Random Factors in IOI 2005 Test Case Scoring

Gordon CORMACK

David R. Cheriton School of Computer Science, University of Waterloo
Waterloo Ontario N2L 3G1, Canada
e-mail: gvcormac@uwaterloo.ca

Received: February 2006

Abstract. We examine the precision with which the cumulative score from a suite of test cases ranks participants in the International Olympiad in Informatics (IOI). Our concern is the ability of these scores to reflect achievement at all levels, as opposed to reflecting chance or arbitrary factors involved in composing the test suite. Test cases are assumed to be drawn from an infinite population of similar cases; variance in standardized rank is estimated by the bootstrap method and used to compute confidence intervals which contain the hypothetical *true ranking* with 95% probability. We examine the relative contribution of easy (so-called *fifty-percent rule*) cases and hard cases to the overall ranking. Empirical results based on IOI 2005 suggest that easy and hard cases are both material to the ranking, but the proportion of each is unimportant.

Key words: informatics olympiad, programming contest, classical test theory.

1. Introduction

The International Olympiad in Informatics (IOI) is a two-day competition in which secondary-school students are presented with several tasks for which they are required to program solutions in an algorithmic language. Students' programs are run on a suite of test cases. Each test case presents to the program one or more instances of one of the tasks[1] which must be solved within a specified time limit; if the program is successful, a score is awarded. A student's score is the sum of the scores awarded for all test cases. IOI rules mandate that approximately the top-scoring 1/12 receive gold medals, the next 1/6 receive silver, the next 1/3 receive bronze, subject to the overall constraint that no more than 1/2 receive a medal. Although not specified by the rules, total rankings have traditionally been published, and the winner has been specifically recognized.

Many games and contests are designed explicitly to include chance as a determining factor; even those that are not – such as the IOI – cannot avoid it entirely. Nevertheless, we seek to minimize its role, so that IOI scores better reflect true achievement. Many chance factors come into play, which we broadly characterize as external and internal. External factors may include contestants' health, prior experiences, distractions and so on. Internal factors are those totally within the control of the competition designers. Although we cannot hope to measure all elements of chance, we know that the overall impact of chance

[1] The problem instances, except for a simple sample case, are unknown to the student prior to evaluation.

cannot be less than any particular one element. It is therefore fruitful to seek to identify and to mitigate the principal individual elements of chance. Our current investigation concerns itself with one specific internal factor – the choice and scoring of test cases.

Factors other than chance may compromise the extent to which IOI scores reflect true achievement or ability. Arbitrary task selection, inaccurate or misleading problem statements, incomplete or inaccurate test cases and subjective judging may compromise students' ability to demonstrate achievement, or the score's reflection of that achievement. Students' perceptions of these factors – whether accurate or not – may also affect their ability to demonstrate achievement (Fisher and Cox, 2006). While these factors are not really attributable to chance, they may sometimes be measured as if they were. Those whose effect exceeds the magnitude of true chance factors bear further investigation; those with smaller effect may be inconsequential. Furthermore, publication of such measurements serves to inform students and coaches, perhaps mitigating factors relating to perception.

2. IOI Test Cases and Scoring

The 2005 IOI competition, as typical, consisted of six tasks to be solved in two five-hour sessions on separate competition days. For each task, each student submitted a separate program which was run on several test cases. Four of the tasks had twenty test cases; two had twenty-four. For each test case a fixed score (4 or 5) was assigned such that the scores for each task summed to 100. This score was awarded to each program yielding a correct result for the case within the specified resource limits; otherwise no score was awarded.

For each task test cases were divided into two equal groups according to the de facto *fifty-percent rule (Yakovenko, 2006).* The easy test cases have smaller test data – dimensions of which are specified explicitly in the task statement – so as to be amenable to simpler, more obvious solutions. The hard cases have larger test data so as to provide more challenge.

