Better rank assignment in multiple-choice entrance exams

Anindya Chatterjee*

Department of Mechanical Engineering, Indian Institute of Technology Kanpur, Kanpur 208 016, India

The Indian educational system administers many multiple-choice exams wherein a small percentage of candidates must receive unique ranks. But discrete marking causes ties, usually resolved by essentially arbitrary criteria. I suggest a better alternative via a probability model for the exam, participants' talent levels, and question difficulties. Monte Carlo simulations suggest that ranking accuracy improves if marks on each question are scaled by the population standard deviation for that question. Maximum marks obtained in the test should approach 100% without any ties. Finally, consideration of a guess-oriented test strategy suggests a target value for the lowest successful score.

Keywords: Multiple-choice exams, probability model, ranking, ties.

THE Indian educational system periodically administers several large, competitive exams. These include entrance exams for professional degree programmes in various popular institutions, including the Indian Institutes of Technology (IITs). Unlike some multiple-choice exams administered internationally, such as the GRE, in India we have some additional constraints. The exam must usually be administered to all candidates at the same time; questions from previous exams cannot in principle be repeated (and so questions cannot be calibrated in advance); a unique rank must be assigned to each candidate; and, in many cases, only a small percentage of the exam-taking population is selected. At the same time, unlike the GRE, these competitive exams do not attempt to make absolute evaluations with inter-temporal validity, i.e. while the same GRE score from two different years is supposed to indicate equal talent or ability, the same rank in two different years need not do the same. These exams also are freed from the need to accurately evaluate the entire exam-taking population as in the case of the GRE: everyone who takes the GRE presumably gets a score, but those not ranked in these entrance exams are not informed regarding where they stand relative to the greater population. Finally, note that ranking is unaffected by the addition of an arbitrary constant to every candidate's score, while the GRE has no such leeway.

*e-mail: anindya100@gmail.com

To appreciate the basic dilemma in rankings, consider a hypothetical exam taken by 300,000 aspirants. The exam has 200 multiple-choice questions with four choices each. Each correct answer fetches one mark, and each omitted answer fetches zero. There is negative marking, with each incorrect answer fetching –0.25 marks. Negative scores on the exam cannot be accepted (attempting nothing fetches zero), and the exam score has a resolution of at best 0.25. Six thousand individual ranks must be given to the top 6000 candidates (top 2%). Yet, clearly even 800 unequivocal ranks cannot be awarded. Tied totals seem unavoidable.

A survey of relevant websites suggests that various ad hoc criteria are presently used to resolve ties. For example, all candidates who score (say) 130.00 would be judged superior to those scoring 129.75 and inferior to those scoring 130.25. Among those scoring 130.00, the tie might be broken by forming subcategories based on, say, scores in first one subject deemed most important (e.g. physics), then another (e.g. mathematics), and so on; and then appeal might be made to high-school board marks; and in some rare cases final ties might be resolved by age. Such ad hoc resolution of ties seems unsatisfactory.

In this article I propose an alternative that seems better. In particular, I use a simple statistical model for the exam; and show that when the individual questions are weighted inversely with the population standard deviation of the scores obtained on those questions, then clashes are virtually eliminated and the ranking accuracy can improve as well. Additionally, simulations using the model provide insights into how existing exams might be improved for better accuracy.

As a non-statistician stakeholder addressing nonspecialists, I will make minimal, clear and simple statistical assumptions, seek no analytical results, and mostly present just Monte Carlo simulations that demonstrate my point. Rigorous justification, alternative assumptions, optimality, etc. of the proposals in this article can hopefully be studied later by more able researchers.

Literature review

There is much prior work on testing and multiple-choice exams. A representative sample of the literature is given here.

The aggressive critique of multiple-choice exams by Hoffmann[1] seems difficult to surpass and is worth a look if only for its remarkable self-belief. However, in my opinion, we have no practical alternative and must do the best we can within the multiple-choice framework.

Many early works theoretically studied the relation between distributions of item difficulties and the reliability of exams[2–5]. Other works have since gone on to study estimation methods and confidence intervals[6], the role of chance on test validity[7], correlations between tests and their relation to test reliability[8], indices of cheating[9], the equivalence of constructed-response and multiple-choice tests[10], empirical tests of guessing behaviour predicted by theory[11,12], models based on latent traits[13], gender-based differences in test-taking[14], evaluations of tests that are combinations of multiple choice and free response[15], fitting polytomous response-theory models to multiple-choice tests[16], the relation between number of item options and test reliability[17], and general issues in the design of good multiple-choice questions[18].

