**Problems with Standards-Based Mathematics Tests**

I want to alert mathematicians to some troubling problems that can affect the standards-based mathematics tests mandated by recent state and federal legislation. These problems became apparent during my work on a New York Board of Regents' special panel investigating the high failure rate on the June 2003 New York mathematics graduation test. Major deficiencies were found both in setting performance standards and in maintaining the performance standards. Many of these problems involve very technical aspects of the psychometric methodology, based on Item Response Theory, for maintaining a constant performance standard over time. The New York State Education Department has only one psychometrician who oversees dozens of tests. Outside vendors do the actual psychometric work. My analysis indicated that instead of requiring a score 51 out of 85 to pass the June 2003 Math A test, the true cutoff should have been around 30. The report of the special Regents panel [2] led to a lowering of the passing score and a total reworking of future math graduation tests. While the panel did not have the time to fully analyze the psychometric reasons for this flawed passing score, subsequent study by this writer was able to supply many of these details (see [4]). The findings of this study are summarized here.

Any test whose problems are not totally predictable is likely to be affected by many of these psychometric problems, especially if the test is aiming to raise the performance of students over time, as mandated by the No Child Left Behind Act.

Here is what Item Response Theory claims to do in a nutshell: it can calculate a consistent passing score on future tests based on the projected performance of a hypothetical student who can solve a certain problem correctly with probability 2/3rds. Clearly, this theory involves a lot of assumptions.

**A. Setting the Performance Standard.** This process is less technical, but still subject to troubling problems. One gives a collection of questions to a sample of students. The questions are ranked by the percentage of students who get them right. An expert committee of mathematics teachers and professors go down the ranking, from easiest to hardest, looking to set a 'bookmark' at an item judged to be of a difficulty that someone meeting the desired performance standard would get the item right, say, 2/3rds of the time. One criterion for setting this bookmark is if the experts think that the next three (harder) items on the list are unlikely to be correctly solved 2/3rds of the time by a student at the borderline for passing.

Most mathematicians are distressed by the premise that a single 'bookmark' question can serve as the foundation for assessing a range of mathematical knowledge and reasoning.

The obvious operational problem with this approach is that people will honestly differ about this criterion. The choice of experts, of problems and of students in the standards setting process all make a difference. Mathematics performance standards vary greatly from state to state [3].

A more subtle but equally serious problem is that students' performances are dependent on which questions the students are familiar with (i.e., were drilled on). Mathematics teachers will frequently rank the difficulty of questions in a very different order than that given by student performance on field tests; this was the case with the New York Math A test. When students and teachers have different standards of difficulty, the whole bookmark process breaks down.

In New York as in other states, the new Math A course involved a more challenging, problem-solving curriculum. The previous curriculum was more procedural. The questions used to set the Math A performance standard were based on the old curriculum and the students who were solved these problems were trained in that old curriculum.

There is also 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 reasonably 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 gradation performance standard needs to move up with these rising expectations, but 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—it was put into effect in 2003-- that was above the ability of the average 1999 student.

**B. Maintaining a Consistent Performance Standard.** According to Item Response Theory (IRT) [1], the mathematical ability of students and the difficulty of test items can be placed on a common β-value scale, frequently chosen so that the average β-value of items is 0 and the β-value units are measured in standard deviations. IRT posits that each item has an item response curve for the probability of a correct answer of the form:

$$p_{\beta,\alpha}(x) = 1/\{1+e^{-\alpha(x-\beta)}\},$$

where $x$ is the ability of a student, β is the difficulty level of the item and α is a scale parameter that implicitly controls the slope of the response curve. Figure 1 shows a response curve with β = 0, α = 1. There are other versions of the item response curves; using one or three parameters.