Fig. 1 shows the distribution of total scores at IOI. The mode is not well defined, and there is a cluster of zero or near-zero scores. The top quartile shows a diminishing tail, with a handful of participants receiving a perfect or near-perfect score. Figs. 2 and 3 show

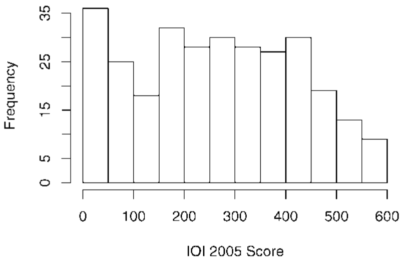

Fig. 1. IOI 2005: score distribution.

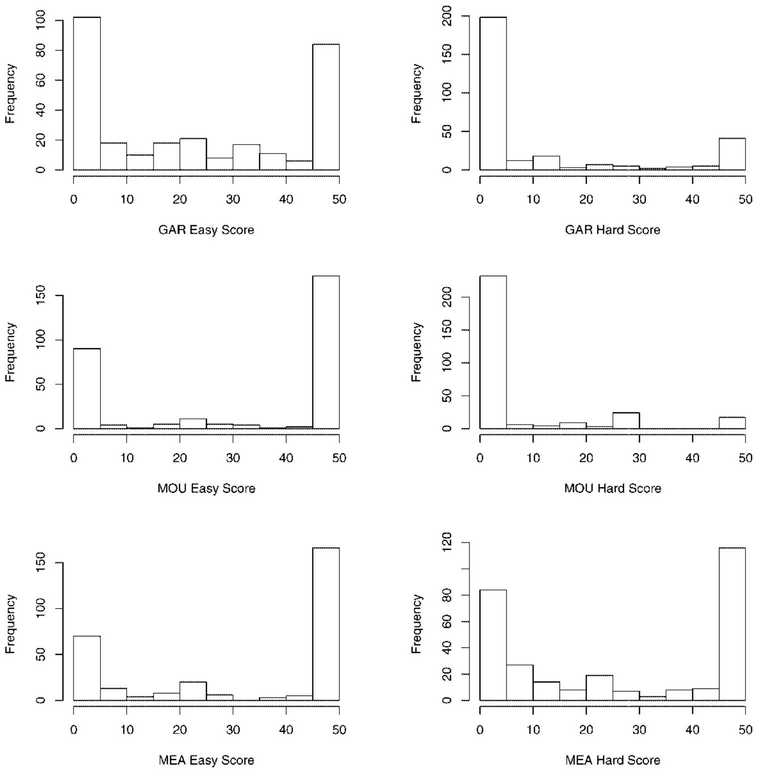

Fig. 2. Day 1: task score distribution.

Table 1

Zero and perfect scores $(n = 295)$

Task	Easy Cases		Hard Cases	
	Zero	Perfect	Zero	Perfect
GAR	92	84	172	41
MOU	77	170	220	16
MEA	63	166	67	116
BIR	57	125	181	48
REC	33	147	104	105
RIV	93	119	224	50
Test cases for all tasks	15	21	39	4

the distributions of scores[2] separated by task and by the fifty-percent rule. Table 1 shows that the vast majority of participants receive either zero or a perfect score on each group of test cases; a substantial number of participants receive zero scores even on the easy

[2] Note that the scales are different, and that the extreme bars may include scores that are neither zero nor perfect. Table 1 shows the exact number of zero and perfect scores for each task in each category, as well as over all tasks in each category.

