An Evaluation of Mathematics Competitions Using Item Response Theory

Jim Gleason

During the past century, mathematics competitions have become part of the landscape in mathematics education. The first major mathematics competition in the United States began with an idea of William Lowell Putnam in 1921 [6], and the mathematics competition named in his honor has become the premier mathematics competition in the United States for undergraduates, with 508 schools and 3,640 individuals participating in 2006 [3]. For high school students, the largest current competition is the American Mathematics Competitions, which are a part of the qualification process for the International Mathematical Olympiad [5].

The American Mathematics Competition began as the "New York High School Contest" in 1950 with over 6,000 students from 238 high schools and was sponsored by the Metropolitan New York Section of the Mathematical Association of America [16, 17]. Since then, the competition has expanded to a nationwide contest with over 413,000 students from over 5,100 schools participating in at least one part of these competitions [5]. In addition to the American Mathematics Competition, there are also hundreds of other local and regional competitions [10].

While these competitions have been in existence for over fifty years and involve over half a million students every year, very little has been done to evaluate these programs [10]. What little has been done has generally involved qualitative studies involving participant surveys [10, 8] or a long term study involving winners of these competitions [12, 13].

Jim Gleason is assistant professor of mathematics education at the University of Alabama. His email address is jgleason@as.ua.edu.

The primary goals of mathematics competitions are often things such as recruiting students, increasing excitement about mathematics, and providing schools with motivation to spend more time with some of their best students. However, within the context of a mathematics competition there is an underlying goal to distinguish the very best students in mathematics apart from the others. In this article, we will look at the effectiveness of two mathematics competitions at achieving the goal of measuring the mathematical ability of students.

Since one of the standard instruments used to make these distinctions among students is the multiple-choice test, we will investigate what is measured by these multiple-choice tests and the extent to which these tests achieve their goal of distinguishing among various participants. This analysis is based upon the psychometric field of *item response theory* combined with an evaluation of individual questions using a knowledge of the mathematics involved.

Another purpose of this article is to introduce mathematicians to item response theory and other tools available for the analysis of multiple-choice tests. The procedures used in this project can be modified to improve tests such as mathematics placement exams given to incoming freshmen or common final exams given to large groups of students.

Overview of Item Response Theory

Classical test theory has been used for years to determine reliability, difficulty, and other characteristics of measurement instruments. Unfortunately, there are some weaknesses of classical test theory. The most inconvenient weaknesses are that the

item indices, such as item difficulty and discrimination, obtained using classical test theory are dependent upon the group of examinees upon which they are tested, and the ability estimates for the participants depend upon the selection of items for the instrument [1]. For example, if a researcher desires to remove certain items from an instrument, classical test theory would dictate that the instrument would need to be completely reevaluated since the initial findings of the instrument and item difficulty are no longer valid.

In order to overcome these difficulties, psychometricians developed a new testing theory called item response theory (IRT). Begun in the 1950s and 1960s, IRT became a fully developed theory in the 1970s and 1980s with the evolution of computers able to handle the large numbers of computations involved in creating examinee ability estimates and item indices [19]. While the initial models dealt primarily with dichotomous data (question is right or wrong), new models have been developed to work with polytomous data including rating scales, multiple-choice models, sub-tests, and various other issues that arise in psychological and educational instrument development [19].

One of the key components of the various IRT models involves the *item characteristic curve* (ICC), which is a graphical display of the information for each item in an instrument generated by the IRT model. This ICC in the basic IRT models is a strictly increasing function that estimates the probability that someone with a certain ability would answer that item correctly. For instance in Figure 1, 10% of examinees whose ability is two standard deviations below the mean would answer this item correctly, while 80% of those examinees with an ability one and a half standard deviations above the mean would have a correct response.

In the Rasch model, also called the 1-parameter logistic model, the ICC is based upon each item having a difficulty parameter. In this model, the ICC for each item is given by

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

where $P_i(\theta)$ is the probability that someone with ability θ answers the i-th item correctly with b_i being the difficulty parameter of item i.

For the 2-parameter logistic model, as shown in Figure 1, each item has a discrimination parameter, a_i, in addition to the difficulty parameter, b_i. The ICC for this model is given by

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

where D is a constant used to make the logistic function similar to the normal ogive function.

