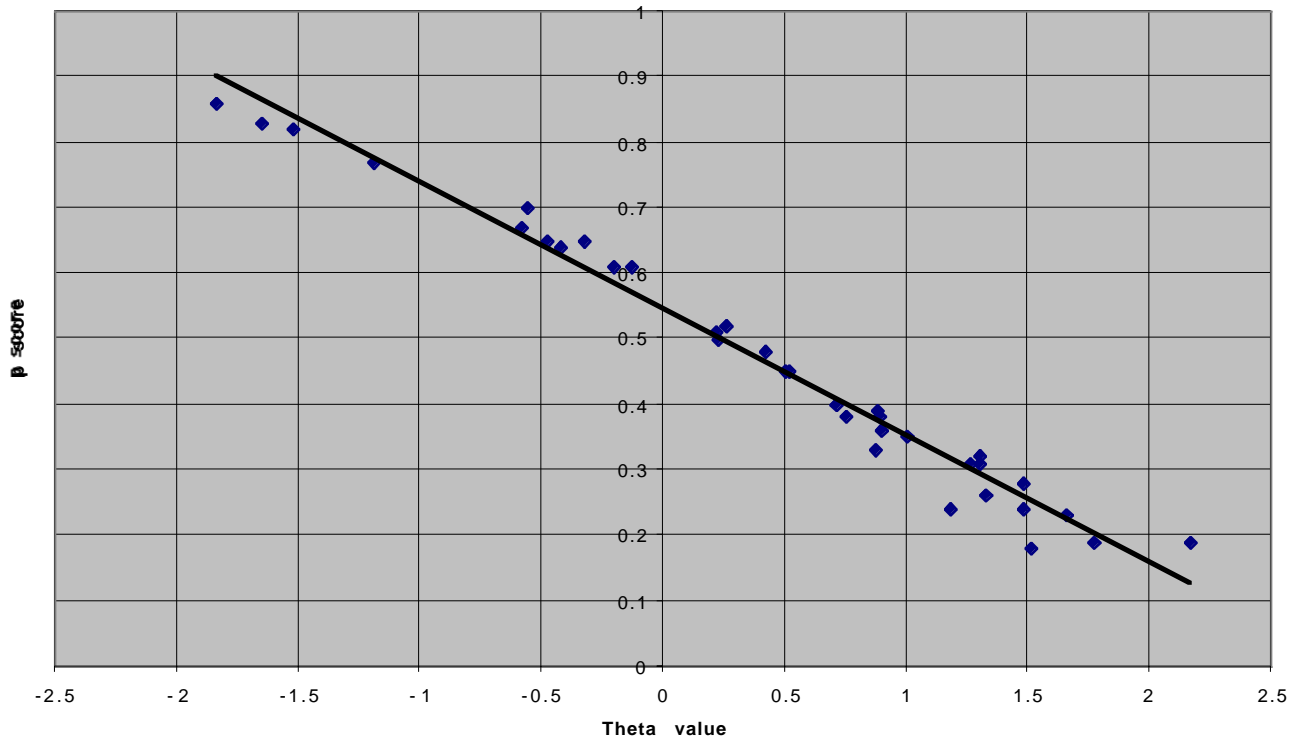
Because the sample sizes in the New York field tests were not large enough, the Rasch IRT model was used in which the scale parameter  $\alpha$  is dropped. The scale parameter indicates the rate at which the probability of a right answer increases as a student's ability increases. Instead, a common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

The critical issue is how well do the Rasch item response curves fit the field test data for Math A. The test statistics that the Math A Panel saw had limited information about the fit of the model, and there was no information about model fit for the June 2003 test. The validity of this assumption could have been judged nicely by comparing the empirical response curve for each question from field test data with the IRT-determined curve for each question. Unfortunately, such data did not exist. Piecing together diverse information about the Math A tests, it is this writer's assessment that the Rasch model fitted field test data reasonably well initially but became increasingly flawed over time. By the time of the 2002 field tests (which included the June 2003 test), much of the IRT data about the tests was becoming inexplicably bizarre.

Note that there is a structural flaw in any IRT response curve. On multiple choice questions, the left end of the curve should not go much lower than  $.25$ , since a very weak student can achieve a probability of success of  $.25$  simply by guessing (there are four possible answers on the Math A multiple choice questions, with no penalty for guessing).

When the IRT model is working properly, the  $\theta$ -values of questions should have an approximately linear relationship with the p-scores (empirical probability of success) of the questions. Figure 2 shows a plot of  $\theta$ -values versus p-scores for the questions on the initial June

Figure 2: June 1999: B-value vs p-score based on 1998 field tests



1999 Math A test (based on 1998 field test data). The regression line fitting the data in Figure 2 is:

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

Recall that the bookmark question had a p-score of .55. Let  $\theta_{\text{book}}$  denote the  $\theta$ -value of this bookmark question. We see from the regression line (\*) that the bookmark p-score of .55 corresponds to  $\theta$ -value of 0; that is,  $\theta_{\text{book}} = 0$ . While normally the  $\beta$ -value scale is set so that the average  $\theta$ -value is 0, the  $\beta$ -value scale for the initial Math A test was chosen so that  $0 = \theta_{\text{book}}$ .

