Objective Scoring for Computing Competition Tasks

Graeme Kemkes, Troy Vasiga, and Gordon Cormack
University of Waterloo
Waterloo, Ontario, Canada

Abstract

Computing competitions like the International Olympiad in Informatics (IOI) typically pose several problems that contestants are required to solve by writing a program. The program is tested automatically on several sets of input data to determine whether or not it computes the correct answer within specified time and memory limits. We consider the controversy of whether and how to award partial credit for programs that fail some of the tests. Using item response theory, we analyze the degree to which the scores from these automatic tests, separately and in various combinations, truly reflect the contestants' achievement.

1 Introduction

The International Olympiad in Informatics (IOI) [1] is a programming contest for secondary school students. Competitors are given a set of tasks (typically six) for which they are to program solutions. The solutions are subject to a number of automatic tests; each test presents the program with a set of data and evaluates automatically whether or not the program produces the correct output within specified time and memory limits. The vast majority of contestants' submissions fail at least one test; a substantial fraction (one-third or more) fail at least one test run for every solution that they submit.

The question we address here is: how should failed tests be scored? Our thesis is that test runs should be categorized according to objective criteria, that points should be awarded according to category, and that a program failing even one test in a category should be awarded no points for the category. Our investigative approach demands careful attention to the difficulty each category; more precicely to the *discrimination* of each category – how much more likely is a contestant with more ability than another to be able to prepare a submission that achieves points in the category? We argue that current practice yields tests with too little diversity of difficulty – too difficult for most and too easy for a few – and hence with poor discrimination among the majority of contestants. We argue that the solution is not to increase the median level of difficulty, as was the trend for several years. We argue that current practice of awarding marks in proportion to the number of tests passed impairs discrimination and underrepresents the difficulty of the tasks. Further, we propose an alternative testing policy which yields much better discrimination ability.

Each test run uses data selected in advance by the contest designers to test various possible sources of error or inefficiency. The data for the test runs is secret; the contestants may not see it before their programs are scored. Prior to IOI 2004, contestants were informed only of the criteria necessary to achieve full marks; that is, the correctness criteria, the maximum input data size, and the memory and time limits. Submissions that were incorrect or inefficient – the vast majority of submissions – could be expected to receive a partial score to the extent that some some test cases would fail to detect incorrect or inefficient programs. A contestant submitting an incorrect or inefficient program had no way of predicting what partial score it might achieve, and therefore little guidance as to how to try to maximize his or her score. To mitigate this problem, the *fifty-percent rule* was implemented in 2004; it states that (weaker) criteria must be specified in the task statement which, if met, will yield a score of at least 50%. Notwithstanding the fifty-percent rule, contestants still experience a great deal of uncertainty as to the score that they might expect to receive for a particular effort.

In a previous study, Cormack [3] suggests that partial scoring be eliminated; that a score should be awarded for a particular group of tests (that is, all the fifty-percent-rule cases for a given task, or all the non-fifty-percent-rule) only if the program passes all of them. Such scoring appears to discriminate better among competitors with high ability, but fails to discriminate at all among the bottom third, who receive a score of 0 under the proposed scheme. It has also been suggested that such all-or-nothing scoring excessively penalizes contestants who make one small mistake, resulting in a zero score for a group of tests. We propose an alternative scheme which accomplishes discrimination at both the upper and lower ability levels.

2 Overview of Item Response Theory

Item response theory (IRT) investigates the ability of a test to discriminate among students with various levels of ability. In contrast, classical theory (and common practice) considers only the difficulty of a test case – as evidenced by the overall success rate – in assessing its contribution to scoring. According to item response theory, a test is composed of several items, each of which is graded as correct or incorrect. For each item, the probability that a student answers correctly is modelled as a function of her ability θ,

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}.$$

This function is called the *item response curve.* The parameters a and b depend on properties of the item: a is the discrimination parameter, b is the difficulty parameter. A good test contains items with a variety of difficulty parameters, covering the range of abilities of the students. The discrimination parameter of each item should be high.