In the more complex models, such as the multiple-choice model, the ICC for each item is made up of curves corresponding to each of the

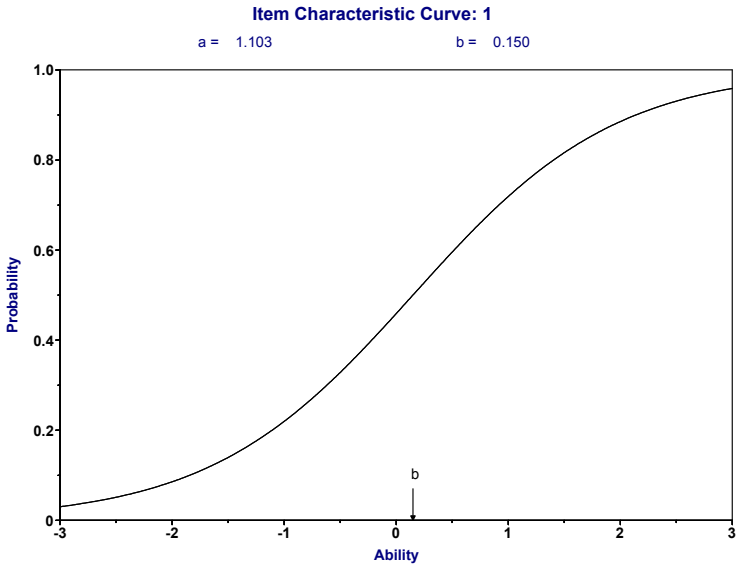


Figure 1. Example Item Characteristic Curve.

choices for the multiple-choice question, the correct solution, and the distractors [15]. This can be seen in Figure 2 where the correct solution, curve 2, starts with almost 30% of participants whose ability level is three standard deviations below the mean making this selection. It then dips down to around 15% of participants with ability around one standard deviation below the mean choosing the correct answer. Finally, over 90% of students whose ability level is at least one standard deviation above the mean make the correct choice. The other curves represent the other choices, or distractors, for the item.

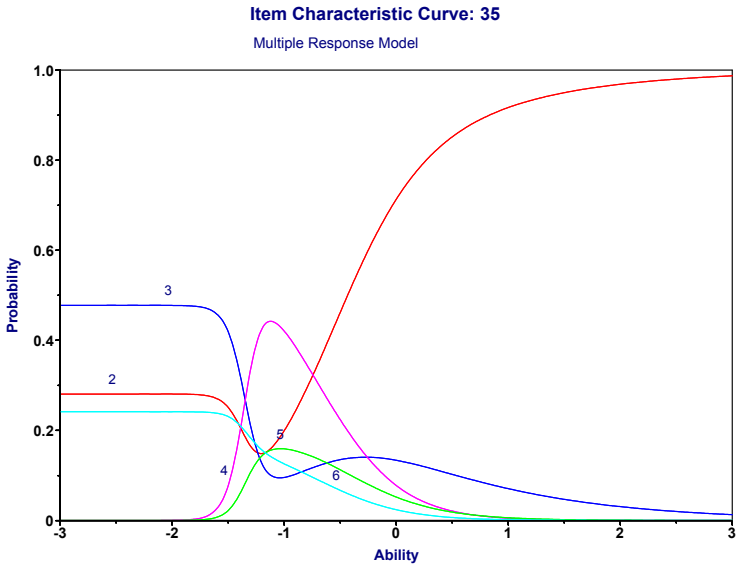


Figure 2. Multiple-choice Item Characteristic Curve.

The generation of the ICC for an item uses various item indices produced through a reiterative process that varies according to the software used. In this study, the software used a *marginal maximum likelihood* estimation procedure that estimated the ability levels of the examinees concurrently with the item parameters using the responses of the examinees to each item of the test.

In addition to the item characteristic curves, each item also has an *item information curve* (IIC), which describes the amount of information the item provides at various points along the ability spectrum. These IICs are then added together to create a *test information curve* that is then used to determine the level of error in differentiating between examinees at various ability levels. While the ideal test information curve would be a horizontal line, in practice, it usually has a peak somewhere around the mean and tapers off near the edges of the ability range. This corresponds to the test being fairly accurate at distinguishing ability levels near the mean, while not as accurate at separating the extremely high or low performing participants.

For more information about item response theory, two very useful books are by Hambleton, Swaminathan, and Rogers [2] and a collection edited by van der Linden and Hambleton [18].

Methodology

This paper is an evaluation of two high school mathematics competitions composed of multiple-choice test instruments. This evaluation focuses on three main questions that are answered for each of the two competitions.

What Does the Instrument Measure?

To answer this question, we must first look at the dimensionality of the instrument by comparing the magnitudes of the eigenvalues of the inter-item correlation matrix [7]. Ideally, each of the instruments are unidimensional, i.e., there is a single ability trait that the instrument is measuring. This dimensionality is demonstrated by the eigenvalues of the inter-item correlation matrix having a dominant eigenvalue.

Once the unidimensionality of the instrument is established, this leaves open the question of what ability trait the instrument measures. For both of the multiple-choice tests analyzed in this study, the questions come from various areas of mathematics such as number theory, algebra, trigonometry, geometry, and probability. Therefore, if the instruments are unidimensional, they are most likely measuring a general mathematical problem solving ability related to questions similar to those asked in this type of a setting. What that might generalize to is an important question to answer, but not one that can be answered in the scope of this study.