However, in these papers, the primary concern is the assessment of the quality or competence of a large population. They are not concerned with assigning accurate ranks to a small percentage of the population, and do not share our indifference to inter-temporal validity. Moreover, in view of my intended audience and in line with my own tastes and skills, my presentation below is informal and more broadly accessible.

Statistical model

All statistical questions are answered in the context of a probability model. We need a probability model here as well.

Need for a model and simulation

Let us say we wish to rank the candidates by some underlying quality, like initial talent. We cannot directly measure talent. All we have is the exam performance. How do we determine whether the ranks we have assigned, based on marks, accurately reflect the underlying quality (namely talent)?

Can we evaluate our ranking scheme by measuring something else later on? I think not. For example, if we follow the careers of individual selected students within (say) the IIT system, then when and what should we measure? Their graduating GPA four years later? But that would eliminate the rejected candidates from evaluation; and the GPA would in any case reflect the combined effects of initial talent, subsequent effort, sustained motivation, academic and personal luck, family circumstances, and the presumably high benefits of studying in the IIT system in the first place. The problem cannot be eliminated by monitoring instead their ability to secure excellent jobs, or their annual salaries 15 years later.

So I suggest that, for evaluating the ranking scheme, we must use the exam data itself, with a suitable probability model. In this article I propose such a model. It shows something useful, and has the advantage of simplicity. Perhaps future work can develop better models.

Model details

Numbers: There are M candidates, N questions on the exam and N_r candidates must be uniquely ranked. Here, we will uniformly consider $N = 200$, $M = 300{,}000$ and $N_r = 6000$. In other simulations not presented here, other numbers gave similar results.

Quality of candidates: Candidate k has a quality $0 < q_k < 1$. Our aim is to rank all candidates in the order of decreasing quality. Note that this quality is an abstract variable, and monotonic transformations of it will not affect the ranking. So presumably some liberty may be taken in assuming its underlying distribution. We assume that q-values for the exam-taking population obey the beta distribution, which is frequently used for random variables on the unit interval, and whose probability density function is given by

$$p(q) = Cq^{\alpha-1}(1-q)^{\beta-1},$$

where the parameters α and β are strictly positive, and the constant C is chosen to ensure

$$\int_0^1 p(q)\mathrm{d}q = 1.$$

Now α strongly influences the distribution only for small q, a regime we are not studying here. The key parameter is β, as it strongly influences the distribution of high-quality candidates. Accordingly, I assume simply that $\alpha = 1$, and take

$$p(q) = \beta(1-q)^{\beta-1},$$

with β being a parameter in the simulation.

In each simulation, a set of M values of q is generated using a random-number generator. The effectiveness of the ranking method can be studied through simulations by comparing the ranks suggested by the exam with the underlying ranks based on q. In an actual exam, q is not known and the rank is all we have.

Difficulty of questions: Question i in the exam has a difficulty level $d_i > 0$. Provided candidate k genuinely attempts question i, her probability of getting it correct is $q_k^{d_i}$. Thus, holding d_i constant, higher q_k increases the probability of success; and holding $0 < q_k < 1$ constant, higher d_i lowers the probability of success.

We assume d is lognormally distributed. That is, $\ln d$ is normally distributed with mean μ and variance σ^2 (two more parameters).

Candidate's strategy: Since there is negative marking, the candidate must make a conscious choice on whether or not to attempt any given question. For simplicity, we assume that the candidate is sufficiently self-aware as to know $q_k^{d_i}$ (see assumption above). If this probability is less than 0.25, she does not attempt the question. Otherwise, she attempts it. On attempting it, she gets a score of 1 with probability $q_k^{d_i}$, and a score of –0.25 otherwise.

Other models in future work might consider introducing a further random component in the decision to attempt or avoid. A brief discussion of guessing after eliminating one option out of four is given in Appendix 1.

Scoreboard: Now the above assumptions can be used in a simple simulation. Given β, the qs are generated. Given μ and σ, a set of random ds is generated. Finally, the exam is simulated candidate by candidate, question by question, to fill in an $N \times M$ scoreboard matrix S, wherein each entry is either 1, 0 or –0.25.