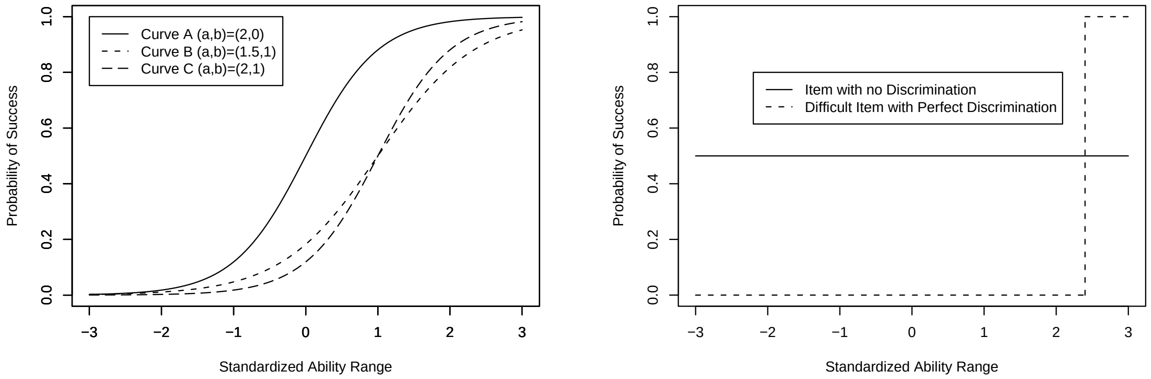

Figure 1: Item Characteristic Curves

These parameters have a quantitative interpretation. The difficulty parameter is the ability level at which the probability of success is 0.5, which is also the inflection point for this curve. The slope of the curve at this point b is $a/4$, which is simply a constant factor times the discrimination parameter. In other words, we concern ourselves with describing the item characteristic curve in terms of the inflection point and the slope at the inflection point. Some sample item response curves are shown in figure 1. On the left are three item characteristic curves (A, B, C) such that

- Curves A and C have the same discrimination parameter (i.e., slope at their respective inflection points),
- Curves B and C have the same difficulty parameter (i.e., inflection point),
- Curve A has a lower difficulty parameter than Curve C (i.e., the curve is shifted to the left),
- Curve B has a lower discrimination parameter than Curve C.

The curves on the right are for items with no discrimination ($a = 0$) and perfect discrimination ($a \to \infty$).

For more information about IRT, see the introductory textbooks [2, 5].

3 Scoring schemes

As stated in the introduction, IOI 2005[1] had 6 problems for students to solve. Each problem had two batches of test cases: *Easy* test cases (the so-called "50% rule" cases), and *Hard* test cases.

We will analyze the IOI 2005 results using three different scoring schemes. These are

1. IOI scoring: as at the IOI, each test case is graded as correct (4 or 5 points) or incorrect (0 points). The score for each batch is simply the sum of the scores of the test cases in that batch. The perfect score on each batch is 50.

2. All-or-Nothing scoring: for each batch, we assign a score of 50 if the program produced correct output on every test case in that batch; otherwise, we assign a score of 0.

3. Significant Progress scoring: for each batch, we assign a score of 50 if the program produced correct output on any test case in that batch; otherwise we assign a score of 0.

The All-or-Nothing scoring scheme is used in well-known programming contests such as ACM ICPC and TopCoder. (See [4] for an overview.) It has also been proposed for use at the IOI [3], but it was noted that many students would receive a total score of zero. Motivated by this observation, we developed the Significant Progress scoring scheme to reward students who make non-trivial progress on any batch.

In order to apply IRT, we need a measure of each student's ability. The only measures available to us are the results from IOI 2005. For a student at percentile rank p, we define her ability to be her standardized rank $\Phi^{-1}(p)$, where Φ is the cumulative normal distribution. This gives a normal (Gaussian) distribution of abilities with mean 0 and standard deviation 1. The item response curves are estimated using logistic regression. The individual item response curves are summed (with appropriate weights) to give the expected score as a function of ability.