Now we can determine the ability level  $\beta_{\text{perf}}$  of a student who just meets the Math A performance standard. Such a student correctly answers with probability  $2/3$  the bookmark question of difficulty  $\theta_{\text{book}} = 0$ . As noted earlier, from Figure 1, we see that  $p_{0,1}(.7) = 2/3$ . Thus  $\beta_{\text{perf}} = .7$ . Note how important the item response curve for the bookmark question is. If the slope of this curve is too high or too low (compared to the actual student performance in the field test), the flawed curve will lead to an incorrect value for  $\beta_{\text{perf}}$ . This problem would not occur if each question had a separate scale parameter. However, since the field test population was not large enough, the Rasch model (with a common scale parameter for all questions) was used.

**6.4 Equating  $\theta$ -Values Over Time.** Next we look at how these response curves are applied to the field test data. In the field testing for year N-1, data are collected on 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. The  $\theta$ -values of the anchor questions are set on the initial field test and kept fixed from year to year, thereby determining the scale on which the  $\theta$ -values of the questions on subsequent tests are placed. There are two ways to do this.

*Equating Method A:* The fixed anchor  $\theta$ -values are used to determine the students'  $\beta$ -values by maximum likelihood estimation, as described above, using data from the current field test. Then the students'  $\beta$ -values can be used to determine the  $\theta$ -values of the test questions.

Method A has the effect of putting the  $\theta$ -values for the questions on every test on the same scale as the original scale used: i) for setting the  $\theta$ -values of the anchor questions, ii) for the bookmark difficulty  $\theta_{\text{book}}$ , and iii) for the questions on the first Math A test.

*Equating Method B (Less Accurate):* The second way to adjust  $\theta$ -values makes separate calculations about test questions and anchor questions. First, determine the  $\theta$ -values of test questions without considering the anchor questions at all. Second, determine how much the average p-score of the anchor questions has changed on the current field test from the 1998 field test. Use this p-score change to compute a  $\theta$ -value change (e.g., using the regression line relating  $\theta$ -values and p-scores; see Figure 2). Shift all the test question  $\theta$ -values by an amount equal to the change in the average anchor  $\theta$ -value.

Method B is geared towards adjusting for changes in the ability of the students taking the field tests, but does not make any adjustments for changes in the difficulty of the test questions. Method B needed to be used in New York because of the design of the field tests (discussed in section 7.5). To make Method B work properly, the average difficulty of questions must be held constant over time, that is, the average p-score of the questions on each test was approximately the same. These p-scores were based from pre-test data. (Pre-tests are organized similarly to field tests and are used to get preliminary p-scores and also to screen potential test questions according to a variety of statistical measures.). When the average p-score on tests is kept constant, Methods A and B should be equivalent.

**6.5 Determining the Passing Raw Score.** After the adjusted  $\theta$ -values of a test's questions have been determined from field test data, then the passing raw score is determined. This passing raw score is the expected score on the test of a student of ability  $\beta_{\text{perf}}$  who represents the performance standard. Recall that  $\beta_{\text{perf}} = .7$ .

The expected raw score of a .7-ability student on a test is found by summing up the probability of success on each question times the point value for that question. On a question with difficulty  $\theta$ , a .7-ability student's probability of a correct answer is  $p_{\theta,1}(.7) = 1/\{1+e^{-(.7-\theta)}\}$ . Technical note: for free response questions, data from field tests make it possible to devise a more accurate estimate for the expected amount of partial credit a .7-ability student will receive, instead of simply multiplying the expected probability of success times the maximum points on the question. For the expected raw score of a .7 ability student to represent the same level of performance on successive Math A tests, it is essential for the equating methodology to keep the scale of  $\theta$ -values of test questions the same over time.

The passing raw score is used to create the mapping function from raw scores to scaled scores. Recall that the passing raw score is supposed to map to a scaled score of 65. The standard setting process also determined a bookmark for a high pass, which is supposed to equal a scaled score of 85. The high passing raw score is determined similarly to the passing raw score. Suppose  $a$  is the passing raw score and  $b$  is the high passing raw score. Then the mapping function must take  $a$  to 65,  $b$  to 85 as well as 0 to 0 and 85 (the maximum raw score) to 100. The mapping function is usually chosen to be the unique cubic polynomial passing through the four points  $(0,0)$ ,  $(a, 65)$ ,  $(b, 85)$ , and  $(85,100)$ .

*Step back and consider the high stakes associated with the IRT model's item response curves. Can students' abilities and questions' difficulties be characterized with single numbers along some scale? How valid are item response curves? Can one build a reliable assessment around the predicted scores on future tests of a hypothetical student who would correctly answer of a single 'bookmark' question with probability 2/3rds? Finally, if the curriculum is changing, as happened with Math A, won't this scale change requiring constant revision of the performance standard?*

*In sum, the margin of error in the overall process of setting the passing raw score on a Math A test appears to be substantial, although very hard to quantify.*

Finally, we close this presentation of Item Response Theory with a 'reality check.' The thinking underlying a performance standard is very idealistic. The reality is that many students, with their teachers' help, will try to get enough points to pass a standards-based test by drilling on easy problems and selected intermediate and harder types of problems from past tests. How these students would perform on a bookmark-level question is irrelevant to these students and to whether they pass a test.

To make the performance standard truly representative of a particular ability level on a  $\beta$ -value scale, one needs tests with unpredictable questions of varying difficulty, like the Math A tests. But such tests are subject to all the problems discussed in the next section. In particular, designing an anchor set that performs as IRT requires seems almost impossible.

## **7. What Went Wrong with Maintaining the Performance Standard**

The preceding section 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 should have lowered the passing raw score. However, the passing raw score actually rose, from 43 in 1999 to 51 in 2003. So, what went wrong? The answer is almost everything. We break our analysis into six parts: test oversight, setting the bookmark, performance of anchor questions, test construction, field tests, and equating calculations.