For example, for candidate k and question i, if $q_k^{d_i} < 0.25$, then we set $S_{ik} = 0$. Otherwise, we generate a random number ξ uniformly distributed on (0, 1) using Matlab's 'rand'. If $\xi < q_k^{d_i}$, we set $S_{ik} = 1$; otherwise we set $S_{ik} = -0.25$.

Rankings: plain and scaled

Each time we generate a scoreboard matrix S as described above, we have simulated one administration of the exam. Adding up the elements of S column-wise gives the total scores (marks) of the M candidates. In place of ad hoc methods used to resolve clashes, we perturb all the total marks by tiny random amounts (on the order of 10^{-9}). Then we rank the candidates according to these total marks: call it the plain ranking.

An alternative to using this method of plain ranking is to normalize the scores question-wise. In this context, Dawes[19] considers the problem of evaluating and ranking candidates in various spheres of activity: interviews to select and rank graduate students, stock selection by investors, and simple formulas to estimate the health of marriages, to name three. Dawes points out, with much data to back him up, that linear regression models used for ranking such options are actually rather robust; and if we get the signs of the coefficients right and assign them otherwise random values, we do a fair job of predictions on average; and if we normalize the variables and then assign them equal weights, then we do better than that average, better than human judges as evidenced by many studies, and typically only slightly worse than the optimal linear models. I fear my brief paragraph cannot do justice to Dawes's paper, and urge the reader to read the original[19].

Accordingly, in our scoreboard matrix S, we normalize question-wise: we divide each row of S by the standard deviation of the elements of that row. Then, as before, we add up the columns and rank the candidates: call this the scaled ranking.

This scaling has the obvious advantage of virtually eliminating clashes; and Dawes's paper suggests it will improve ranking accuracy. However, the only way to check if ranking accuracy improves is to study, using a reasonable model, a suitable numerical measure of ranking errors; and therein lies the contribution of this article.

Measure of ranking errors

In a real exam, we do not know the true ranks of the candidates. However, in our simulation we do. Accordingly, for each ranking method, when we identify the top N_r candidates with the highest ranks, we also know their true ranks. For defining our error measure, we assume that mistakenly assigning rank 10 to what should be 30 is about as bad as assigning rank 100 to what should be 300, or 300 to 100. Accordingly, we define a logarithmic error measure as follows:

$$E = \frac{1}{N_r} \sum_{k=1}^{N_r} \left| \ln \left(\frac{k}{\text{True rank of person assigned rank } k} \right) \right|.$$

If this error E turns out to be 0.3, say, then it would mean that a typical assigned rank is off by a factor of $e^{0.3} = 1.35$. Assigned rank 100 might typically reflect a true rank between 74 and 135 (high confidence error bounds would include a larger range). Obviously, a ranking method which gives a smaller value of E is superior.

The Matlab code used for these simulations is given in Appendix 2. One minor issue is that dividing by the standard deviation is problematic if it is exactly zero. However, no matter how we scale those problems' marks, nobody gets any advantage in the rankings anyway. Here, I simply added 10^{-9} to the standard deviation in all cases before division.

Simulations

While N, M and N_r were fixed at 200, 300,000 and 6000 respectively, it is not a priori obvious what values might be suitable for μ, σ and β. I think $\beta > 1$ is appropriate because it lowers the density near $q = 1$, which is consistent with extreme talent being rare. Similarly, $\mu > 0$ seems appropriate because it makes the median difficulty greater than unity. Note that for any given set of parameter values, the outcome of each simulation is random. In order to understand what the model predicts, many

simulations have to be conducted for many different parameter values.

Accordingly, I first conducted 10,000 simulations. In these simulations, μ was chosen randomly every time, uniformly distributed between 0 and 6; σ was uniformly distributed between 0 and 4, and β was uniformly distributed between 1 and 3. For each simulation, the plain score E_p, the scaled score E_s, the maximum and minimum plain scores (among the ranked, or top 6000 candidates) were recorded. Of these 10,000 simulations, those where the minimum score was zero or less were discarded; those where the maximum score was 200 were also discarded. There remained 7011 simulation results, which are presented below. Subsequently, results are presented from other, more focused, simulations.

Results

Initial simulations

From the above-mentioned initial 7011 random simulations, a histogram of the ratio E_p/E_s is given in Figure 1. The scaled ranking method is most often better than the plain ranking method. The minimum values obtained are small for both E_p (0.1094) and E_s (0.1082). A histogram of E_s is shown in figure 2. It is seen that a more typical value is 0.33.

