An Evaluation of the 2008 Alabama Statewide Mathematics Competitions

By Jim Gleason

On March 29, 2008, 824 students gathered at testing centers throughout the state of Alabama to participate in the 27th annual statewide mathematics competition. The first round of competition involves the students taking a multiple-choice test in one of three areas, Geometry; Algebra II and Trig; or Comprehensive. In this article, we will evaluate each of the multiple-choice tests using techniques of item response theory.

Item response theory is an alternative to classical test theory that creates a model based upon test data that analyzes each item (question) independently of the other items. In the January 2008 Notices of the American Mathematical Society, a detailed overview of item response theory and its use for mathematics competition is given (Gleason, 2008). For a general overview of item response theory, the reader is referred to (Hambleton. Swami-nathan, & Rogers, 1991) or (van der Linden & Hambleton, 1997).

Methodology

In order to analyze the 2008 Alabama State Mathematics Competition multiple-choice tests, we first needed to choose the appropriate item response theory model. Each of the tests are assumed to be unidimensional, in that they are generally measuring simple mathematical ability, and so multidimensional models were decided to be necessary. A large majority of students are not influenced by the time limitation and so a more complicated model taking the time issue into account was not necessary. Students are discouraged from guessing due to a guessing penalty and so a 3-parameter model is not appropriate.

Since an incorrect response counted as 0 points (guessing penalty), a blank response as 1 point, and a correct response as 5 points, the tests fit within the partial credit models of item response theory. Due to the hypothesis that the discrimination parameters of the items likely differ, they were freed from constraint and so the Generalized Partial Credit Model was used (Muraki, 1997).

The Generalized Partial Credit Model is based upon the Nominal Categories Model (Bock, 1997). This model takes into account all possible point values for each item and gives each of these point values difficulty and discrimination parameters. The computer software MULTILOG (Thissen, 2003) estimates these parameters for each item of the test and the participants' ability levels using a maximum marginal likelihood method . The parameters are then used to generate a set of item characteristic curves for each of the items which estimate the likelihood of individuals of various ability levels to answer that item correctly, answer incorrectly, or leave the item blank. In addition to the item characteristic curve, the parameters are also used to generate a test information function and standard error curve that are used to describe the reliability of the entire test.

Results

Geometry

One can see from Figure 1 that the ability estimates of the Generalized Partial Credit Model fit very well with the participants' actual scores. This verifies that the model does indeed fit the actual data and is therefore appropriate to evaluate the test.

Out of the 50 items included in the geometry test, only two did not add significant information to the test. These two items were number 8 and number 33. Number 8 was one of the easier questions regarding the sum of the angles of nested triangles with 30% answering correctly. Number 33 is a question that asks for the number of paths between two points with five choices. The fact that the correct answer is "more than 13" might have made the correct answer to this problem easier to guess than the correct answer to other problems. With only two questions having little information, this test proves to be well written with a good problem selection.

The overall test proved to have a large amount of information and a low standard error as shown in Figure 2. In fact, the marginal reliability was an extremely high 0.95 which shows an overall effectiveness at measuring the participants' geometric ability. One particularly interesting point is that among the top participants, the test did an excellent job of discriminating between individuals with the standard error being less than 0.2 for individuals who performed better than 50 points, the score achieved by leaving all answers blank.

Therefore, the geometry test successfully accomplished the goal of distinguishing between the best geometry students in the state of Alabama with 50 strong questions. The only possible changes would involve shortening the test since the reliability and standard error are much better than necessary for such a competition.

Algebra II and Trigonometry

The ability estimates of the Generalized Partial Credit Model fit the participants' actual scores very well (See Figure 1). Therefore, the model is appropriate to evaluate the test.

Of the 50 items on the Algebra II and Trigonometry test, 6 items provided little to no information. For all six of the items, the difficulty level of the item was too high for the participant population. The items were numbers 8 (5% correct), 13 (5% correct), 14 (4% correct), 32 (11% correct), 33 (6% correct), and 44 (4% correct). None of these items provided more than 0.10 of information and therefore could be removed without changing the overall ability of the test to distinguish between participants.

The test as a whole provided adequate information (marginal reliability of 0.91) as shown in Figure 2. Within the range of participant scores, the test of less than 0.32 with the least amount of error at the top of the range as is desired with a mathematics contest.

Therefore, the Algebra II and Trigonometry test was also very successful with the only possible modifications being the removal of a few extremely difficult items.

Comprehensive Division I

As with the previous two tests, the Comprehensive Division 1 test actual data matched the model's estimated ability level (Figure 1).