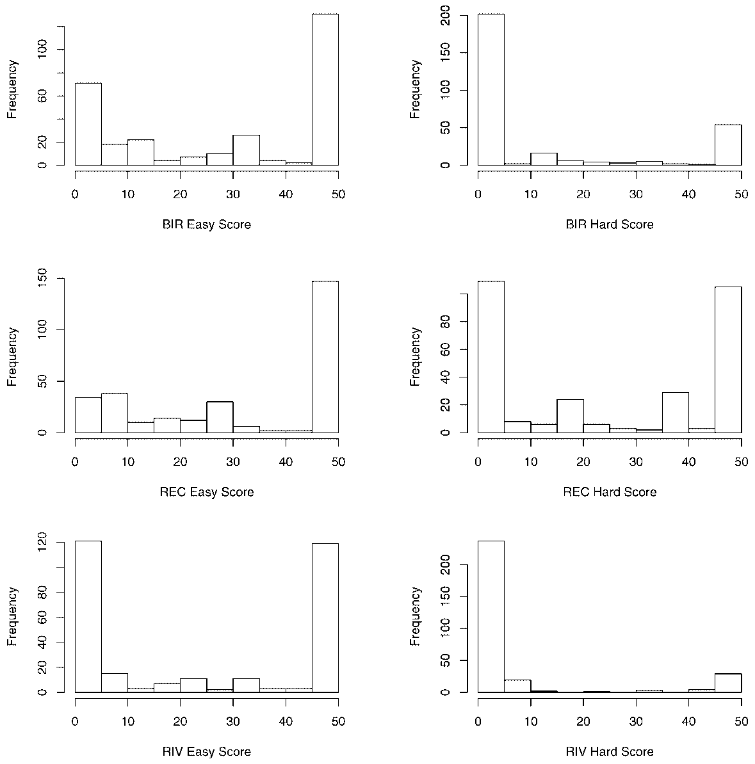

Fig. 3. Day 2: task score distribution.

cases; half or fewer receive perfect scores. A small fraction receive partial scores within a particular group.

3. Measuring Achievement

Achievement is an abstraction that can be measured only indirectly(Baker, 2001). To this end, IOI poses tasks which, if solved correctly, indicate achievement. Test cases are used to indicate the extent to which tasks are solved correctly. Test-case scores for a student s are summed to yield a raw score R_s which is taken as an indicator of the student's true achievement. A single raw score has little meaning; however we take $R_{s_1} > R_{s_2}$ as evidence that s_1's true achievement exceeds s_2's.

The issue of whether tasks and test cases reflect achievement is one of *validity.* The likelihood that $R_{s_1} > R_{s_2}$, given s_1 whose true achievement exceeds that of s_2, is one of *precision.* IOI scoring must be both valid and precise to be an accurate measure of achievement. While validity is appropriately the subject of current debate (Cormack *et al.*, 2006; Forisek, 2006; Verhoeff, 2006; Yakovenko, 2006), our primary concern is the precision of IOI scoring.

Since we are concerned only with relative achievement, the magnitude and distribution of raw scores are unimportant. For each R_s we compute an equivalent standardized score[3] $N_s = \Phi^{-1}(r_s)$ where Φ is the cumulative normal distribution, $r_s = \frac{|\{s' \in P | R_s > R_{s'}\}| + \frac{1}{2}|\{s' \in P | R_s = R_{s'}\}|}{|P|+1}$ and P is the set of all participants. The standardized score, also known as z-score, has a normal (Gaussian) distribution with standard deviation $\sigma = 1$. Although z-scores cover the interval $[-\infty, \infty]$, we report the range $[-3, 3]$ which contains all values of practical interest in the current context. IOI gold medals are awarded[4] for standardized scores in the range $(0.967, \infty]$; silver for the range $(0.674, 0.967]$; bronze for $(0, 0.674]$.

We define the hypothetical true standardized score T_s to represent achievement: $T_s = \Phi^{-1}(\rho_s)$ where ρ_s is the proportion of students actually having lower achievement than s. Our scores may be considered accurate to the extent that we can argue that $N_s \approx T_s$ for all $s \in P$. Classical test theory (Novick, 1966) assumes validity and that chance is the only source of error. It assumes, furthermore, that the magnitude of the error is the same for all s. More formally, $N_s = T_s + E$ where E is a random variable with expected value 0 and standard deviation σ_E. Given σ_E – the standard deviation of the error – one may compute for any given student s a confidence interval $N_S \pm 1.96\sigma_E$ that, with 95% probability, contains T_s.

4. Modelling Test Case Selection

We assume that the test cases are drawn from an infinite hypothetical population[5] of cases materially similar to the ones used at IOI 2005. The test cases actually used are considered to be a random sample of this population; on another day a different sample might have been chosen.