Figure 3 shows a scatter plot of scaled rank E_s against marks range. The lowest value of E_s is seen as an outlier at bottom right. A higher marks range generally gives more accurate ranking.

Accordingly, consider the scatter plot of E_p/E_s against marks range in Figure 4. For a sufficiently large marks range, the scaled ranking has a high probability of being better than the plain ranking; even when it is worse, it is so by very little. This observation suggests a guideline for exam design: the marks range among the top N_r candidates should be large, preferably greater than 60% (see also Appendix 1).

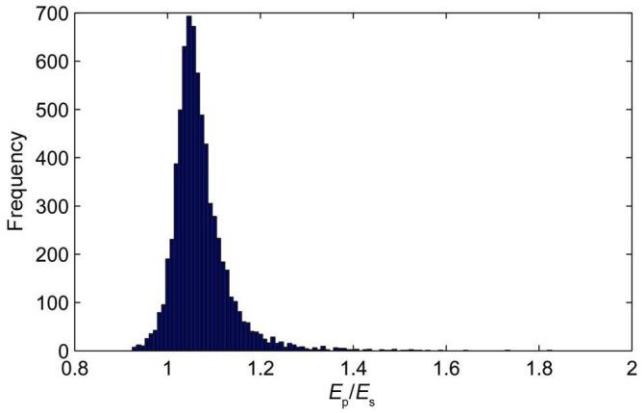

Figure 1. Histogram of E_p/E_s from 7011 simulations.

A well-known result of testing theory is that low variability in the difficulty of questions is desirable. In this light, see the scatter plot of E_s against the standard deviation σ of the logarithmic difficulty, as shown in Figure 5. The figure indeed suggests that lower variability improves ranking accuracy. Note, however, that several high values of E_s (poor rankings) are seen in Figure 5 for σ between 1 and 2. These points resemble others from the

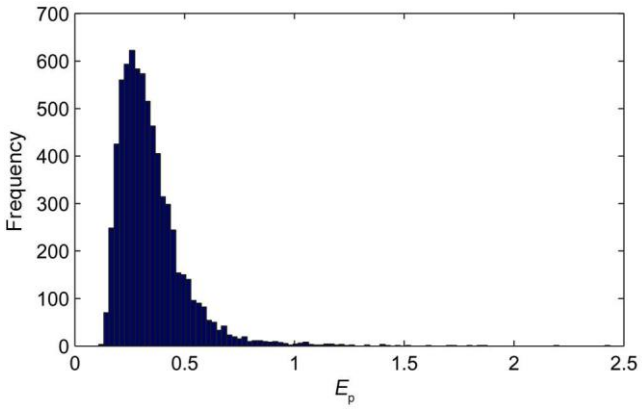

Figure 2. Histogram of E_p from 7011 simulations.

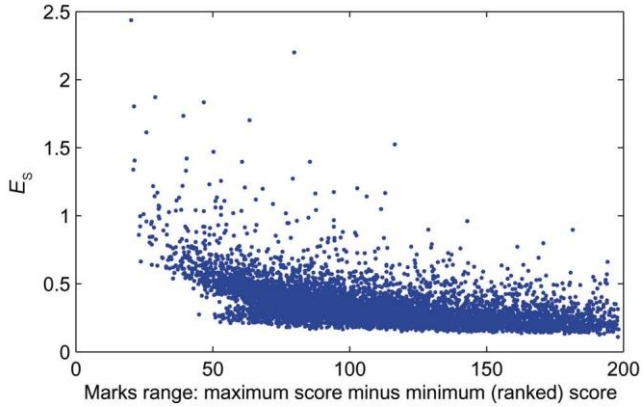

Figure 3. E_s plotted against marks range for 7011 simulations.

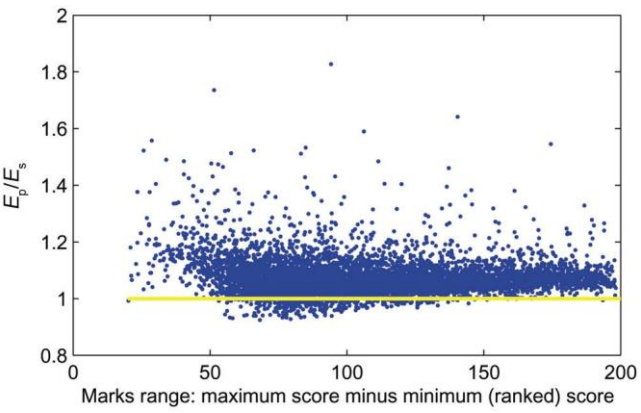

Figure 4. E_p/E_s plotted against marks range for 7011 simulations.

top left portion of the figure, which got eliminated when results with extreme scores (zero and full marks) were dropped. Thus, assuming that the exam is neither too easy nor too hard, low variability in difficulty is desirable.