Upon analysis of the initial run, it was discovered that problem 39 was incorrectly graded since high ability students appeared to do much poorer than low ability students. After changing the grading of the problem, the model was run again. Of the 50 problems on the test, only problem 11 was too easy to provide any information with 84% answering it correctly. Problems 36 and 49 did not provide much information because they were too difficult. However, problem 49 provided more information than problem 36 since many participants guessed incorrectly on problem 36 with very few guessing incorrectly on problem 49.

The overall test information (marginal reliability of 0.93) was excellent with the standard error staying below 0.26 for all participants (See Figure 2). Therefore, the Comprehensive Exam for Division 1 was an excellent exam with the minor problem of one item being graded incorrectly.

Comprehensive Division II

The same incorrect item from the Comprehensive Division I was also included in the Comprehensive Division II as item number 41. The grading for this item was changed and the computer model re-estimated.

The overall test information (marginal reliability of 0.88) was less than desired, but still adequate for this type of competition since the standard error was below 0.30 for those individuals at the top of the ability range.

There were several items from this test that contained little or no information. The first such item was number 2, which asked for the positive root of the equation

$$\left(\sqrt{200}+\sqrt{56}\right)x^2+10x-2\left(\sqrt{50}-\sqrt{14}\right)=0.$$

While this involved only a complicated application of the quadratic formula, only 24% of the participants answered the question with less than 9% answering correctly. On the other hand, for item number 4, which involved evaluating

$$\cos\left(67^\circ\right)\cos\left(22^\circ\right)+\cos\left(23^\circ\right)\cos\left(68^\circ\right)$$

using trigonometric identities, more answered correctly, but most of those who gave an answer (37%) chose the incorrect solution of 0 (18%). This in itself is interesting since all four values in the entry are positive and should therefore have generated a positive solution.

The next item that contained little information involved absolute and relative error in measurement. The reason for the lack of information is that the participant's response to the item had no relationship to their estimated ability.

Item number 33, which is a question involving probability along the lines of the classic “Let's Make a Deal” problem, caused those who had a high estimated ability to answer incorrectly. Therefore, this item seems to be measuring a distinct construct separate from the overall construct that the remainder of the test is measuring.

This could possibly lead to such a question being removed from the test.

Only 2 out of 104 participants answered item number 42 correctly. This showed that either the participants were unable to understand the question, or to solve a system of linear equations. Either way, this item proved to be too difficult for the tested population and could have been removed without changing the test information.

For item number 43, those who marked an answer appeared to do so randomly since all answers, other than "none of these", were answered with equal frequency. Therefore, this item led the majority of students to leave it blank and only measured if students followed the instructions and left the problem blank when they did not know the answer.

Since item number 48 involved a combination of knowledge about increasing functions and properties of the natural log function, it was too difficult for this population and also provided no information for the test.

Because of the number of items that produced little information, one item that appeared to measure a second construct, and a small sample size (104), the model did not match the actual data as well as the other tests. (See Figure 1)

Comprehensive Division III

The ability estimates of the Generalized Partial Credit Model fit the participants' actual scores very well (See Figure 1). Therefore, the model is appropriate to evaluate the test.

With 59% of the responses being blank, the Comprehensive test for Division III did not have nearly as much information as the instruments for Divisions I (39% blank) or II (53% blank). This decrease in information resulted in a marginal reliability of 0.865 with the standard error between 0.3 and 0.4 within the ability range of the participants. This leads to the conclusion that this test was too difficult for the participants to get reliable data about their ability levels.

Twelve problems on the Comprehensive Division III test provided no information, three of which were problems from the Comprehensive Division II test that provided no information. The nine new problems with no information had low response rates (more than 80% of the participants left the question blank), and those who did answer correctly were scattered across the ability range. These nine problems were numbers 4, 11, 19, 28, 30, 35, 36, 39, and 43. The topics for these problems included complex square roots, multiple combinatoric steps, properties of exponentials, and geometry.

Discussion

Overall, the multiple-choice portion of the 2008 Alabama Statewide Mathematics Contest was well written with appropriate reliability to distinguish between the participants' ability levels. This was particularly true for the Algebra/Trigonometry, Geometry, and Comprehensive (Division I) tests.

The major issues with the Comprehensive tests for Divisions II and III involved the questions being too difficult for the participants and causing a high occurrence of blank responses. One of the goals for the test writers should be to have the blank response rate below 40% since this produced marginal reliabilities above the 0.9 desired to distinguish between individuals. An additional benefit of reducing the blank response rate is that the participants will enjoy their experience more and not be as discouraged at the end of the test.

References

[1] Bock, R. D. (1997), The Nominal Categories Model, In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory,* (p. 33-50), New York: Springer-Verlag.

[2] Gleason, J. (2008), An Evaluation of Mathematics Competitions Using Item Response Theory, *Notices of the American Mathematical Society,* 55(1), 8-15.