**7.1. Test Oversight.** Educational Testing Service has great expertise in developing tests whose scores are comparable over time. There is widespread public acceptance of the consistency of tests such as the SAT's and AP tests. ETS has impressive in-house expertise in testing and it charges a hefty fee on tests to support its high standards. ETS estimates that it spends about \$600 to create a test question. On the other hand, state tests are free for students. The New York State Education Department had suffered massive cuts in its testing staff over the past 15 years. Its number of mathematics specialists had declined from 6 to 1. It is estimated that New York spends \$20 to create a test question for a Regents test.

Almost all components of test development in New York, including the psychometric equating, are contracted out to a variety of specialized testing consultants. For example, when teachers come to Albany to review field test results on free response questions, e.g., to modify the guidelines for partial credit or change the wording of questions, an outside contractor oversees this committee's work following prescribed procedures, while a State Education staffer sits in as an observer. The 40 members of the State Education test development division are responsible for about 50 tests a year, many administered in multiple languages. The State Education staff lack the time to look closely at whether the procedures to maintain a constant standard are behaving properly. Further, staff turnover is a problem. For example, the one psychometrician position has had considerable turnover. Another deficiency is a lack of adequate and timely data

about how students perform on the actual tests. Some of the problems discussed below might have been spotted with more timely reports on test data.

Such a shortage of resources and extensive staff turnover are the norm in most states. Hence, if something is not working right with a state test, as happened with New York's Math A test, no one may spot the problem until it makes newspaper headlines. Given the array of new state mathematics tests mandated by the No Child Left Behind Act and the complexity of psychometric procedures to maintain a constant standard of passing, every state education department needs to have adequate staff in their testing division who understand mathematics and the relevant psychometric methods for standards-based tests.

With education department staff in most states lacking the resources and time to investigate the occasional serious problems that do arise, external advisory committees that aggressively look for anomalies will be critical to the smooth and fair implementation of the new mandated mathematics tests. New York is now considering whether to set up such a committee.

**7.2. Setting the Performance Standard.** Here is where a major structural flaw in standards-based testing arises. The performance standards for the Math A course-- how much students needed to know about each topic-- were to be defined by the Math A test. Thus, the passing bookmark for the Math A test was supposed to be set *before* teachers could develop a syllabus for the Math A course. However, the bookmark setting process required as input field test data showing how students performed on new Math A problems. This created 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. In reality, the set of questions used in the bookmarking procedure, coming out of the initial field tests in 1998, were based largely on what was currently being taught in the old Math I and II courses. The State Education staff and question writers planned to phase in more appropriate Math A test questions over time as the Math A course evolved.

So rather than being an absolute performance standard over time, the initial performance standard was a future expectation involving an extrapolation for Math A learning from student performance on Math I,II problems. Also, the initial bookmark served as a reasonable starting point for the psychometric equating process, that is, to obtain  $\theta_{\text{book}}$ , the bookmark  $\theta$ -value, and from it  $\beta_{\text{perf}}$ , the ability  $\beta$ -value of a student whose expected raw score would be the passing raw score on each test; in this case,  $\beta_{\text{perf}} = .7$ . The steady-state Math A test and its passing standard would be phased in over time. The State Education staff believed that the equating procedures would be able to maintain the original bookmarked performance standard as the difficulty of tests increased. It turned out that the equating method did not make the appropriate adjustments for harder tests.

Another major problem with this phasing-in strategy was that teachers were given inadequate guidance about what the steady-state form of Math A would be. The Mathematics Resource Guide and initial sample Math A test indicated only what would be on the first Math A test and gave no sense of what future Math A tests would look like. There was no written plan that the Math A Panel saw describing what the steady-state Math A test would be like. The evolution of the tests was in the hands of the teachers making up test questions and the directions that they got from State Education staff. At the same time, the passing raw score was evolving in an unplanned fashion due to failures of the equating methodology.

Finally, 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. While there is no direct information about the abilities of these students, there is the following indirect information. 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 of those questions on the actual June 1999 test. There was extensive anecdotal evidence that only better students took the first Math A test in June 1999. Thus, it seems likely that the standard setting relied on data involving above average students. Recall that the bookmark was set at an ability level above the average performance of those above average students in the 1998 field test.

Following the Math A Panel report in fall 2003, New York attempted to address these problems by establishing a new Math A performance bookmark based on student work on Math A questions from recent field tests. However, this time the standard setting problems described in Section 5 became more apparent.

A survey by the Math A Panel of graduation tests in other states found that most of the harder questions are of template, predictable types, e.g., given one basic geometric shape, e.g., rectangle, triangle or circle, with another basic geometric shape inside it, find the area of the larger shape that is not part of the inside shape. Thus, the mathematics graduation standards in other states are largely defined with respect to a specified class of questions. The old Math I graduation test was in this predictable format (focussing on basic algebra skills) and New York teachers knew that virtually any student could be drilled to pass the old test. The decision to raise both the difficulty of problem-solving and the unpredictability of a test's questions was laudable, but, in light of the experiences described here, may have been too ambitious.

**7.3. Anchor Questions.** It was the Math A Panel's understanding that a subset of questions in the bookmark study were chosen to be the anchor questions. While the Math A Panel never saw the set of questions used to set the bookmark, it did see 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 over time (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. Without a change in performance on anchor questions, the equating procedure could not detect the increasing difficulty of the test questions.