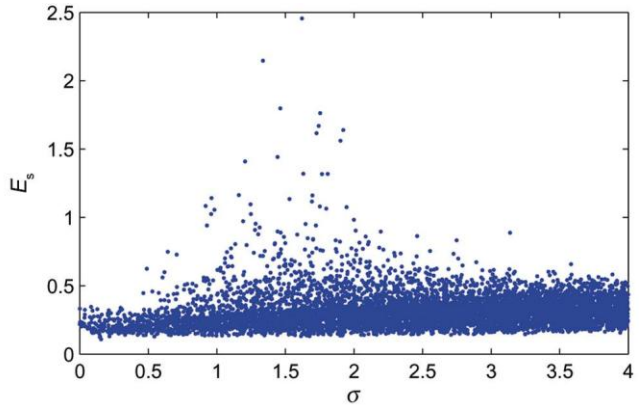

Figure 5. E_s plotted against σ for 7011 simulations, where σ is the standard deviation of the logarithmic difficulty of questions. Low variability in difficulty improves ranking accuracy, provided the maximum and minimum marks are below 200 and above zero respectively.

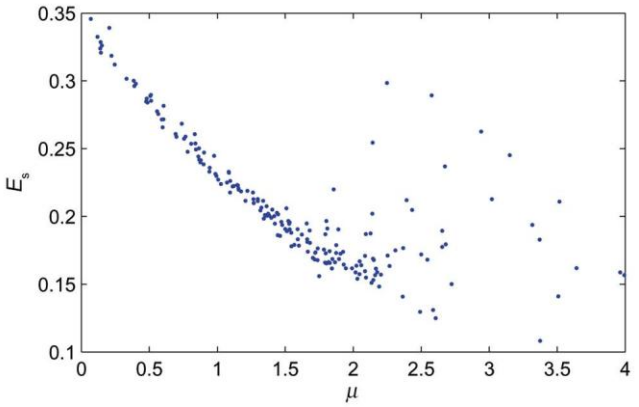

Figure 6. E_s plotted against μ for 190 simulations (with $\sigma < 0.45$), where μ is the mean logarithmic difficulty of questions. See also Figure 7.

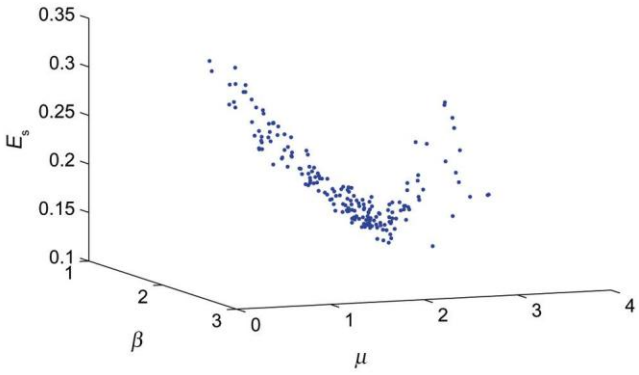

Figure 7. E_s plotted against μ for 190 simulations (with $\sigma < 0.45$). The points lie approximately on a surface with a 'valley', seen roughly sideways in this view.

Accordingly, we now choose those points among our 7011 simulations where $\sigma < 0.45$. There happen to be 190 such points. A plot of E_s against μ is given in Figure 6. A more interesting view of the same data is given in Figure 7, which shows that E_s for this subset of points lies approximately on a surface with a well-defined 'valley'. To study these optimal parameter values, we consider further simulations.

Further simulations

Further simulation results are given in Figure 8. Here, $\beta = 1.6$; $N = 200$, $M = 300{,}000$, $N_r = 6000$, $\sigma = 0.2$ in all cases; and μ is as shown. In the simulation results shown in Figure 9, all parameters remain the same, except that σ is 0.5. Finally, results for $\sigma = 0.8$ and all other parameters unchanged are given in Figure 10.