We consider two test cases to be *materially similar* if, for every student s, the two test cases yield the same score. Our hypothetical population contains cases materially similar to those actually used, and in the same proportion. That is, if there are t test cases, our population has in proportion $\frac{1}{t}$ cases materially similar to each. For a student s receiving actual score N_s, we wish to estimate N'_s, the score that the student might have received had a different set of test cases been drawn from the hypothetical population. We assume that the true achievement $T_s = \mathcal{E}(N'_s)$, the expected value of N'_s. That is, $N_s = N'_s + E_s$. Under the classical test theory assumption that $E_{s_1} = E_{s_2} = E$ for all s_1 and s_2, we regard each E_s as a separate estimate of E and compute the approximation $E' = \frac{1}{|P|}\sum_{s \in P} E_s \approx E$. Our estimate of the standard deviation of the error is therefore $\sigma_{E'} = \left(\frac{1}{|P|}\sum_{s \in P} \sigma^2_{E_s}\right)^{0.5}$.

[3] Normalization is common and necessary (Novick, 1966) to transform the score so as to have an appropriate distribution for analysis. Neither absolute score nor percentile rank has such such a distribution, being subject to hard upper and lower limits.

[4] Subject to minor ad-hoc adjustments.

[5] As characterized by Fisher (Fisher, 1925; Lenhard, 2006).

5. Bootstrap Simulation

The bootstrap (Efron and Tsibirani, 1994) is used to simulate IOI scoring using different sets of test cases. The bootstrap uses the actual test cases and results as a proxy for an infinite population of test cases – for each sample a number of test cases are selected, with replacement from the actual test cases. That is, a given case may be selected more than once. In terms of our model these multiple selections represent different but materially similar cases, yielding identical scores for all s. For each sample and for each s, we compute an instance of N'_s by summing the scores and standardizing. The standard deviation of these examples closely approximates $\sigma_{N'_s}$ and hence σ_{E_s} (Efron and Tsibirani, 1994).

Fig. 4 shows achievement and standard deviation of the error as a function of IOI scores, as estimated using 100 bootstrap samples. The estimate is given as a solid point; the bounding lines are 95% confidence intervals.[6] The left curve – plotted in terms of percentile rank – illustrates the range of true ranks likely to account for any given ρ_s. The confidence intervals for many students includes a medal cutoffs[7]; however, none includes two such boundaries. The right curve shows the standard deviation of the error as a function of standardized score. From this representation we may assess the magnitude of the error as a function of score. Except for the end points, which represent zero or perfect scores for which we are unable to compute σ_{E_s}, our assumption of a common error term appears to hold. Our estimate for the standard deviation of the error is $\sigma_{E'} = 0.11$.

Figs. 7 and 8 show the effect of eliminating the easy and hard cases entirely. Eliminating the easy cases introduces a very large error – much larger than chance – at the low end of the achievement range. Eliminating the hard cases introduces similar error at the high end.

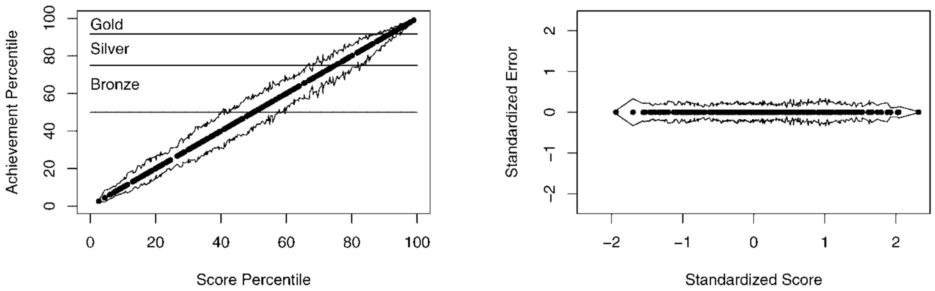

Fig. 4. Percentile and standardized 95% confidence intervals.

[6] Note that the confidence intervals are vertical, not horizontal or perpendicular to the curve.