How Much Information Is Provided by the Instrument?

In order to determine the amount of information provided by each of the tests, one must first determine which model from item response theory best fits the situation. After testing for the unidimensionality of each of the tests, the same model was used for each of the tests since the testing situation for each was very similar.

Each of the tests was given during a timed situation, but one in which a majority of the participants completed the test during the time allotted. Therefore, models for time limit tests were considered, but rejected since these models were designed for tests in which all questions could be answered by all participants given enough time [20, 9].

Another model that was eliminated for an evaluation of each of the complete tests is the multiple-choice model of Thissen and Steinberg [15]. This decision was based on a desire to evaluate the test in its entirety instead of focusing on individual items. However, this model was used in the analysis of each test to determine which items provide the most information and how to improve these items.

One additional model that was considered but not used is the three-parameter logistic model of Birnbaum as described in the chapters contributed to the work of Lord and Novick [4, Chaps. 17-20]. The reason for not using this model was that guessing on each of the tests was discouraged by informing the participants that each correct answer counted as four points, each blank answer as one point, and each incorrect answer as zero points.

After an initial run using the data in the one-parameter and two-parameter logistic models, it was decided that the two-parameter model best fit the data due to large discrepancies in the discrimination parameter of the items generated by the one-parameter model. Thus the data from each of the tests was coded as dichotomous data with correct answers as 1, incorrect answers as 0, and no answer as 9, and the two-parameter logistic model in MULTILOG [14] was used to analyze the data.

What Types of Items Provide the Most Information?

In addition to an analysis of each test in its entirety, it is important to analyze each test item to determine which items generate the most information and how various items can be refined to either generate more information, or similar amounts of information, while taking less time for participants

to answer. This analysis was completed using three main tools.

First, by studying the eigenvector associated with the dominant eigenvalue of the inter-item correlation matrix it may be possible to determine which items contribute the most to the dominant trait that is being tested. Using a similar procedure, items were analyzed for their lack of contribution to the dominant trait.

Furthermore, by studying the item characteristic curves (ICC) and item information curves (IIC) generated for each of the items during the analysis of the tests using the two-parameter logistic model, each item was categorized as either contributing or not contributing based upon the amount of information generated by that model. An item was deemed as contributing if the item information curve had a maximum value above two units in the graphical interface of MULTILOG.

Finally, the multiple-choice model of Thissen and Steinberg [15] was used to further analyze the items by studying the ICCs of each of the distractors (incorrect choices) as well as the correct solutions. Through this analysis, certain items whose item characteristic curves for the correct response are not strictly increasing or whose ICCs of the distractors do not fit an appropriate form were further analyzed.

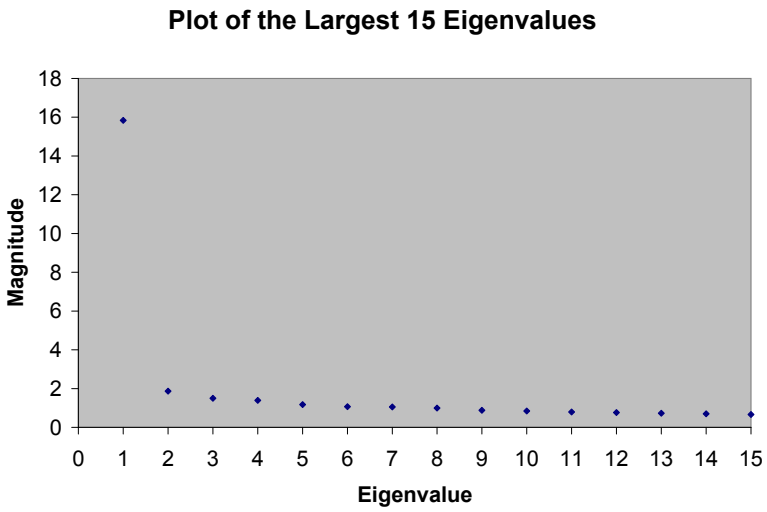
Results

University-Sponsored Comprehensive Competition

Conducting a factor analysis of the results of the university-sponsored comprehensive competition instrument generated seven eigenvalues of the inter-item correlation matrix greater than or equal to one. However, since the largest eigenvalue is more than eight times larger than the second largest, this instrument is unidimensional. (See Figure 3.)

Additional information gathered from the factor analysis is that only six of the thirty-nine items contributed less than 50% of their information to this primary component. Therefore, it would appear that the exam measures a general mathematical ability and not any particular topic within mathematics.