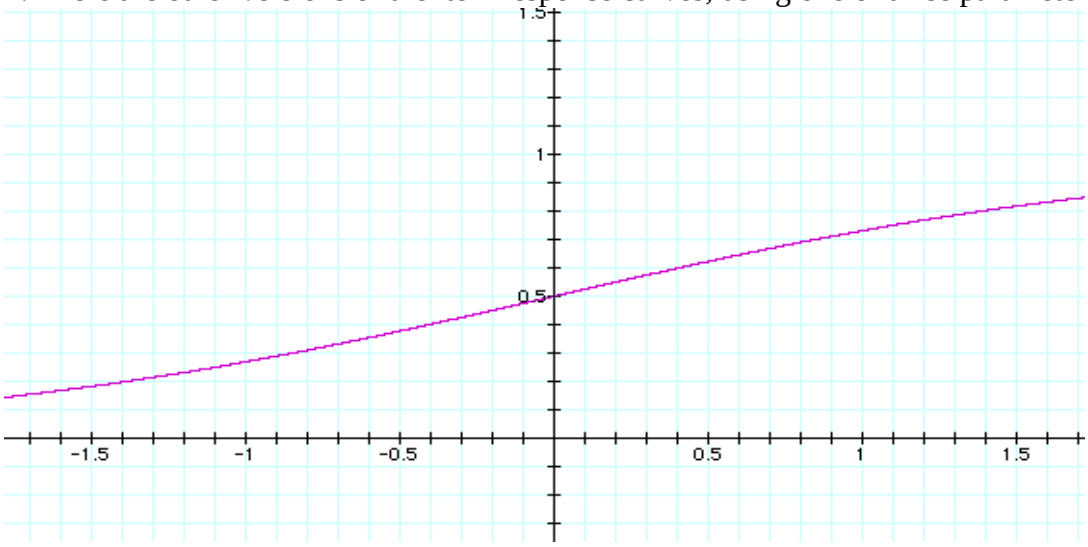


Figure 1: $p_{0,1}(x) = 1/(1+e^{-1(x-0)})$,

A question is assigned a difficulty rating of β if a student of ability β has a .50 probability of correctly answering it. The β- and α-values for items are determined from field test data by maximum likelihood estimation (this simultaneously involves calculating the ability levels of students in the field tests).

Suppose the β-value of the bookmark item in the performance standard (see A. above) is β=0, as was the bookmark on the original Math A test. If the performance standard is to get such a bookmarked item correct with probability 2/3, then the performance standard will be met by a student of ability β = .7, according to the Figure 1; i.e., $p_{0,1}(.7) = 2/3$ (assuming α=1). The passing score for a test is the expected score of a .7-ability student on the test. This expected score is the

sum of the expected number of points earned on each test item, $\Sigma\, p_{\beta_i,\alpha_i}(.7)\times n_i$, where item *i* has $n_i$ points and item response curve parameters $\beta_i$, $\alpha_i$.

For the performance standard-- the expected score by a .7-ability student on a test—to stay the same over time, the β-value scale must be invariant over time. To equate the β-value scale of test items from year to year for changing student ability and item difficulty, a set of anchor items is used in every field test. The β-values of anchor items are set in the first field test at the same time that the bookmark is set and are used to fix the β-value scale on future tests.

This is how Item Response Theory proposes to assess a prescribed performance standard in a consistent manner over time. Now we examine the possible problems with this theory, particularly those that arose with the Math A test.

*1. Model Assumptions:* The assumption of a one-dimensional scale on which to place students and questions is a risky simplification given the multiple domains of mathematical knowledge and the distortions in student performance caused by drilling for tests. Item response curves are reasonable approximations to student performance on problems but they fit field test data too imperfectly to be able to determine with a high degree of accuracy (i) β-values of test items; (ii) the ability level of the performance standard; and (iii) the passing scores on future tests. Note that item response curves are inherently unsuited to multiple choice questions where random guessing yields a probability of success of .25.

*2. Slope Parameter:* The sample size in field tests was too small for determining both α and β. The Rasch IRT model, used in New York, drops the scale parameter $\alpha$. Instead, the units on the β-scale were adjusted, which had the effect of giving a common scale parameter to all questions. This scale was fixed on the initial field test. Over time as the Math A tests got harder, the scale should have been changing. Further, because of problems with the field test design (see 4. below), psychometric calculations involving the scale parameter became distorted. While difficult to quantify, the scale problems introduced a substantial uncertainty in the calculation of β-values of test questions and of the passing scores.