[7] 35% of gold medallists' confidence intervals contain the gold/silver cutoff; 25% of silver medallists' contain the gold/silver cutoff while 40% contain the silver/bronze cutoff; 30% of bronze medallists' contain the silver/bronze cutoff while 30% contain the no-medal cutoff. 20% of the confidence intervals for those awarded no medal contained the bronze medal cutoff.

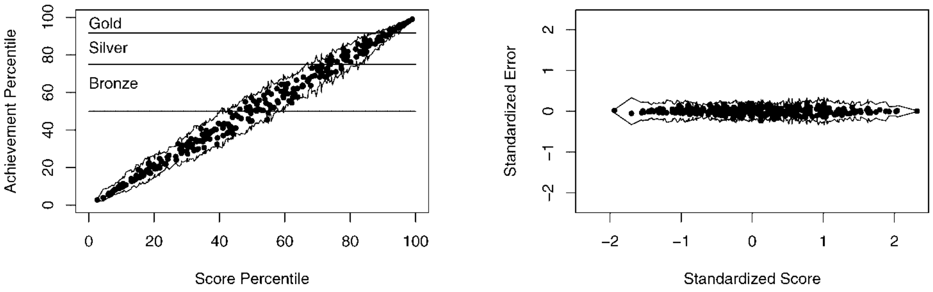

Fig. 5. Easy-to-hard test cases in ratio 1:2.

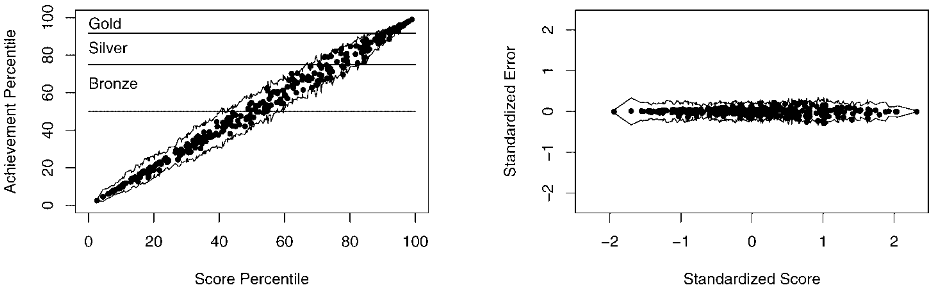

Fig. 6. Easy-to-hard test cases in ratio 2:1.

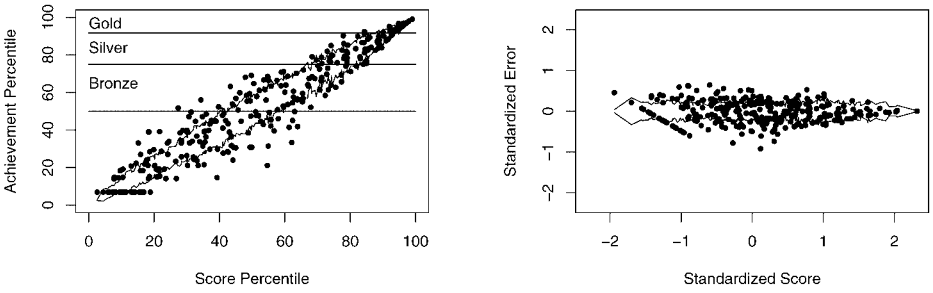

Fig. 7. No easy cases.

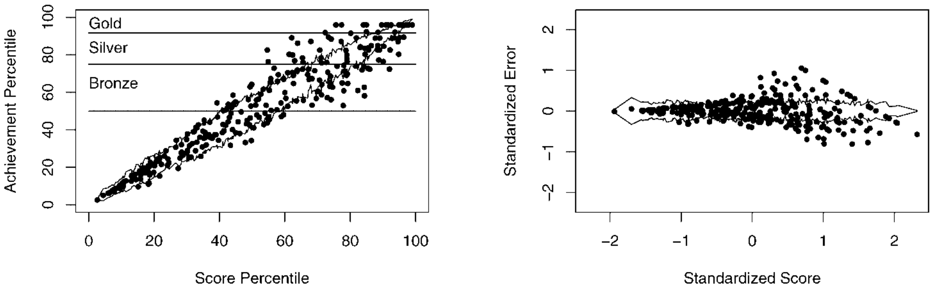

Fig. 8. No hard cases.