Figures 8 through 10 together show that (a) all else held constant, there is an optimal value of mean logarithmic difficulty μ, (b) increased logarithmic variability σ slightly lowers precision in rankings at the optimal point, and (c) the superiority of the scaled ranking method over the plain ranking is small, but robust.

These figures also suggest that it is desirable for the maximum score obtained to approach 100% without causing ties there. For larger μ (more difficult exam on average), higher variability seems to be better, possibly because it allows some easier questions to remain in the exam. Interestingly, while low variation in difficulty gives the most accurate rankings at the optimum, it appears that somewhat higher variation makes the optimum more robust with only a modest decrease in optimal ranking accuracy. In this sense, perhaps it may be practically desirable to aim for a somewhat difficult exam on average, with also a somewhat larger range in the difficulty levels of questions.

One thing that is not clear from the above simulations is whether or not there is an optimal lowest score among the ranked candidates (see Appendix 1).

Concluding remarks

In this article I have presented, through extensive Monte Carlo simulations of an assumed probability model for a multiple-choice exam, results that indicate the following.

Such exams have an optimal level of difficulty in questions, whereby the maximum score, obtained by the top-ranked candidate, approaches full marks. Variation in question difficulty should not be too large, but if median question difficulty errs on the higher side, then somewhat higher variability can be used to make the ranking more robust. Generally, a larger spread of marks leads to better accuracy. Finally, normalizing each question's marks with the population standard deviation of marks obtained on that question (the scaled ranking method) improves

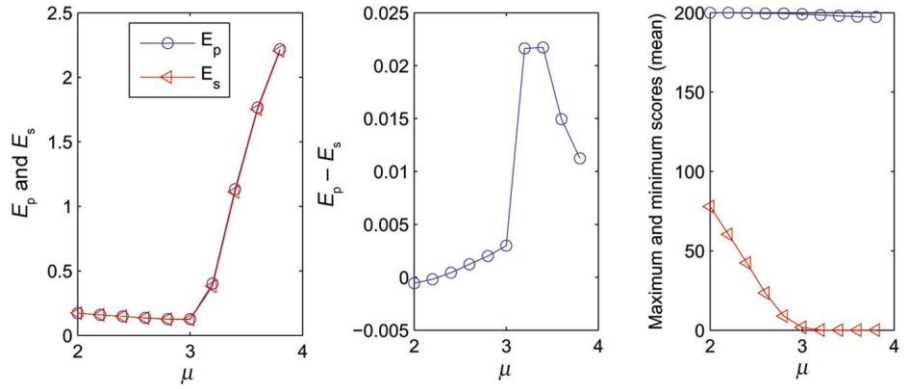

Figure 8. (Left) E_p and E_s versus μ (averaged over 100 simulations each). There is an optimum at $\mu \approx 3$. (Middle) Average value of $E_p - E_s$ shows that the scaled ranking is slightly superior (see also Figures 9 and 10). (Right) Mean values of maximum and minimum scores obtained (among 6000 ranked candidates).

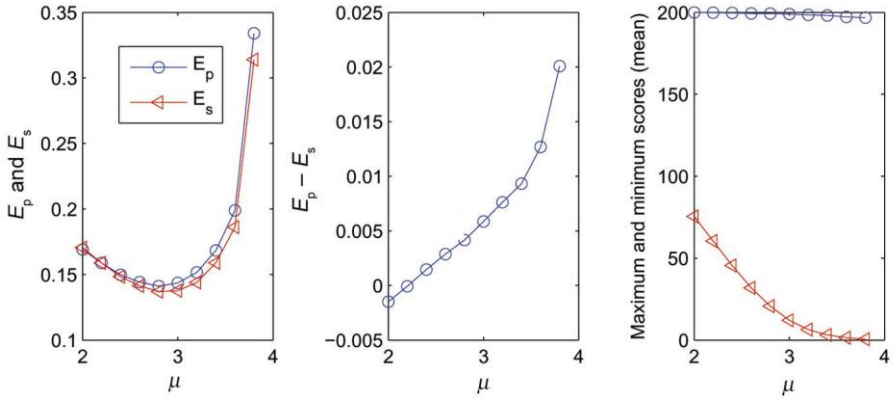

Figure 9. Same as Figure 8, except $\sigma = 0.5$. See also Figure 10.