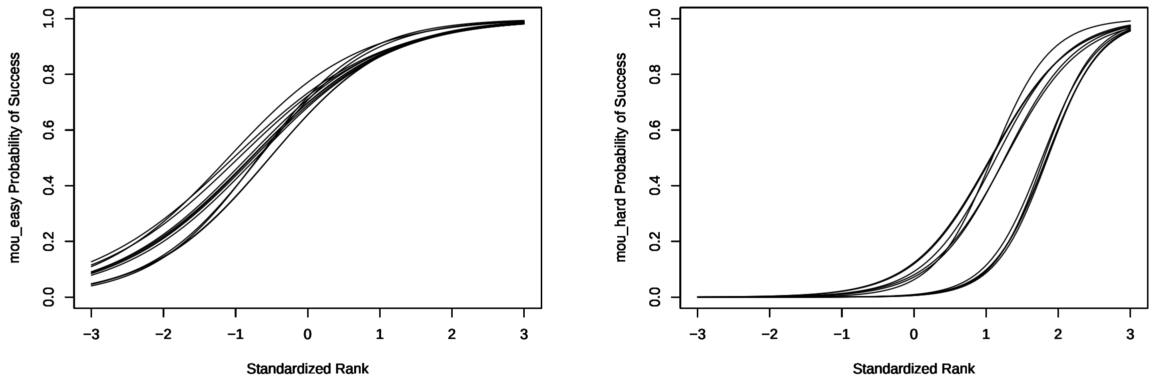

Figure 2: Item Response Curves for Mountain Task (easy, hard), IOI Scoring

[1] It is not the intention of this paper to criticize the IOI 2005 organizing committee in any way. Our evaluation is being peformed on this IOI since it is the only available data set at the time of writing. The authors would be very willing to perform a similar analysis on data from other IOI competitions.

3.1 IOI scoring

According to the IOI scoring scheme, each test case is an item for which we fit an item response curve. The item response curves for all of the test cases of the "Mountain" task are shown in figure 2. Notice that the test cases of the easy batch (left-hand graph) all have similar shapes; their discrimination and difficulty parameters are very similar. These difficulty parameters are, of course, lower than the difficulty parameters of the hard batch (right-hand graph), as the curves for the hard test cases are further to the right.

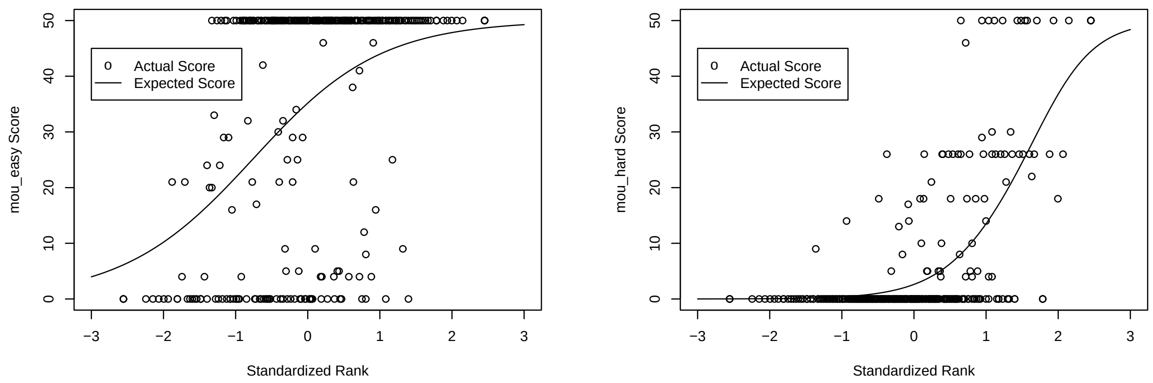

Figure 3: Expected Score for Mountain Task (easy, hard), IOI Scoring

The expected score for a batch is a function of ability. We compute this function by taking the sum of the item response curves of its test cases, weighted by the value (4 or 5) of each test case. The resulting curves, together with the actual contestants' IOI scores, for the easy batch and hard batch of the Mountain task are shown in figure 3.