A further problem was that in 2001, half of the original 35 anchor questions were dropped. While current 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. This change in itself is troublesome, but more significantly, most of the discarded anchor questions were free response questions. There were only 3 free response questions among the remaining 18 anchor questions. On one hand, because of the value judgements by graders in interpreting the scoring rubrics for free response questions, they are inherently less reliable than multiple choice questions. On the other hand, it was the free response part of the Math A test that was getting harder over time. The Math A course gave increasing emphasis to challenging multi-step word problems in the free-response part of the Math A test. So it is not surprising that the free response anchor questions, based on the old Math I,II courses, misbehaved. But the response of the psychometric

contractor should have been to ask for new anchor questions rather than ‘shoot the messenger’ by discarding most of the free response anchor questions.

As a result of the Math A Panel report, new anchor questions have been chosen that are better aligned with current Math A test questions. However, the structural problems with anchor questions discussed in 5.2 now became evident. Students’ performance on anchor questions is too dependent on how much practice they have had with problems like those in the anchor set. This factor confounds accurate assessment real changes in students’ mathematical ability. In addition, the critical part of the Math A test is the free response questions, and free response anchor questions have reduced reliability because of the subjective component in their grading. *It is questionable whether any static set of anchor questions can form a sound basis for psychometric efforts to adjust for improving mathematical abilities of test takers and associated changing instruction for these improving students.*

**7.4. 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. This 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. Keeping the average difficulty of questions constant over time was critical to the proper functioning of the equating method (Equating Method B) being used.

How does one reconcile constant average p-scores for tests over time with the claim that the tests were getting harder? The answer is crucial— students were getting stronger over time. For example, NAEP data show that in New York and nationally, the percentage of 8<sup>th</sup> graders performing at the ‘proficient’ level arose from 15% in 1990 to 26% in 2000. This pattern could be documented in Math A field tests by looking improving p-scores on the multiple choice questions. The Math A panelists judged that the multiple choice questions had gotten a little harder over time, but their average p-score in field tests had increased by 14% from .52 in 1998 to .59 in 2002. In sum, keeping the average p-score on the tests constant had the effect of making the Math A tests harder at the same rate that students were performing better.

Recall that the passing ability level, having a 2/3 probability of success on a bookmark level questions, was set substantially above the ability of the average student in the original 1998 field test who had a .55 probability of success at the bookmark level. Without the appropriate upward psychometric adjustment of the  $\theta$ -values-- which did not occur because of faulty anchor questions-- making the tests harder in lockstep with students’ increasing ability had the effect of keeping the passing performance the same distance above the performance of the average student and thus guaranteeing a high failure rate.

**7.5. Field Tests.** New York faces an extra hurdle as a truth-in-testing state. The questions on all state-wide tests, and their solutions, must be released two days after a test is given. Field tests were the only venue where standards equating can be done. The State Education Department depended on the goodwill of schools and teachers to run field tests. Many schools refused to participate, not wanting to lose the class time. The result was that the sample sizes were sometimes much smaller than desirable for statistical validity. Further, while the group of high schools asked to participate in a field test was a demographically representative sample of NY high schools, there was limited data on the high schools that actually did participate and which

groups of students within those schools took the test. For example, there was concern among some State Education staff that schools gave the tests mostly to classes of honors students because the schools thought 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). Note that with 16 questions on each subtest, there could be some common questions among the subtests to use to equate  $\beta$ -values among the subtests. 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. Until recent revisions, each subtest was taken by 250 to 600 students.

Breaking a test into three subtests makes it difficult to build a valid scale of  $\theta$ - (and  $\beta$ -) values for the questions on the whole test. There can be up to 6 questions that may appear on all three subtests of a given test, but this is a dangerously thin overlap for combining the  $\theta$ -value scales of individual subtests into a single  $\theta$ -value scale for a whole test (especially when several of the common questions are hard free response questions on which most students have very low scores). Another difficulty is that the anchor subtests are taken by different students from the students taking subtests of the real test questions. This is why Equating Method B (see subsection 6.4) had to be used.

SED officials knew that the field test data was unreliable. For example, they saw no cause for alarm in the low scores on the June 2003 Math A test during 2002 field test (the average was 36, while passing was 51), because field test performance was weaker than actual exam performance. Nonetheless, they used the field test data as input to critical psychometric calculations!

While there is no easy way to assess how accurate the  $\theta$ -values of questions on a test may be when the  $\theta$ -value scale is assembled from the  $\theta$ -value scales of three subtests, there were some clear signs of problems in the field tests arising from inadequate sample sizes and the voluntary nature of which high schools and students participated. 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 varying p-scores among the subtests. In 2002, there were large variations in the average unadjusted  $\theta$ -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 both the field test design and the tests themselves, as certainly happened with the June 2003 test (the other tests that were field tested in 2002 were scrapped after the problems with the June 2003 test).

Because students involved in the field tests knew that these tests did not count for a grade, they did not try as hard, and up to 15% of even the easiest problems on some field tests were left blank. Students had not seriously studied for these tests. Breaking up a full test into three subtests for field tests alters the psychological impact of a long hard test. In sum, numerous deficiencies in the field tests made them a poor basis for the psychometric procedures required to determine a passing raw score.

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 for is being 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 raw 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.

**7.6. Equating Calculations.** Here is where the most significant problems with the Math A tests arose. Based on the types of flaws discussed in this subsection, it appears that by June 2003 the passing raw score was at least 15 points higher than would have been if equating procedures had been working properly to maintain the 1999 passing performance standard.