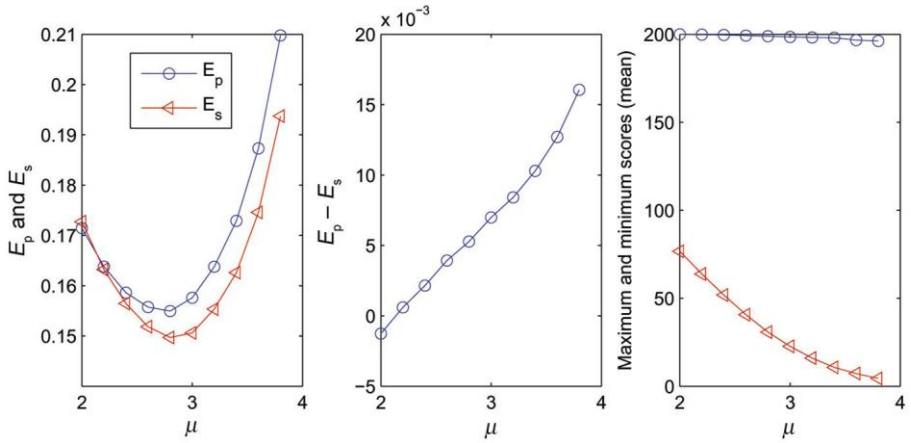

Figure 10. Same as Figures 8 and 9, except $\sigma = 0.8$. The superiority of the scaled ranking method over the plain ranking method is now clearly seen.

ranking accuracy. The scaled ranking method, which may offer increased accuracy of a few percentage points on even a well-set exam, has the additional appeal of being transparent, and in my opinion more rationally motivated than the presently used methods. A separate criterion for a suitable minimum score on the exam is given in Appendix 1.

An issue not addressed in this article is that even the best candidates often skip a few questions in such exams; yet these questions might be attempted by others. Such

behaviour is not accommodated in the model studied here. A simple way to do so is to say that every candidate has a certain fixed and uniform probability p (say, 0.4 or 0.1) of not even considering a given question, and skipping it without considering it at all; if she considers it, then she decides whether or not to attempt it, according to the model studied above. Several simulations with such an added random-skipping step in the model showed an overall reduction in ranking accuracy, but the small superiority of the scaled ranking over the plain ranking method persisted. Detailed study of question-skipping is left to future work.

I close with an interesting question raised by a colleague. If a question is relatively quite easy, so that most candidates get it right, then the standard deviation of that question's score will be low. Conversely, a question that is relatively quite difficult may be omitted or marked incorrectly by many, again leading to a low standard deviation. A typical question, in contrast, may have a somewhat higher standard deviation. In our scaled ranking scheme, it may seem odd that we assign large weights to both the easy and the difficult question, while giving a relatively smaller weight to the average question. But the point, as mentioned at the start of the article, is that we are interested in ranking and not absolute evaluation. For the difficult question, the high weightage is intuitively obvious. For the easy question, which many get right, the high weightage acts indirectly as a penalty on the ones who did not get it right. Presumably, a well set exam will have no questions where the success rate is either extremely large or extremely small.

Appendix 1. Lowest score among qualifying candidates

The simulations have not identified a clearly advantageous lowest score among the ranked candidates. For example, the lowest score for an optimal exam need not be zero (see Figure 10 with μ just under 3).

To this end, note that my simplistic model does not account for guessing-based strategies. Such strategies may be important in the training imparted by coaching classes that specialize in such exams. Consider a highly coached candidate who cannot actually answer any of the questions asked, but has developed the undesirable talent of being able to eliminate one choice from each question. It is then advantageous for this candidate to randomly guess for every question from among his three remaining choices.

Working with $N = 200$ questions, $M = 300{,}000$ candidates, and $N_r = 6000$ candidates (top 2%), we note that the probability of such a guesser getting n questions correct is

$$P_n = \binom{N}{n}\left(\frac{1}{3}\right)^n\left(\frac{2}{3}\right)^{N-n}.$$

Summing P_n for $n = 81, 82, \ldots, 200$, the probability of getting more than 80 answers correct out of 200 is about 2%. Getting 80 answers correct (upon attempting all) gives a score of 50 out of 200, or 25%. Thus, if coaching can aid in such elimination of options, then the lowest score in the exam (for the candidate ranked 6000) needs to be higher than 25%. Otherwise, the poorest of the legitimate performers risks being ousted by guessers.