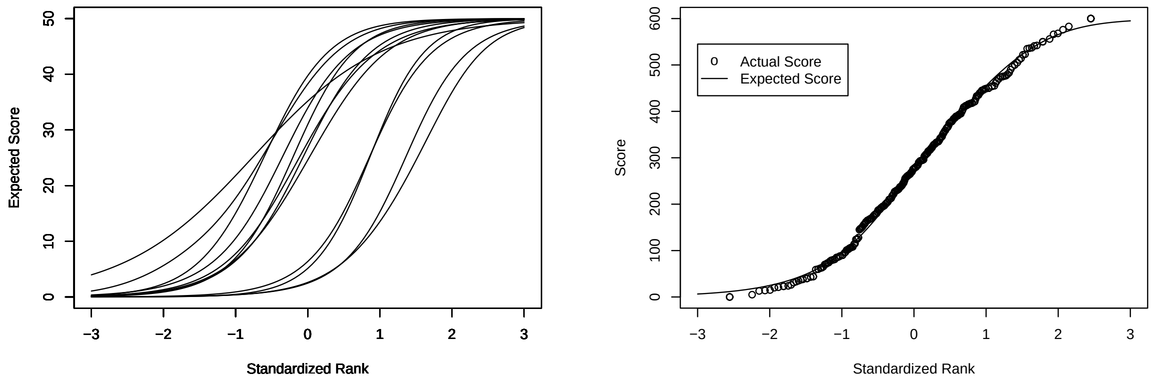

Figure 4: Expected Batch Scores and Total Scores, IOI Scoring

The IOI 2005 competition was composed of 12 batches (6 problems with 2 batches per problem). We repeat the above steps for all of the other batches to get an expected score curve for each batch. These curves are shown in figure 4 (left-hand graph). Finally, in the right-hand graph, we sum these twelve expected score curves to get the expected overall score. The actual scores are also plotted.

3.2 All-or-Nothing scoring

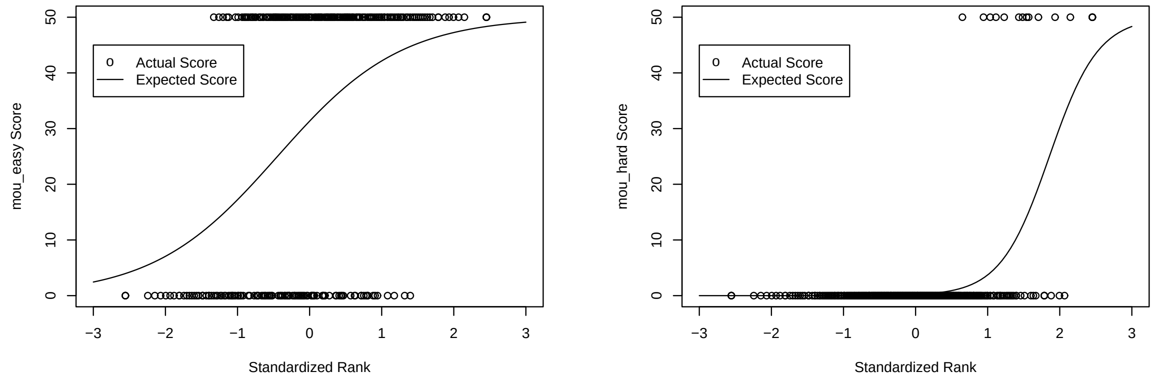

Figure 5: Item Response Curves for Mountain Task (easy, hard), All-or-Nothing Scoring

For the All-or-nothing scoring scheme, each batch of test cases is an item. The batch is marked correct if the student passes every case in that batch; otherwise the batch is marked incorrect.

To apply IRT to this scoring scheme, we fit an item response curve for each batch. The item response curves (scaled to give a maximum score of 50) for the easy Mountain batch and the hard Mountain batch are shown in figure 5. The actual scores awarded under this scheme are also plotted.