*7.6.A. Shift in  $\beta$ -values.* The most serious error in equating involved a change in the scale for the  $\beta$ -values, which occurred when psychometric contractors were switched in 2000. In theory, the scale for the  $\theta$ -values (and  $\beta$ -values) is supposed to be set so that the average of questions'  $\theta$ -values is 0. As noted above, on the initial field test, the scale of the  $\theta$ -values was chosen so that 0 was  $\theta_{\text{book}}$ , the  $\theta$ -value of the bookmark level of difficulty. This choice of 0 on the  $\theta$ -scale meant that the average  $\theta$ -value of questions was .46. From the choice of  $\theta_{\text{book}} = 0$  it followed that  $\beta_{\text{perf}}$ , the ability of a student expected to get a bookmark-level problem correct with probability 2/3, equaled .7. Recall that the expected raw score of .7-ability student would set the passing raw score on a test.

In 2000, a new psychometric contractor was employed. In the next year's field test, this contractor dropped half of the anchor questions, as previously noted, and also changed the 0 on its  $\theta$ -scale to its traditional value of the average  $\theta$ -value of a test's questions. This change lowered  $\theta$ -values of all future test questions by .46. In a major mistake,  $\beta_{\text{perf}}$  was not also lowered by .46, from .7 to .28. This error had the effect of raising the passing raw score by about 7 points. Because of the changes in the set of anchor questions and other (unknown) modifications by the new contractor, the passing raw score increased by only 4 points from June 2001 to June 2002. However, the 'high passing' raw score jumped by 9 points from 63 to 72. The Math A test was still being phased in at this time, and the technicalities of setting the passing scores were too abstruse for anyone to question these jumps, except perhaps the one overworked psychometrician. (The error was only uncovered recently by this author.)

*7.6.B. Proper calculation of the passing raw score.* There are two components to this problem. The first involves the error introduced by dropping the scale parameter  $\alpha$  in the item response curves. Recall that the performance standard  $\beta_{\text{perf}}$  equals the ability  $\beta$ -value  $x$  such that  $p_{\theta_{\text{book}},1}(x) = 2/3$ . The Math A Panel was unable to get any data to estimate what the true scale parameter should have been for the bookmark question's response curve. Unfortunately, changes in the choice of scale parameter can have major consequences. For example, if  $\alpha$  should have been 2 instead of 1 (meaning that the chances of success increased twice as fast with increasing student ability), then  $\beta_{\text{perf}}$  would drop from .7 to .35 and this in turn would cause the passing score on a typical Math A test to drop by 5 points.

This problem in the determination of  $\beta_{\text{perf}}$  has been corrected in the new Math A bookmark which is set at the difficulty of a problem that would have to be correctly answered 50% of the time instead of 2/3rds of the time. This means that  $\beta_{\text{perf}} = \theta_{\text{book}}$ , and so the scale parameter has

no role in setting  $\beta_{\text{perf}}$ . On the other hand, some psychometricians argue that a 50% success rate on the bookmark question is too low, because it is too close to the 25% success rate obtained by guessing.

The second problem concerns the failure of the equating procedure to lower the passing raw score as the tests got harder. From 1998 to 2002, students seemed to be getting stronger in Math A skills. The average p-score of multiple choice problems increased from .52 on the 1998 field test to .59 on the 2002 field test, although the multiple choice questions were getting slightly harder. However, as discussed in subsection 7.4 above, the tests were being constructed with a constant average p-score of about .47. Since the students were getting stronger, keeping the average p-score constant meant that *on average* harder questions had to be used.

There was no change in students' performance on anchor questions as they were performing better on Math A types of questions. The equating method did not raise the average  $\theta$ -value of the test questions, as it should have. Instead, the psychometric calculations took the improved student performance on the multiple choice questions to mean that those questions had gotten easier.

Here is one reasonable way to estimate how much this equating flaw affected the passing raw score. The .07 increase from June 1999 to June 2003 in the average p-score of the multiple choice questions translates into a decrease of about .4 in  $\theta$ -values, since the slope of the regression line of  $\theta$ -values versus p-scores was -.19. This in turn translates in to a drop in the passing raw score of about 6 points.

*7.6.C. The evolving bi-modal distribution of  $\theta$ -values.* Over time the 3-point and 4-point free response questions got considerably harder, while the multiple choice questions stayed about the same. However, to keep the overall average  $\theta$ -value of test questions constant, the average  $\theta$ -value of multiple choice questions decreased from .3 to .9 from the June 1999 test to the June 2003 test, and the average  $\theta$ -value of the 3- and 4-point free response questions (which carry greater weight) increased correspondingly from +.5 to +1.1. The breakdown of  $\theta$ -values by the four sections of the test is as follows.

Test	<u>Mult. Choice <math>\theta</math></u>	<u>2-pt Free Resp <math>\theta</math></u>	<u>3-pt Free Resp. <math>\theta</math></u>	<u>4-pt Free Resp. <math>\theta</math></u>
June 1999	-.26	.14	.14	.76
June 2003	-.88	-.30	.74	1.43

(The original June 1999  $\theta$ -values have been decreased by .46 in this table to compensate for the contractor's change of the  $\theta$  scale in 2000.) Note that the 2-point and 3-point free response questions had the same average  $\theta$ -value in the June 1999 test, while there was a gap of 1 unit between those  $\theta$ -values in the June 2003 test. Figure 3 shows the plot of  $\theta$ -values versus p-scores for the June 2003 test. Compare this plot with the June 1999 plot in Figure 2 (recall that the June 1999 plot is shifted to the right by .46 units, because its average  $\theta$ -value is .46 instead of 0). In the June 2003 plot, there are only two questions with  $\theta$ -values between -.5 and +.5, while there are 12 questions with  $\theta$ -values in the corresponding unit interval centered around  $\beta = .46$  in the June 1999 plot. Common sense tells one that a bi-modal distribution like this is inherently undesirable in any test.