Following the factor analysis, the data from the university-sponsored mathematics contest was evaluated using a two-parameter logistic item response theory model using MULTILOG [14]. The appropriateness of this model was verified through a comparison of the model's ability estimates with the actual scores of the students with a correlation of 0.91. Since the two-parameter model shows that the instrument has a marginal reliability of 0.92 and a p value of less than 0.005, this instrument provides enough information to make claims about individuals' levels of knowledge. Furthermore, the



Extraction Method: Principal Component Analysis. [11]

Figure 3. University Comprehensive Test Eigenvalues.

IRT model showed that the instrument has a standard error of less than 0.5 for those who are between three standard deviations below the mean and almost three standard deviations above the mean as is shown in Figure 4. Therefore, for the majority of those participating in the contest, the instrument was capable of differentiating ability levels among individuals.

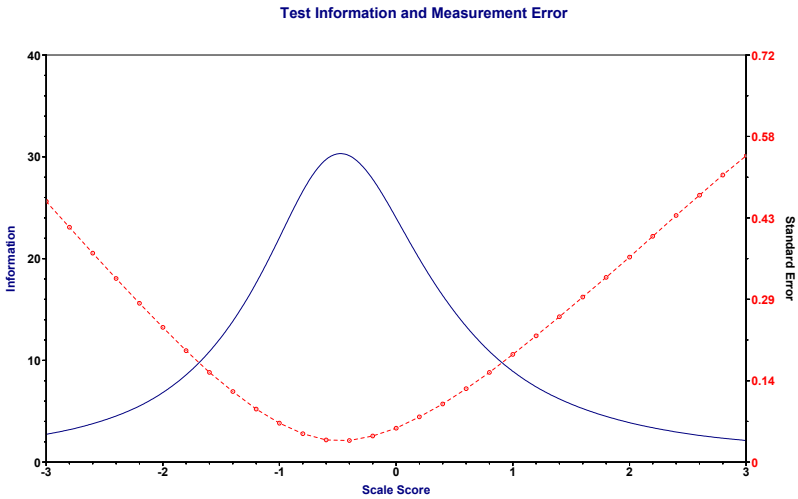


Figure 4. University Comprehensive Test Information.

After an evaluation of the entire instrument using methods from item response theory, we evaluated individual items using the the inter-item correlation matrix and the two-parameter logistic and multiple-choice models. Since the instrument

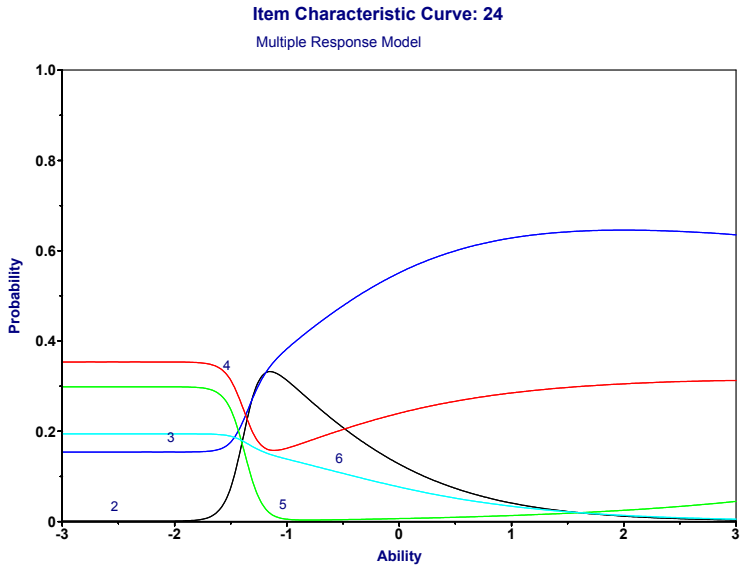


Figure 5. ICC for Item 24 of the University Comprehensive Test.

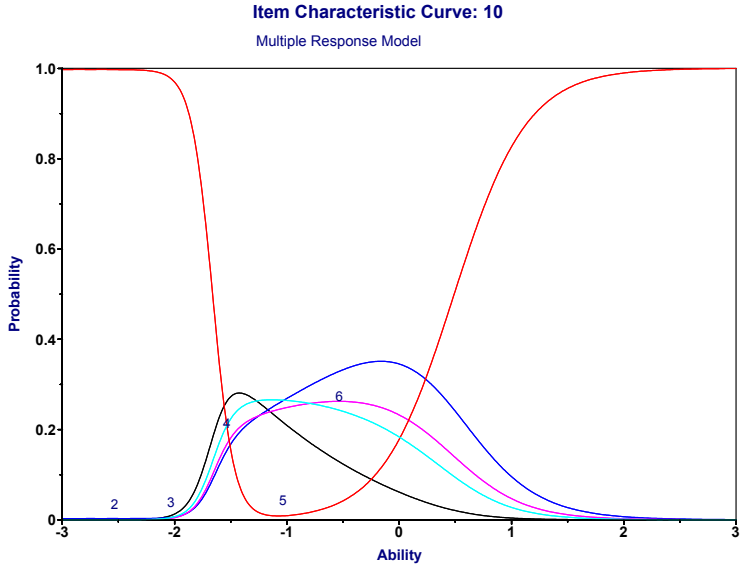


Figure 6. ICC for Item 10 of the University Comprehensive Test.