[3] Hambleton, R., Swaminathan, H., & Rogers, H. (1991), *Fundamentals of Item Response Theory,* Newbury Park, CA: Sage.

[4] Muraki, E. (1997), A Generalized Partial Credit Model, In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory,* (p. 153-164), New York: Springer-Verlag.

[5] Thissen, D. (2003), *MULTILOG for Windows (version 7.0),* Mooresville, IN: Scientific Software International, Inc.

[6] van der Linden, W. J., & Hambleton, R. K., (1997), Item Response Theory: Brief History, Common Models, and Extensions, In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory,* (p. 1-28), New York: Springer-Verlag.

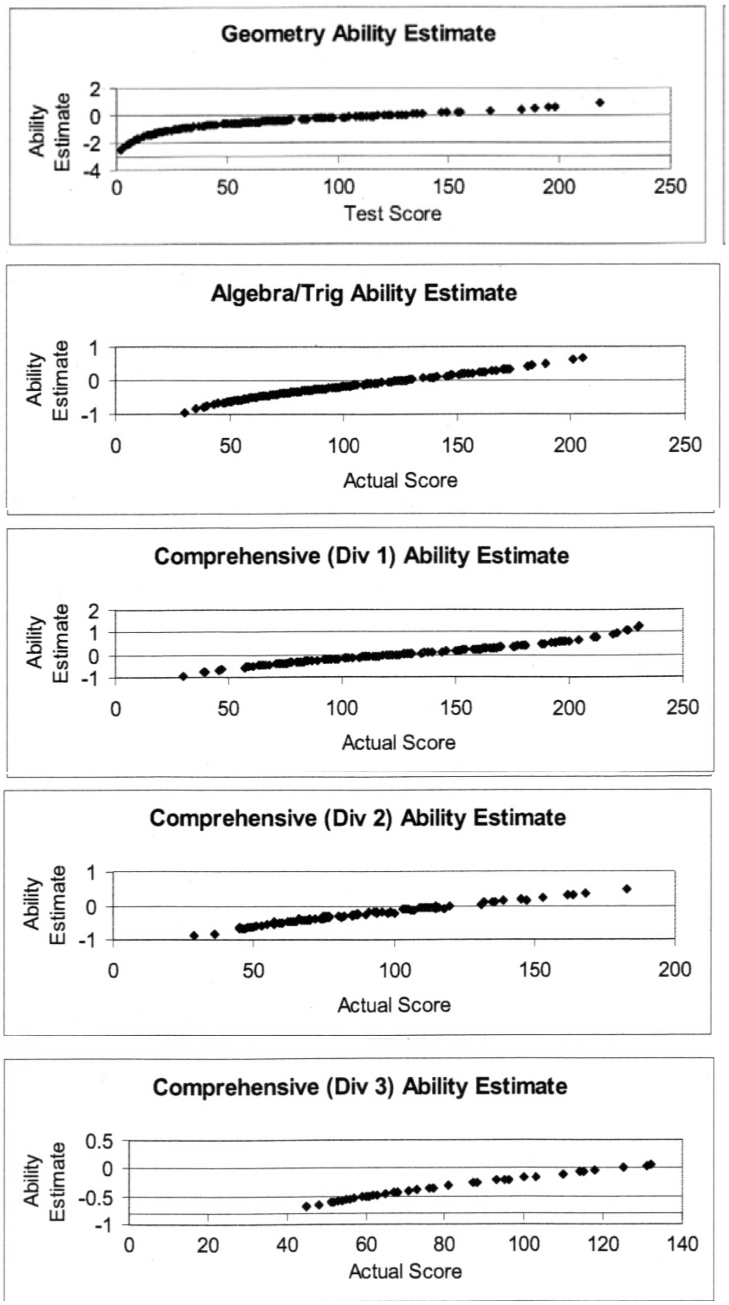


Figure 1. Comparisons of Estimated Ability Level and Actual Test Score