*3. Misperforming Anchor Items.* The New York Math A test involved more problem-solving than in the previous curriculum. The items in the anchor set were drawn mostly from problems based on the previous curriculum. Students' performances on anchor items (and test items) is a function not just of their intrinsic ability but also of how similar those items are to problems on which students have been drilled. The increasingly out of-date Math A anchor items did not pick up the improving performance of students on Math A-type questions. Without the needed adjustment of β-values, keeping the average difficulty of a test's items constant over time required putting harder items on tests. The misperforming anchor items are estimated to have raised the passing score over time by at least 6 points (see [4] for details).

In 2000, most of the written-response anchor questions were dropped from the anchor set because they were generating poor psychometric statistics. The subjective nature of grading on written-response questions makes them less reliable from a psychometric viewpoint. On the other hand, challenging multi-step problem solving is the focus of Math A, and so written-response anchor questions were essential for measuring the improving problem-solving skills of students.

*4. Flawed Test Fields.* The psychometric calculations are done with field test data. Participation in field tests was voluntary, and the NY State Education Department had no idea which students in a school were taking the test fields and how they were prepared for them. To fit in 45-minute class periods, a three-hour test was broken into three subtests (with a few common items). Students'

average scores were about 15% worse on items on the field tests than on the operational tests. Further, there evolved strange, unexplained variations in psychometric data for different subtests and for the anchor test. IRT calculations based on data from such imperfect field test data was produced inexact β-values of test times and hence inexact passing scores. Breaking the hard June 2003 test into three subtests in the 2002 field test hid the destructive psychological impact of the succession of challenging written-response items on that test.

Because of these flaws in the field test design and in anchor items (mentioned above), the Math A Panel recommended an adjustment in the scoring of the June 2003 test that lowered by passing score from 51 to 36 (out of 85). Following the Math A report, New York totally redesigned field tests to eliminate these flaws.

*5. Bi-modal distribution of items.* The multiple-choice Math A items were fairly easy, and so to keep the average difficulty of items constant as students' performance improved, much harder written-response items were used. The Math A tests evolved to a bi-modal distribution of items, with multiple choice items grouped around a β-value of –1 and written response items around a β-value of +1. There were almost no items around the original bookmark level of β=0. The nominal performance standard of getting a bookmark item correct with probability 2/3 was being determined by a student's performance on much harder and much easier items. The determination of the passing score now became highly dependent on the shape of the (imprecise) item response curves. In particular, the inaccuracy (discussed in 2. above) in the common scale parameter of these curves now had a significant impact on the calculation of the passing score. This impact was additional factor in the Math A Panel's recommended adjustment in the June 2003 passing score.

*6. Vendor Mistake.* In 2000, New York changed the vendor who did the psychometric calculations of β-values and the passing score. The new vendor used a slightly different version of IRT which shifted the β-value scale for test items. Accidentally, the vendor did not change the ability level (β = .7) of the performance standard, raising the passing score by 7 points.

**C. Conclusion** A major component of the No Child Left Behind Law is sanctions for schools at which one or more defined cohort of students does not improve suitably over time. The reality is that passing scores on standards-based tests over time are very unlikely to be comparable at the level of precision that justifies high-stakes consequences, unless the tests consist of highly predictable questions (for which students can be drilled). From my study of one such standards-based test in New York, reliability is in its early infancy for standards-based tests that can measure the sort of mathematical reasoning that a high-quality K-12 mathematical education should develop in future citizens. While many of the problems with the original Math A test are being fixed, some problems such as reliable anchor items seem very hard to resolve. An extensive collaboration between mathematicians and psychometricians is needed before such standards-based mathematics tests can become a valid basis for assessing schools and students over time in mathematics.

References:

1. Frank Baker, *The Basics of Item Response Theory*, ERIC Clearinghouse for Assessment and Evaluation, 2001. Available on the web at ericae.net/irt/baker.

2.Math A Panel Report to the New York Board of Regents, October 2003, www.regents.nysed.gov/2003Meetings/October2003/1003brd3.htm.

3. New York Times, "How to Measure Student Proficiency", Dec. 18, 2003, page B8.

4. Alan Tucker, The New York Regents Math Test Problems, preprint available at www.ams.sunysb.edu/~tucker/MathA.htm.

-Alan Tucker, SUNY-Stony Brook