was very unidimensional, the evaluation of the eigenvectors of the inter-item correlation matrix provided no useful information. Also, the analysis of the instrument using the two-parameter logistic model showed that only item number 24 had an item information curve that did not rise above two units. Therefore, in order to provide a better understanding of the individual items, an analysis of the test was conducted using a multiple-choice model in MULTILOG that had a marginal reliability of 0.97 with $p < 0.005$. This model was further verified to be appropriate for this instrument since the correlation between the students' actual scores and the IRT model's ability estimates for the students was 0.94.

In order to fully analyze item 24, it is helpful to see the actual item in question in addition to its multiple-choice model ICC given in Figure 5.

24: How many distinct points of intersection are there for $r = 1 - 2\cos\theta$ and $r = 1$?
(A/2) 1 (B/3) 2 (C/4) 3 (D/5) 4
(E/6) none of these

Looking at the item characteristic curve (Figure 5) for this item which did not provide much information, one can see that a problem is that distractor (B/3) was chosen more frequently as the ability level increased. At the same time, the correct choice, (C/4), was chosen less frequently as the ability levels increased. Mathematically, the reason behind this difficulty is that if the participant ignores the possibilities of negative values for r, algebraic methods produce two distinct points.

Therefore, while this item does not provide much information for IRT models, it may be useful for measuring the ability of the top students to think in a non-traditional manner. Whether or not such an item should be included should then be left up to the individual contest coordinator to decide according to his or her desired outcomes. However, one adjustment to the item should be the removal of distractor (A/2) and possibly (E/6) since they provide no real information and take up valuable time for the assessment.

Another item of interest is the following:

10: A certain complex number, w, satisfies $w^2 = w - 1$. Determine w^{99}.
(A/2) i (B/3) $-i$ (C/4) 1 (D/5) -1
(E/6) none of these

whose ICC is given in Figure 6. This item, even though it contributes information according to the IRT models, should be removed from the instrument.

The reason for the removal is that a remarkably large number of contest participants whose ability levels were quite low chose the correct response, (D/5), while those of mid-level ability seemed to choose the other options at random. While the reasons behind these choices are worth pursuing, the inclusion of such a question on a mathematics contest is not appropriate.

Overall, this analysis shows that the multiple-choice test used for this mathematics competition is successful at discriminating among individuals, at least within two standard deviations of the mean. The analysis also shows that while two questions from the forty provided little or unreliable information, the majority of the items were highly appropriate for the situation.

Statewide Comprehensive Competition

By running an initial two-parameter logarithmic IRT analysis of the statewide comprehensive competition exam, six of the fifty items in the exam were determined to be questionable as to the information provided. After further review, it was found that these six items either had severe issues with their typesetting, or they had incorrect solutions. Therefore, these six items were removed and the analysis of the test was completed on the revised results.

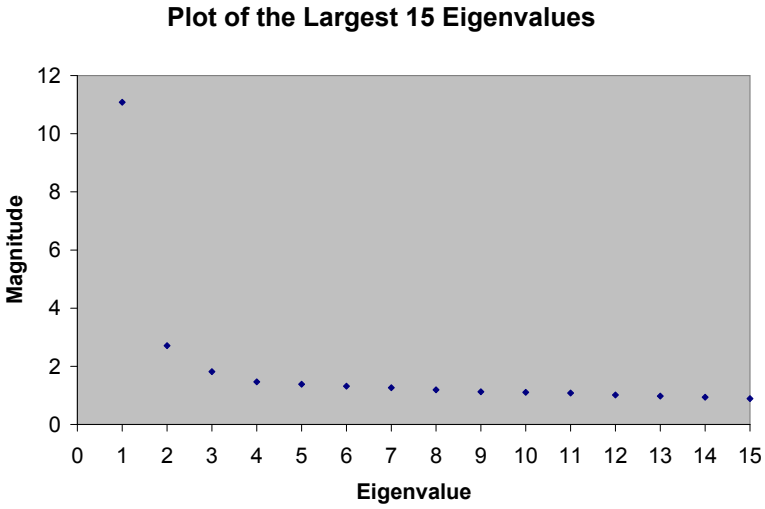
After running a factor analysis on the results of the revised exam, a dominant eigenvalue of the inter-item correlation matrix was evident as can be seen in Figure 7. It was also noted that all items contributed at least 27.3% of their information to this primary component. From this one can assume that the test instrument is unidimensional with all items contributing primarily to this latent ability trait.