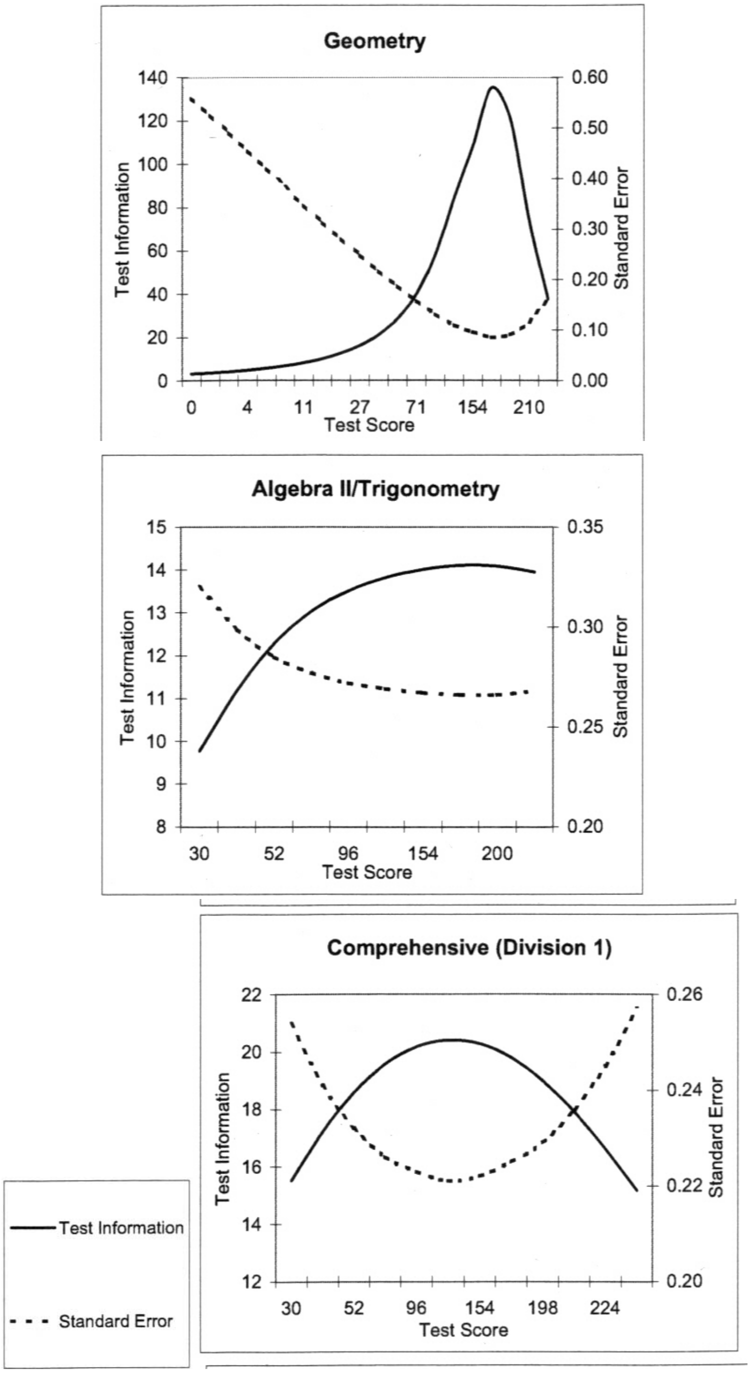
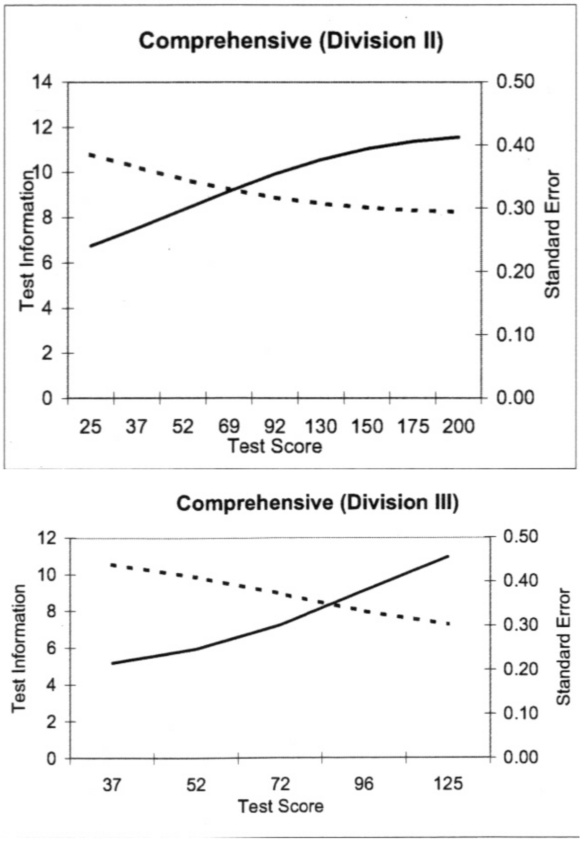


Figure 2a. Test Information and Standard Error Graphs



Comprehensive (Division II)

Comprehensive (Division III)

Figure 2b. Test Information and Standard Error Graphs

Department of Mathematics
The University of Alabama
Tuscaloosa, AL 35487-0350
(205)348-1975
jgleason@as.ua.edu
jgleason@bama.ua.edu

ACTM Fall Forum 2009

Exploring Math from Many Angles

October 15-16

Auburn University Montgomery

Thursday, October 15 - Starting at 1 pm (Registration Begins at 12)

For more information go to http://www.alabamamath.org

Speaker proposal forms are also available at the address above.