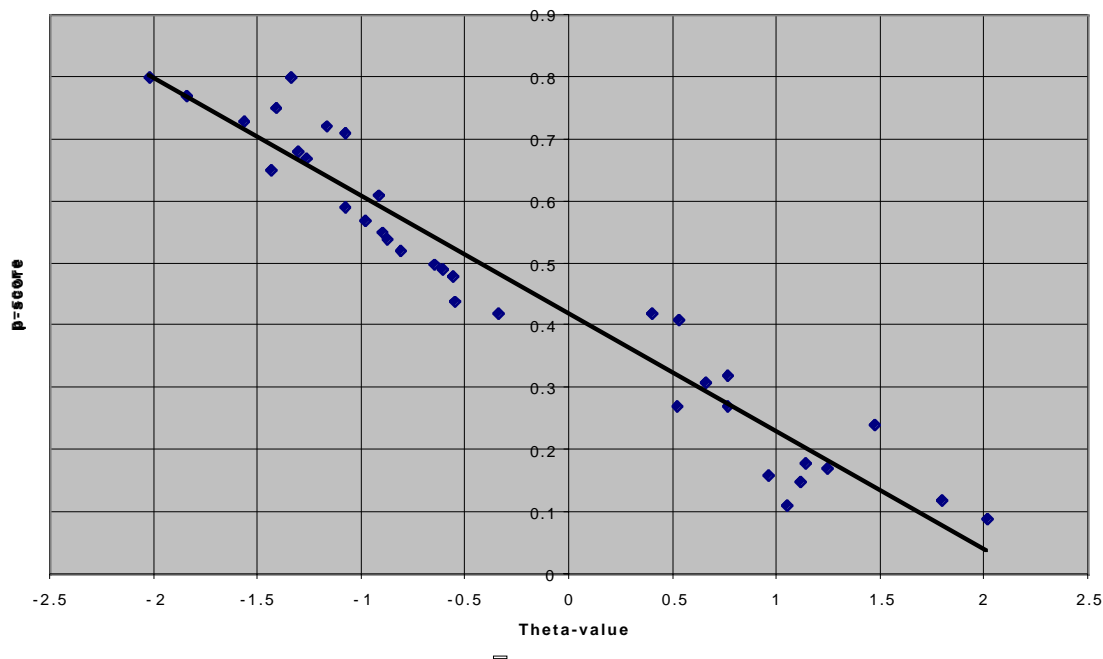
Recall that the performance standard was based on a bookmark question with difficulty  $\theta = 0$ . *It is nonsensical to estimate a student's probability of success of a bookmark question by asking questions that are much easier or much harder.*

Also note the poorer fit of the regression line in Figure 3 as compared to Figure 2. This is evidence that student performances on the June 2003 test questions (in the 2002 field test) were not fitting the item response curves well.

The assumption of a common scale parameter for all questions may not be serious if the  $\theta$ -values of questions were nicely distributed around the bookmark difficulty, as Figure 2 shows was occurring in the first Math A test. However, with the evolving bimodal distribution in Figure 3--  $\theta$ -values of multiple choice questions clustering near -1 and the  $\theta$ -values of free response questions clustering near 1--the scale parameter becomes very important in determining the expected score of the student of ability  $\beta_{\text{perf}}$ .

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  $\theta = 0$ , because there are no questions with those  $\theta$ -values. When a new bookmark was set in December 2003, using a

Figure 3: June 2003 test-- Theta-value vs p-score based on 2002 field test



sample of recent test questions, the bookmark question chosen had a  $\theta$ -value that was .25 less than the next harder question. If the next harder question were chosen for the bookmark, the  $\theta$ -value increase of .25 would translate into an increase of 3 or 4 points in the passing score on future Math A tests, which in return translates into a like decrease of 8-10% in the percentage of students who will pass the test.

### 8. Explaining the Decline in Student Performance from June 2002 to June 2003.

The preceding sections have dissected the evolving nature, intended and unintended, of the content standards and performance standards for the Math A test. Now we return to the event that triggered the investigation of the Math A test, namely, the high failure rate on the June 2003 Math A test. While reasons for changes in the difficulty of test questions and increasingly flawed calculation of the raw passing score have been described, they do not account the

dramatic jump in the failure rate from June 2002 to June 2003. The new psychometric vendor's erroneous shift in  $\beta$ -values was not relevant because it occurred earlier and thus affected June 2002 and June 2003 tests equally.

Some of the high failure rate was surely due to the presence of weak junior and seniors who had put off taking the Math A test after failing the easier Math I test in their freshman or sophomore year. On the other hand, there was a 12 point drop in the average raw score of 9<sup>th</sup> graders taking the Math A test, from 63 to 51, according to a large SED sample of student scores. Since Math A was a three-semester course normally taken in 9<sup>th</sup> grade and the first half of 10<sup>th</sup> grade, the 9<sup>th</sup> graders taking this test were generally among the stronger students (and had not put off taking it). Their dramatic drop in performance from June 2002 to June 2003 seemed to indicate that the test was really much harder. Only 60% of the 9<sup>th</sup> graders in the sample passed the June 2003 Math A test.