Additionally, this new 44 item test was evaluated using a two-parameter logistic IRT model and was found to have marginal reliability of 0.93 with $p < 0.005$. Furthermore, the test had a standard error of less than 0.5 for those contest participants within two standard deviations of the mean as can be seen in Figure 8.

Through a study of the two-parameter logistic IRT model of the 44 item test, four items were determined to not contribute any significant amount of information to the test. Upon removal of these items, the new abbreviated statewide comprehensive test was processed using a two-parameter logistic IRT model and was determined to have a marginal reliability of 0.93 with $p < 0.005$. Therefore, very little information was lost through this abbreviation of the exam. To verify the appropriateness of this two-parameter logistic model, MULTILOG generated ability levels that corresponded with the participants' actual scores on the test with a correlation of 0.89.

Following the initial analysis of the test that resulted in a new abbreviated instrument of 40 items, the instrument was evaluated using the multiple-choice model to analyze these 40 items. The student ability estimates from this model corresponded to the actual scores with a correlation of 0.92, which verifies the reliability of the model.

Of the 40 items on the test, two items stood out for further evaluation. From the ICC for Item 18 (see Figure 9), one can see that this item provides a great deal of discrimination for those contest participants who are near half a standard deviation above the mean. The difficulty with this item includes the type of information this item provides for participants whose ability level is outside this small range. One possible way to make this item more effective is to make some adjustments to the distractors.



Extraction Method: Principal Component Analysis. [11]

Figure 7. Statewide Comprehensive Test Eigenvalues.

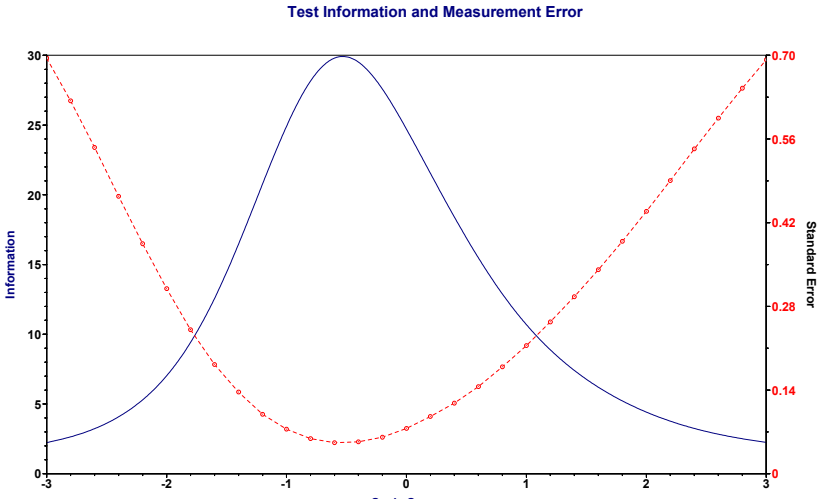


Figure 8. Statewide Comprehensive Test Information.

Item 18 is one whose distractors appear to be not providing much strength to the quality of the item. This may be due to having none of the distractors (A/2)-(D/5) related to the actual solution to the problem.

> **18:** The number $\sqrt{18+\sqrt{308}}$ can be written in the form $\sqrt{a}+\sqrt{b}$ where a and b are whole numbers and $a > b$. What is the value of $a - b$?
>
> (A/2) 8 (B/3) 12 (C/4) 16 (D/5) 20 (E/6) None of these

When one solves the problem using algebraic techniques, one finds that $a = 11$ and $b = 7$, thus

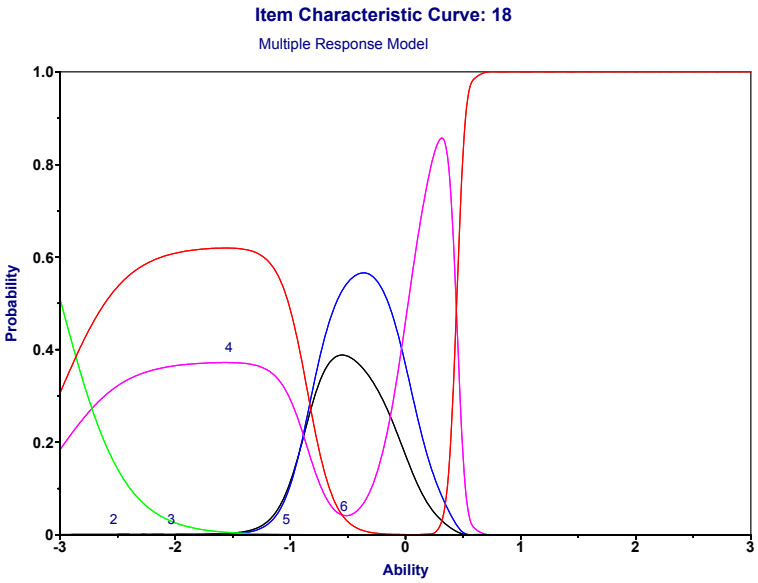


Figure 9. ICC for Item 18 of the Statewide Comprehensive Test.