6. Sensitivity to the Fifty-Percent Rule

We investigated the effect of changing the proportion of easy and hard test cases. Such a change is arbitrary rather than random; however, we may extend the model to measure this effect. Let F_s^r be a random variable representing the effect of changing the easy-to-hard ratio to r. That is, $N_s = N_s' + E_s + F_s^r$. Suppose that $\mathcal{E}(F_s^r) > 0$ for some student s. Our standardization procedure implies that there is a corresponding student s' such that $\mathcal{E}(F_{s'}^r) < 0$. So a common effect may be modelled by $F^r = \frac{1}{|P|}\sum_{s \in P} F_s^r$ where $\mathcal{E}(F^r) \approx 0$ and $\sigma_{F^r} \approx \left(\frac{1}{|P|-1}\sum_{s \in R} \mathcal{E}(F_s^r)^2\right)^{0.5}$. The combined error due to chance and altering the ratio of easy-to-hard cases is $E + F^r$ with standard deviation $\sigma_{EF^r} = (\sigma_E^2 + \sigma_{F^r}^2)^{0.5}$. We estimate F_s^r by altering the bootstrap procedure so that the probability of selecting easy and hard cases is determined by r.

Fig. 5 shows the effect of reducing the number of easy test cases by half $(r = 1 : 2)$. The points in the left and right graphs represent the mean of our bootstrap computations. The confidence ranges are as computed before. We see that all the points fall within the confidence intervals, indicating that the effect of reducing the number of easy cases is less than that due initially to chance. Table 2 confirms this impression, showing σ_E, σ_{F^r} and σ_{EF^r} for several values of r. In this case $(r = 1 : 2)$ we see that $\sigma_{F^{1:2}} = 0.08$ is less than $\sigma_E = 0.11$. Fig. 6 and Table 2 show a similar effect when the number of hard cases is halved $(r = 2 : 1)$. Figs. 7 and 8 show that when easy and hard cases are eliminated altogether, the effect is greater than chance and concentrated at low and high scores respectively.

7. Discussion

Our simulation and model indicate that IOI 2005 rankings involve a considerable degree of chance due to test case scoring. This chance error is, of course, in addition to that

Table 2

Effect of easy-hard ratio r on error standard deviations

r		σ_E	σ_{F^r}	σ_{EF^r}
0:1	(0%)	0.11	0.29	0.31
1:4	(20%)	0.11	0.14	0.18
1:2	(33%)	0.11	0.08	0.14
3:4	(42%)	0.11	0.03	0.11
1:1	(50%)	0.11	0	0.11
4:3	(58%)	0.11	0.03	0.11
2:1	(67%)	0.11	0.08	0.14
4:1	(80%)	0.11	0.15	0.19
1:0	(100%)	0.11	0.29	0.31

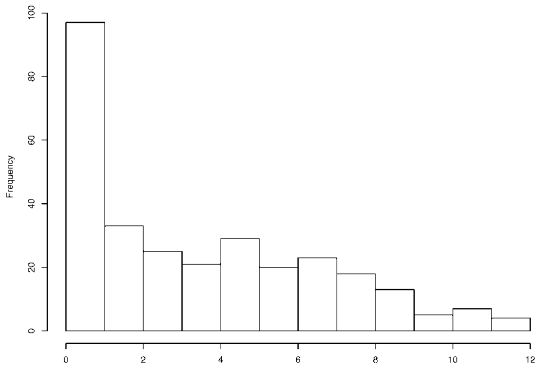

Fig. 9. All-or-nothing score distribution.

which may be due to other sources of chance or invalidity. The 95% confidence interval for most recipients contains a medal boundary. Ranking is not particularly sensitive to the relative number of easy and hard cases. Most students achieve either zero or full marks on each identified group of test cases – a substantial number of students fail to solve even the easiest ones. We believe that effort should be spent to identify more classes of easy and hard test cases, rather than to add (or subtract) more cases to (or from) a particular class.