**8.1 Quantifiable Aspects.** The field test data explained some more of the difference. The expected raw score on the June 2003 test was 35 in the 2002 field test, about five points lower than the expected raw score on the June 2002 test on the 2001 field test. This drop appears to have been caused by an increase in the relative difficulty of the 3-point and 4-point free response questions. The average  $\theta$ -value of 3- and 4-point free response questions in June 2003 was .42 higher than in June 2002. The average  $\theta$ -value of the 2-point free response questions stayed about the same, and the average multiple choice  $\theta$ -value dropped from -.57 in June 2002 to -.89 in June 2003. However, the change in difficulty of free response questions was actually greater. The multiple choice problems on the June 2003 and June 2002 tests had almost the same average p-score (percentage of students getting a problem correct) and so their  $\theta$ -values should have stayed the about same from June 2002 to June 2003. The drop in the average multiple choice  $\theta$ -value by .32 was due solely to the requirement that the average of all the  $\theta$ -values on a test had to stay at about 0, as the free response questions got harder. If the anchor questions had picked up the improving student performance, then the  $\theta$ -values of multiple choice questions would have stayed the same (or increased slightly) and the average  $\theta$ -value of the 3-point and 4-point free response questions would have increased by about .70— a substantial shift that resulted in the observed decrease of 5 points in the average overall raw score from June 2002 to June 2002.

**8.2. Psychological Aspects.** The harder 3- and 4-point questions account for almost half of the drop on the 9<sup>th</sup> graders average raw score. The rest of the shift was judged by the Math A Panel to be due to the psychological impact of the harder 3-point and 4-point questions. In the field test, the full test was broken into three subtests, each of which typically included only 4 of the hard 3-and 4-point questions and so their impact was not as greatly felt. Teachers on the Math A Panel reported that many of their students said that when they came to the 3-point questions and found the first few very hard, they gave up. Here are the first two 3-point questions from the June 2003 Math A test.

#26. Seth has one less than twice the number of compact discs (CDs) that Jason has. Raoul has 53 more Cs than Jason has. If Seth gives Jason 25 CDs, Seth and Jason will have the same number of CDs. How many CDs did each of the three boys have to begin with?

#27 Tina's preschool has a set of cardboard building blocks, each of which measures 9 inches by 9 inches by 4 inches. How many of these blocks will Tina need to build a wall 4 inches thick, 3 feet high, and 12 feet long?

Neither of these types of problems had appeared on previous tests. Three-point question #29 was also quite complex in its lengthy (92-word) presentation.

#29. A certain state is considering changing the arrangement of letters and numbers on its license plates. The two options the state is considering are:

Option 1: three letters followed by a four-digit number with repetition of both letters and digits allowed.

Option 2: four letters followed by a three-digit number without repetition of either letters or digits.

[Zero may be chosen as the first digit of the number in either option.]

Which option will enable the state to issue more license plates? How many more different license plates will that option yield?

The psychological impact of such challenging questions is the most likely reason why the field test data, obtained from the three subtests, underestimated the difficulty of the actual June 2003 test. These hard questions also probably made students more prone to misreading easier questions that were phrased differently. For example, students had seen many questions which asked them to plot on graph paper a linear revenue function and a linear cost function and then asked for the point where the revenue equaled the cost. Question #35 asked for students to plot the *net profit* from a dance as a function of attendees where there was a fixed cost of \$40 for the disc jockey and revenue from tickets sold for \$2 a person. Then the question asked, “How many tickets must be sold to break even?” Students were confused by this change of language—and many did not know how to interpret ‘break even’ in mathematical terms. Instead, many students plotted  $y = 40$  and  $y = 2x$ , and got the right answer by determining where these two lines intersected. However, the scoring rubric gave no credit for this approach.

How did so many harder and differently worded questions get placed on the June 2003 test. At a human level, the answer is probably that the question writers and test construction committees were ramping up the difficulty of questions and trying not to repeat previous types of questions. At the time the questions on the June 2003 test were being written in the fall of 2000 and assembled into tests in fall 2001, high scores on recent Math A tests encouraged this strategy. Furthermore, as students performed better, harder free response questions were needed in the construction of tests whose average p-score of questions was held constant over time.

## 9. Other Issues

There were a number of issues indirectly related to the Math A test that the Math A Panel also addressed. Along with higher standards for students, the Panel recommended higher standards for teachers in both their pre-service preparation and in-service professional development activities. In particular, it recommended that all future elementary school teachers be required to take 9 credit hours of mathematics specifically related to the foundations of elementary mathematics. Currently, professional development in New York is vaguely defined and, especially at the elementary school level, rarely content-specific. Many school superintendents prefer to have single district-wide professional events, e.g., workshops on disruptive students. It is much more complicated and expensive to organize a number of subject-matter events which would require different sessions for different sets of grades. This will be a hurdle to in-school professional development plans in any state. Happily, the Board of Regents is very enthusiastic

about strengthening teacher education in mathematics and has stated its intention to issue regulations that strengthen the pre-service and in-service mathematical education of teachers in New York.