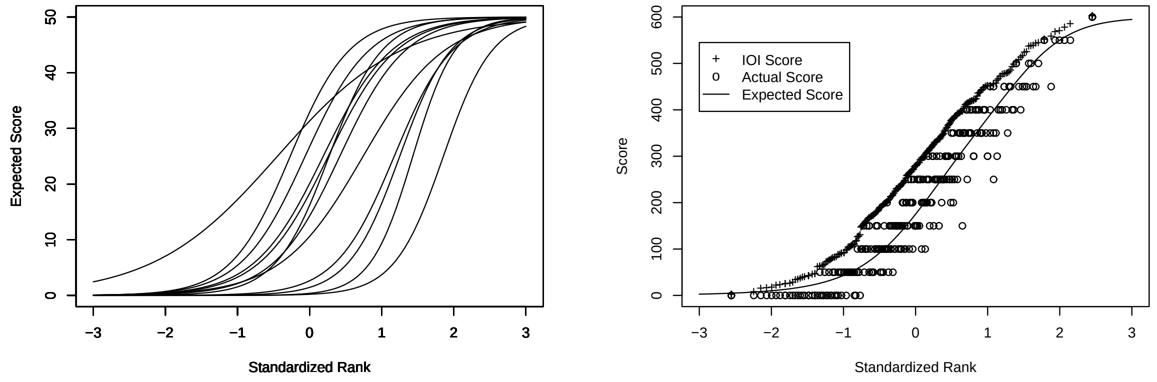

Figure 6: Expected Batch Scores and Total Scores, All-or-nothing Scoring

Repeating this procedure for each batch, we get the 12 curves shown in figure 4 (left-hand graph). When we compare these curves to the ones for the IOI scoring scheme (in the previous section) we see that the discrimination parameters are similar. The difficulty parameters of these curves are noticeably higher. The right-hand graph shows the expected total score, formed by summing these curves. The actual scores awarded under this scheme are also plotted, along with the IOI scores. The IOI scores are noticeably higher.

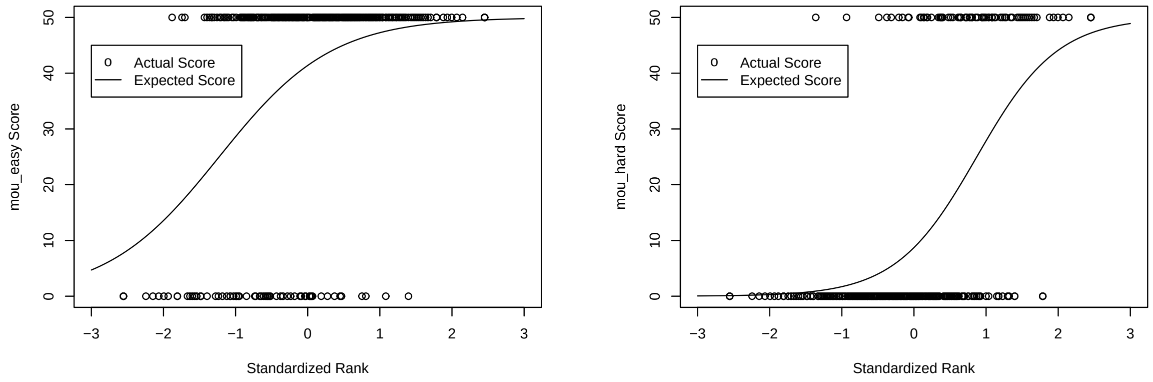

Figure 7: Item Response Curves for Mountain Task (easy, hard), Significant Progress Scoring