Our statistical results, and also concern as to the validity of awarding part marks for incorrect programs, led us to do one concluding experiment. We scored each group of tests categorically as correct or incorrect – a program solving all the cases was given a score of 1 for the set; a program failing to solve any was given a score of 0. Fig. 9 shows the resulting total score distribution. Although there are only thirteen possible values, the distribution appears to have a smooth tail, as appropriate for picking medalists. However, this scheme awards ninety-five students – approximately one-third – a score of zero. We suggest that awarding of part marks simply adds random noise rather than addressing the inherent problem that the tasks and test cases are unable to distinguish among the bottom third of achievement. We have shown that the number of test cases within each category is not critical. We argue that the overall precision of the contest, and the satisfaction of the contestants would be improved, if one-third of the test cases were removed from each of the existing categories, and used to form a new category – truly easy – for which some eighty percent of the participants would be expected to succeed.

References

The International Olympiad for Informatics (2006).
`http://www.ioinformatics.org/`

Baker, F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

Cormack, G., G. Kemkes, I. Munro and T. Vasiga (2006). Structure and purpose of programming competition. In *Perspectives on Computer Science Competitions for (High School) Students*. Dagstuhl. *Informatics in Education*, **5**(1), 15–36.

Efron, B., and R.J. Tsibirani (1994). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Fisher, M., and A. Cox (2006). Gender and programming contests: Mitigating exclusionary practices. In *Perspectives on Computer Science Competitions for (High School) Students*. Dagstuhl. *Informatics in Education*, **5**(1), 47–62.

Fisher, R.A. (1925). Theory of statistical estimation. In *Proceedings of the Cambridge Philosophical Society*, vol. 22, pp. 700–725.

Forisek, M. (2006). On suitability of programming competition tasks for automated testing. In *Perspectives on Computer Science Competitions for (High School) Students*. Dagstuhl. *Informatics in Education*, **5**(1), 63–76.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science*.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, **3**, 1–18.

Verhoeff, T. (2006). The ioi is (not) a science olympiad. In *Perspectives on Computer Science Competitions for (High School) Students*. Dagstuhl. *Informatics in Education*, **5**(1), 147–158.

Yakovenko, B. (2006). 50% rule should be changed. In *Perspectives on Computer Science Competitions for (High School) Students*. Dagstuhl.

G.V. Cormack is a professor in the David R. Cheriton School of Computer Science, University of Waterloo. Cormack has coached Waterloo's International Collegiate Programming Contest team, qualifying ten consecutive years for the ICPC World Championship, placing eight times in the top five, and winning once. He is a member of the Canadian Computing Competition problem selection committee. He is currently an elected member of the IOI Scientific Committee. Cormack's research interests include information storage and retrieval, and programming language design and implementation.

2005 m. tarptautinės informatikos olimpiados vertinimo atsitiktiniai veiksniai

Gordon CORMACK

Straipsnyje nagrinėjamas Tarptautinės informatikos olimpiados dalyvių spręstų uždavinių vertinimo tikslumas. Užuot nagrinėjus pačius uždavinius ar jų parinkimo veiksnius, pagrindinis dėmesys sutelkiamas į klausimą, kaip dalyvių gautieji taškai iš esmės atitinka sprendimų teisingumą. Į konkrečias dalyviams pateiktas užduotis žvelgiama kaip į panašių užduočių begalines aibes tam tikra išraiška. Standartinių užduočių įvairovė įvertinama pagal pakopinį metodą, skaičiavimams naudojami patikimumo intervalai, kurių hipotetinis išdėstymo teisingumas siekia 95%. Tiriama lengvų (vadinamoji penkiasdešimties procentų taisyklė) ir sudėtingų užduočių santykinė sąveika, atsižvelgiama į bendrą išdėstymą. Vadovaujantis empiriniais duomenimis (2005 m. Tarptautines informatikos olimpiados rezultatais), daroma prielaida, jog tiek lengvos, tiek sudėtingos užduotys deramai pasitarnauja vertinant dalyvių sprendimus, tačiau šių užduočių proporcija tam nėra svarbi.