Incidentally, if we change the number of questions from 200 to 300, then the probability of getting more than 117 answers correct by such guessing is under 2%, corresponding to a score of about 24%. Increasing the number of questions to 400 makes it 23%. In other words, increasing the number of questions does not help much. The problem is made slightly worse when the percentage of candidates to be selected is somewhat higher (say, 3% instead of 2%).

Thus, paper-setters must take pains to set questions where one option cannot be easily eliminated, so that a larger range of marks is available for ranking candidates accurately. Failing that, I suggest we should aim for the lowest score in the exam (among the selected candidates) being greater than 25%: perhaps 30%? This in turn suggests a marks range of about 70%, not inconsistent with the discussion following Figure 4.

Appendix 2. Matlab code

```
function C = jeesim(N, M, Nr, muu, sigma,
beta)

d = exp(randn(N, 1)*sigma+muu);
scores = zeros(N, M);
q=fliplr(sort(1-rand(1,M).^(1/beta)));

for  n = 1 : N
     dd = d(n); p_right = q.^dd;
     r = rand(size(p_right));
     scores(n,:) = ((r < p_right)
      *1.25 -0.25).* (p_right > 0.25);
end
plain_scores=sum(scores)'+1e-9*randn(M,1);

for  n = 1 : N
     s = std(scores(n,:));
     scores(n,:) = scores(n,:)/(1e-9 + s);
end
scaled_scores = sum(scores)' ...
   + 1e-9*randn(M, 1);

[sc, scaledrank] = sort(-scaled_scores);
[sp, plainrank] = sort(-plain_scores);
n = [1:Nr]';
C = [mean(abs(log(plainrank(n)./n))), ...
    mean(abs(log(scaledrank(n)./n))), ...
    plain_scores(plainrank(1)), ...
    plain_scores(plainrank(Nr))];
```

1. Hoffmann, B., Multiple choice tests. *Phys. Educ.*, 1967, **2**, 247–251.

2. Gulliksen, H., The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, 1945, **10**, 79–91.
3. Brodgen, H. E., Variation in test validity with variance in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, **11**, 197–214.
4. Lord, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17**, 181–194.
5. Cronbach, L. J. and Warrington, W. G., Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika*, 1952, **17**, 127–147.
6. Lord, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18**, 57–76.
7. Plumlee, L. B., The predicted and observed effect of chance success on multiple-choice test validity. *Psychometrika*, 1954, **19**, 65–70.
8. Horst, P., The maximum expected correlation between two multiple-choice tests. *Psychometrika*, 1954, **19**, 291–296.
9. Frary, R. B., Tideman, T. N. and Watts, T. M., Indices of cheating on multiple-choice tests. *J. Educ. Behav. Stat.*, 1977, **2**, 235–256.
10. Traub, R. E. and Fisher, C. W., On the equivalence of constructed-response and multiple-choice tests. *Appl. Psychol. Meas.*, 1977, **1**, 355–369.
11. Cross, L. H. and Frary, R. B., An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *J. Educ. Meas.*, 1977, **14**, 313–321.
12. Bliss, L. B., A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *J. Educ. Meas.*, 1980, **17**, 147–153.
13. Nicewander, W. A., A latent-trait based reliability estimate and upper bound. *Psychometrika*, 1990, **55**, 65–74.
14. Ben-Shakhar, G. and Sinai, Y., Gender differences in multiple-choice tests: the role of differential guessing tendencies. *J. Educ. Meas.*, 1991, **28**, 23–35.
15. Thissen, D., Wainer, H. and Wang, X. B., Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *J. Educ. Meas.*, 1994, **31**, 113–123.
16. Drasgow, F., Levine, M. V., Tsien, S., Williams, B. and Mead, A. D., Fitting polytomous item response theory models to multiple-choice tests. *Appl. Psychol. Meas.*, 1995, **19**, 143–166.
17. MacCann, R. G., Reliability as a function of the number of item options derived from the 'knowledge or random guessing' model. *Psychometrika*, 2004, **69**, 147–157.
18. Nicol, D., E-assessment by design: using multiple-choice tests to good effect. *J. Further Higher Educ.*, 2007, **31**, 53–64.
19. Dawes, R. M., The robust beauty of improper linear models in decision making. *Am. Psychol.*, 1979, **34**, 571–582.

ACKNOWLEDGEMENTS. Atanu Mohanty independently verified some results and helped streamline my Matlab code. Devlina Chatterjee and Vikranth Racherla provided useful comments on the manuscript.

Received 26 April 2013; accepted 15 May 2013