Passing the Math A test is very highly correlated with passing the New York 8<sup>th</sup> grade mathematics test, whose difficulty is close to the that of the high school mathematics graduation test in other states. This is good news, since it indicates that the Math A test was measuring the cumulative mathematical learning in all grades up to 10<sup>th</sup> grade, and that students could not pass the Math A test by cramming for this test for a few semesters. The bad news is that half of New York students have been failing the 8<sup>th</sup> grade mathematics test. As a consequence, the Math A Panel made several recommendations about K-8 mathematics instruction. In addition to strengthening the pre-service and in-service mathematical education of elementary teachers, the Panel recommended that a carefully structured curriculum be designed for grades K-8 to replace the current content and performance standards (similar to Math A), which are now laid out by pairs of grades (1-2, 3-4, 5-6, 7-8). This recommendation was made in part to lay the proper foundation for the future mathematics tests coming in grades 3, 5, 6, and 7, in addition to current tests in grades 4 and 8, all mandated by the No Child Left Behind Act.

## **10. Implications for No Child Left Behind Act**

The No Child Left Behind Act (NCLB) requires that each state establish performance standards in mathematics and English, assessed by a test each year, for students in grades 3 through 8 plus a high school graduation standard. Each year, a school is required to have 10% more of its students meet each performance standard than did the year before. Improvement is required not only overall but also in specific cohorts, such as Afro-American and Latino-American students. A school that fails two years in a row to meet the improvement targets for any test and any cohort can be labeled ‘failing’ with a number of associated penalties, including allowing parents to transfer their students to another school. These tests are meant to measure schools, not students. However, in New York, Math A is a graduation requirement for students as well as a component of the NCLB assessment of schools.

While the overall goals of NCLB are hard to fault, the requirement of continual 10% improvement on standards-based tests raises serious psychometric difficulties, given the above problems in standards-based tests in mathematics. When a school is required to improve from, say, 40% of students meeting the standard in year N to 44% in year N+1, the determination of the passing cut-score will play at least as great a role in success as the actual improvement in student learning. Every one point change in the pass cut-score typically translates into a 2 or 3% change in the percentage of students passing. Given the difficulties discussed above in maintaining a consistent performance standard from year to year, the margin of error in setting the passing cut-score will often introduce a level of psychometric ‘noise’ whose impact on the year-to-year percentage change in the passing rate exceeds the NCLB-mandated 10% improvement.

In short, the annual improvements mandated by NCLB are scientifically meaningless over short time spans. Even over the intermediate term, they are suspect given that the performance standard of the New York Math A test became about 20 points too high over a 4 year period.

## 11. Concluding Remarks

When asked to be a member of the Math A Panel, this reviewer assumed that the task would consist of making subjective judgements about the difficulty of problems on the June 2003 Math A test in comparison to earlier Math A tests. He also hoped 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. Doing the ‘data mining’ to identify useful information and figuring out what else to ask for was a time-consuming and daunting—but essential—part of the work. Some key information was missing or was withheld because it might appear on future tests, and had to be inferred indirectly. In the end, the only useable information was the p-scores and  $\theta$ -values of test questions and anchor questions from the field tests, along with the SED sample of the passing rates on the June 2002 and June 2003 tests and the value of  $\beta_{\text{perf}}$ . 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 many months.

At first it seemed as if only a little knowledge would be needed of the psychometrics underlying standards-based tests, but incrementally it became necessary to understand this subject in depth. The entire process was much more intellectually challenging than seemed possible. Some of the explanations presented here only became clear to this writer months after the Math A Panel presented its report. It is important to note that the explorations by this mathematician were welcomed and strongly supported by New York State Education Department officials.

The ideal of establishing and maintaining a consistent performance standard in mathematics for high school graduation—or for mathematical proficiency at earlier grades-- may not be attainable, given the questionable assumptions of psychometric equating methodology and the difficulty of developing valid performance bookmarks and anchor questions. The psychometric problems that arose in New York’s Math A test, as laid out in Section 7, are daunting.

Nonetheless, state-wide tests of some form are probably needed to pressure our K-12 educational systems— school administrators, teachers, parents, and students-- to raise the mathematical skills of the next generation of our workforce. If one looks at ‘standards-based’ graduation mathematics tests in other states, one tends to find a lot of standardized types of questions (that appear on the test every year) on which students can be drilled. This writer feels that mathematics tests of this form may well do more harm than good. It is imperative that mathematics high school graduation tests require mathematical reasoning to solve new problems along with mastery of the solutions of standardized problems. The challenge is to develop fair tests with this mixture.

What is clear is that Ph.D. mathematicians need to be integral participants in efforts to develop and monitor these mathematics tests. Their disciplined reasoning and creative minds will be fully challenged. This article was written to encourage other mathematicians to get involved, while alerting them to the complex issues they will face.

\*\*\*\*\*

ACKNOWLEDGEMENT: The author would like to express his appreciation to his 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. We struggled together. Our complementary insights made the Panel's investigations rewarding and highly productive. 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 Frank Baker, *The Basics of Item Response Theory*, ERIC Clearinghouse for Assessment and Evaluation, 2001. Available on the web at <http://ericae.net/irt/baker>.
2. Gregory Cizek, editor, *Setting Performance Standards*, Lawrence Erlbaum, Mahwah, NJ, 2001.
3. Paul Holland, On the Sampling Theory Foundations of Item Response Theory Models, *Psychometrika*, **55** (1990), p. 577-601.
4. Math A Panel Report to the New York Board of Regents, October 2003, [www.regents.nysed.gov/2003Meetings/October2003/1003brd3.htm](http://www.regents.nysed.gov/2003Meetings/October2003/1003brd3.htm).
5. New York Times, "How to Measure Student Proficiency", Dec. 18, 2003, page B8.