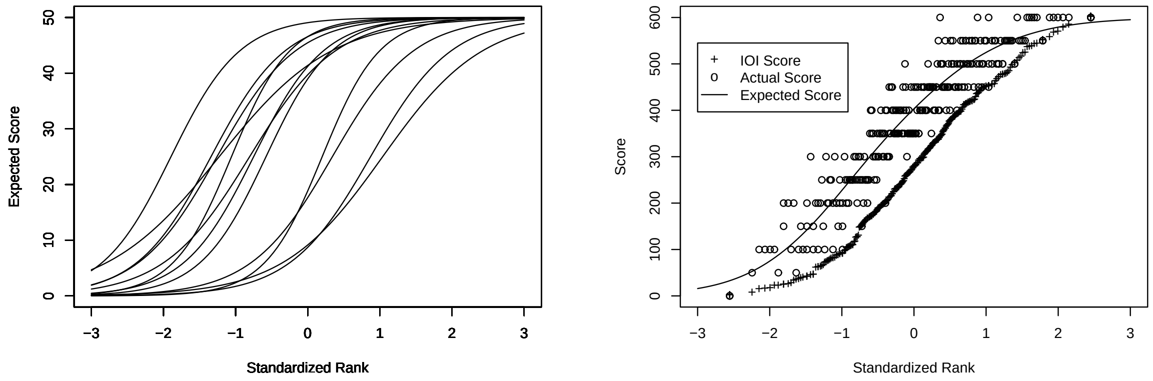

Figure 8: Expected Batch Scores and Total Scores, Significant Progress Scoring

3.3 Significant Progress scoring scheme

The significant progress scoring scheme marks a batch as correct if the student passes any test case in that batch. For each batch, we fit an item response curve. The item response curves (scaled to give a maximum score of 50) for the easy Mountain batch and the hard mountain batch are shown in figure 7, together with the actual scores awarded under this scheme. The curves for all twelve batches are shown in figure 8 (left-hand graph). When compared with the IOI scoring scheme, these curves have similar discrimination and lower difficulty. The curve of the expected total score, shown in the right-hand graph, shows scores that are much higher than the IOI scores.

3.4 Combining All-or-nothing with Significant Progress

The All-or-Nothing and Significant Progress scoring schemes give good discrimination among high and low ability levels, respectively. It makes sense to use them both in a Combined scoring scheme. Each batch produces two items: one is marked as correct if any one test case is solved correctly, the other is marked as correct if all of the cases are solved. The resulting 24 item response curves are shown in figure 9 (left-hand graph); their sum produces the expected score curve shown on the right. We see that this Combined scoring

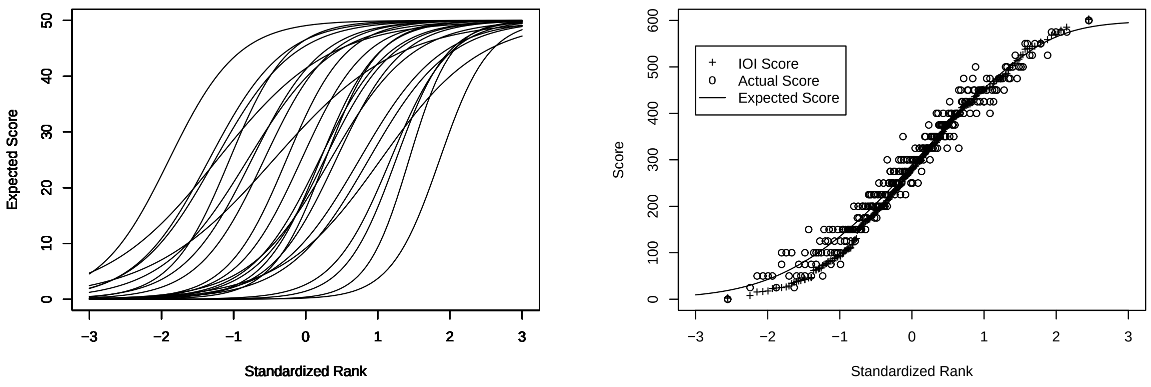

Figure 9: Expected Batch Scores and Total Scores, Combined All-or-nothing and Significant Progress Scoring

scheme yields good discrimination over a broad range of ability levels and produces an expected score curve that is nearly linear over the entire spectrum.

To quantify these observations, we computed the range of difficulty parameters for the item response curves of the IOI scoring scheme and the Combined scoring scheme. For the IOI scoring scheme, the parameters range from a low of -1.896 to a high of 1.855. For the Combined scoring scheme the range is nearly identical, from -1.896 to 1.856.