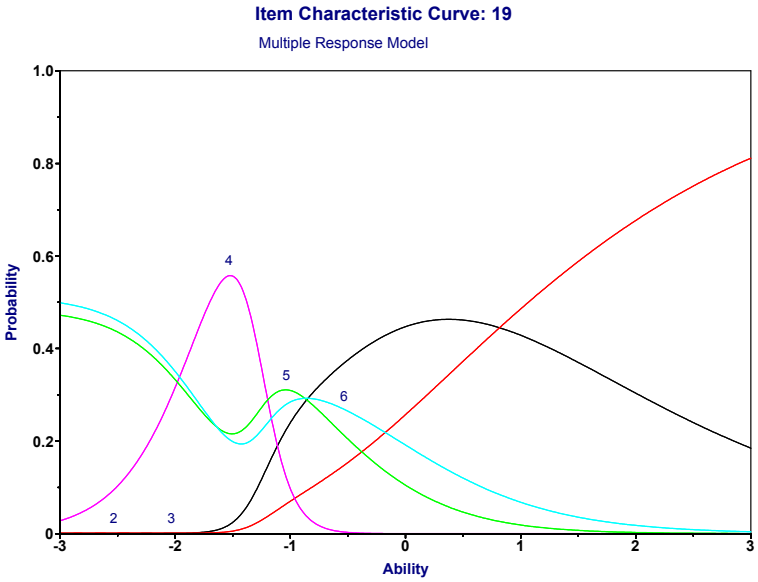


Figure 10. ICC for Item 19 of the Statewide Comprehensive Test.

$a - b = 4$. Perhaps by changing the choices to be:

$$(A/2)11 \quad (B/3)4 \quad (C/4)7$$

the amount of information that this item provides would improve.

The item characteristic curve for item 19 in Figure 10 draws attention to two interesting features. First, choices (D/5) and (E/6) have the same basic shape and they both appear to be based on guessing. Thus if one of these choices were removed, it is likely that the item would produce the same amount of information with the amount of time taken being reduced.

Another feature of the ICC for Item 19 is that choices (C/4) and (A/2) seem to target specific ability levels as distractors. This type of pattern in an item's ICC is useful if one is to use a multiple-choice model to predict the participants' ability levels. For instance, if the participant responds with (C/4), the evaluator of the test could predict that the participant falls between one and two standard deviations below the mean.

We now turn our attention to the mathematics involved in Item 19 given below.

19: Find the sum of the solutions of the equation: $\log x + \log(x + 30) = 3$.
(A/2) -30 (B/3) 20 (C/4) -40
(D/5) 30 (E/6) There are no solutions

If one uses simple mathematical manipulations without a deep understanding of the logarithmic function, the participant in the contest would state that the equation $\log x + \log(x + 30) = 3$ is equivalent to the equation $x^2 + 30x - 1000 = 0$. This latter equation has two solutions, namely 20 and -50. However, with the original equation, the values of x are restricted to positive numbers due to the domain of the logarithmic function. Therefore, this item is useful in testing algebraic skills as well as knowledge of logarithmic functions.

Overall, this analysis shows that the multiple-choice exam used for this competition was not efficient in discriminating among participants. This is due to ten of the fifty items providing limited or incorrect information due to improperly typeset items, no correct choices being given, or poor choices of distractors. However, after these ten items were removed, the instrument was shown to provide accurate and useful information.

Conclusions

For Mathematics Competitions

Overall, the analysis of multiple-choice tests involved in these two mathematics competitions showed that the tests provided enough information to discriminate ability levels at the individual participant level. However, it also showed several things to keep in mind when designing a mathematics competition that involves a multiple-choice test.

One thing that a contest organizer should keep in mind when deciding to include a multiple-choice test as a part of the competition is that multiple-choice tests provide the most information to discriminate between participants whose ability levels are near the mean. This would then imply that the multiple-choice portion of the competition would be most useful as an initial round of the competition to reduce the number of contestants

so that a later round could include free response questions or other possible formats.

Furthermore, the writer of the multiple-choice test should keep in mind that often a test could include fewer questions and still provide sufficient information about the contestants to satisfy the goals of the test. This shortening of the multiple-choice portion of the competition would enable the inclusion of other activities during the day of the competition. Also, in an effort to reduce the time involved in the multiple-choice test, fewer distractors should be included for each question since many of these distractors seem to provide overlapping information as seen above.

Finally, the competition coordinator should think purposefully about the role of a multiple-choice test in the mathematics competition. These tests are useful to achieve certain goals and can be an integral portion of the competition.