3.5 Partial Credit

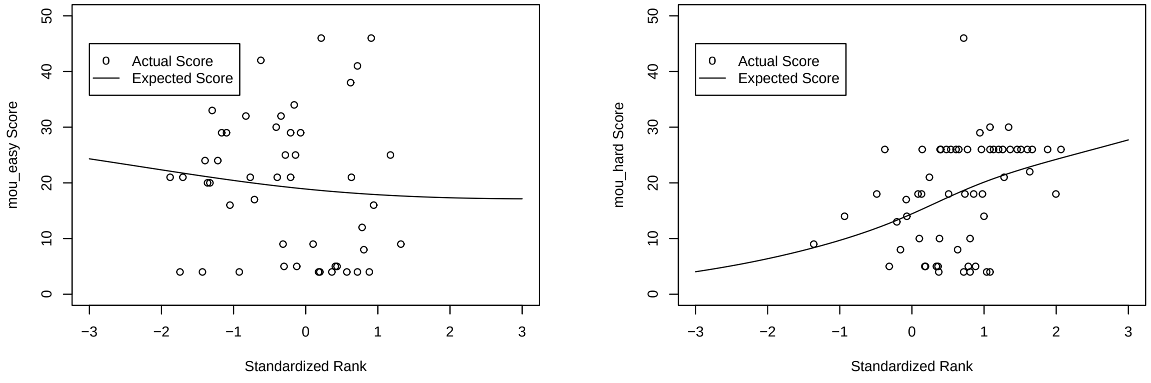

Figure 10: Expected Batch Scores (easy and hard), Mountain Task, Part Marks Only

The Combined scoring scheme uses only two pieces of information about each batch: “Did the student score 0?” and “Did the student earn a perfect score?” The scheme does not use the partial scores for each batch. Do the partial scores provide any additional discrimination? To investigate this question we performed an additional experiment. For each batch, we isolated the population of students whose score was neither zero nor perfect. Then we created item response curves for each test case. By summing these curves (with appropriate weights) we get the expected score on this batch for this population. The expected score curves for the easy and hard batches of the Mountain task are shown in figure 10. They appear to have little

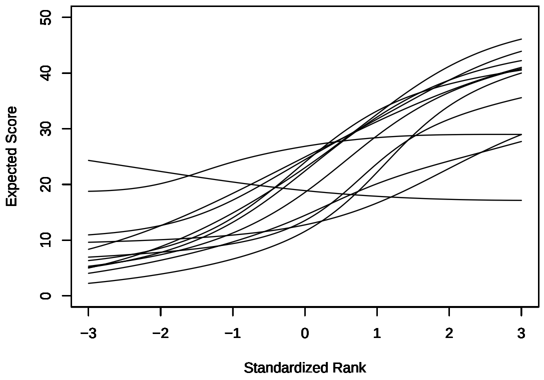

Figure 11: Expected Batch Scores, Part Marks Only

discriminatory value. In fact, the curve for the easy batch suggests a negative correlation between ability and score! The expected score curves for all of the batches are shown in figure 11. In general, these appear to have poor discriminatory value.

4 Discussion

Presently there is no practical alternative to automated scoring at computing competitions such as the IOI. It is difficult for an automated system to determine the magnitude of the flaws in incorrect programs, which form the majority of the submissions at the IOI. Using item response theory, we have examined the discrimination and difficulty of the tasks at IOI 2005 under various automated scoring schemes. The current scheme awards scores in proportion to the number of test cases for which the program produces correct output. This suffers from the drawback that there is no objective standard for measuring the number of cases solved by a program; it depends entirely on the authors of the test cases. Furthermore, we have shown that this scoring scheme yields poor discrimination among the submissions which receive partial credit.

We examined a Combined scoring scheme which awards credit for a batch of test cases according to whether the program solves all, some, or none of the cases. This scheme gives good discrimination over a broad range of ability levels. We argue that it assesses objective criteria which could (and should) be explicitly stated in the problem statements. The scoring would be improved by adding more batches of various difficulties, each evaluating specific, explicitly-stated criteria. Ideally, there should be an objective standard for the awarding of every mark in the contest.

This scheme makes the scoring more objective: two judges creating their own test data would likely award similar marks when evaluating a program. Contestants could predict the score that would be awarded for various submissions. Weaker students would have achievable goals. By making the scoring more predictable at all difficulty levels, contestants will have more confidence in the scoring and a more enjoyable contest experience.

References

[1] The International Olympiad of Informatics (IOI), `http://www.ioinformatics.org/` (2006).

[2] Baker, F. *The Basics of Item Response Theory* ERIC Clearinghouse on Assessment and Evaluation, College Park, MD (2001). (Also available at `http://edres.org/irt/baker/`)

[3] Cormack, G. Random Factors in IOI 2005 Test Case Scoring. In *Perspectives on Computer Science Competitions for (High School) Students* (Dagstuhl, 2006).

[4] Cormack, G., Kemkes, G., Munro I., and Vasiga T. Structure, Scoring and Purpose of Computing Competitions, *Informatics in Education* (**5**), 2006, 1-22.

[5] Partchev, I. Visual guide to item response theory. Available at: `http://www2.uni-jena.de/svw/metheval/irt/VisualIRT.pdf`