Other Applications

While item response theory is useful for such things as evaluating mathematics competitions, its possible uses among the mathematical community are extensive. One of the main potential areas to incorporate the process discussed in this study is in mathematics placement exams used by colleges and universities to determine into which mathematics classes students should enroll. Item response theory can improve these instruments by removing unimportant distractors, removing items that provide little or no information, and providing the information needed to create a computer adaptive test that can provide the necessary information about the examinee's ability using less time by only asking questions that are expected to provide information for that examinee's estimated ability level.

While this study used item response theory to evaluate a multiple-choice test, there are other models useful for analyzing open response tests using well-defined rubrics. These models can also be combined to analyze common final exams for large universities.

As mathematicians, we must not ignore the benefits that research and developments in other fields of sciences and social sciences can provide to improve the educational and administrative work that is an integral part of our lives.

References

[1] R. K. HAMBLETON, Principles and selected applications of item response theory, *Educational measurement*, (R. L. Linn, ed.), 3rd edition, Macmillan, New York, 1989, pages 147–200.

[2] R. K. HAMBLETON, H. SWAMINATHAN, and H. J. ROGER, *Fundamentals of Item Response Theory*, Sage, Newbury Park, CA, 1991.

[3] L. F. KLOSINSKI, G. L. ALEXANDERSON, and L. C. LARSON, The Sixty-Seventh William Lowell Putnam Mathematical Competition, *American Mathematical Monthly* **114**(8) (October 2007), 714–724.

[4] F. M. LORD and M. R. NOVICK, *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, MA, 1968.

[5] MAA, *What's What: AMC's Programs and Contests*, Retrieved October 24, 2007, from `http://www.unl.edu/amc/whatswhat.shtml`.

[6] WILLIAM LOWELL PUTNAM, A suggestion for increasing the undergraduate interest in studies, *Harvard Graduates' Magazine* **30** (December 1921), 210–212.

[7] M. D. RECKASE, Unifactor latent trait models applied to multi-factor tests: Results and implications, *Journal of Educational Statistics* **4** (1979), 207–230.

[8] BETTY BAKER REITER, *A survey of the effects of Mathcounts and Olympiad mathematics competitions on students' attitudes and feelings about mathematics, science, and careers*, Ph.D. thesis, University of South Carolina, 1999.

[9] EDWARD E. ROSKAM, Models for speed and time-limit tests, *Handbook of Modern Item Response Theory*, (Wim J. van der Linden and Ronald K. Hambleton, eds.), Springer-Verlag, New York, 1997, chapter 11, pages 187–208.

[10] LAURIE SOMERS and SUSAN CALLAN, An examination of science and mathematics competitions. Technical report, Westat, Rockville, MD, June 1999, retrieved October 24, 2007, from `http://www.wmich.edu/evalctr/competitions.pdf`.

[11] SPSS 13.0 for Windows, 2004.

[12] R. F. SUBOTNIK, R. A. DUSHCL, and E. H. SELMAN, Retention and attrition of science talent: A longitudinal study of Westinghouse Science Talent search winners, *International Journal of Science Education* **15** (1993), 61–72.

[13] R. F. SUBOTNIK, A. D. MISERANDINO, and P. OLSZEWSKI-KUBILIUS, Implications of the Olympiad studies for the development of mathematical talent in schools, April 1996, paper presented at the annual meeting of the American Educational Research Association in New York, NY.

[14] DAVID THISSEN, *MULTILOG for Windows (version 7.0)*, Scientific Software International, Inc., Mooresville, IN, 2003.

[15] DAVID THISSEN and LYNNE STEINBERG, A response model for multiple-choice items, *Handbook of Modern Item Response Theory*, (Wim J. van der Linden and Ronald K. Hambleton, eds.), Springer-Verlag, New York, 1997, chapter 3, pages 51–66.

[16] NURA D. TURNER, A historical sketch of the Olympiads, national and international, *American Mathematical Monthly* **85**(10) (December 1978), 802–807.

[17] ———, A historical sketch of the Olympiads: U.S.A. and international, *College Mathematics Journal* **16**(5) (November 1985), 330–335.

[18] WIM J. VAN DER LINDEN and RONALD K. HAMBLETON, *Handbook of Modern Item Response Theory*, Springer-Verlag, New York, 1997.

[19] ———, Item response theory: Brief history, common models, and extensions, *Handbook of Modern Item Response Theory*, (Wim J. van der Linden and Ronald K. Hambleton, eds.), Springer-Verlag, New York, 1997, chapter 1, pages 1–28.

[20] N. D. VERHELST, H. H. F. M. VERSTRALEN, and M. G .H. JANSEN, A logistic model for time-limit tests, *Handbook of Modern Item Response Theory*, (Wim J. van der Linden and Ronald K. Hambleton, eds.), Springer-Verlag, New York, 1997, chapter 10, pages 169–186.