On multiple choice tests and negative marking

Rajeeva L. Karandikar

We critically examine the impact of marking schemes in multiple choice tests on the outcomes. We postulate reasonable models for the distribution of marks as well as of the guessing behaviour of the candidates when they do not know the correct answer. Through simulation, we show that the impact is significant. We suggest an alternative for improving the outcome.

Keywords: Gatecrasher, multiple choice test, negative marks, random guess.

Multiple choice tests have been used for screening candidates for a specific objective. Increasingly they are being used as a single test for final selection for admission to a course, award of fellowship, or for a job.

By a multiple choice test we mean a traditional test where each question has exactly one correct answer (among several choices, typically four or five) and to get credit the candidate needs to tick the correct answer (under the assumption that there is exactly one correct answer).

In a multiple choice test, when an answer is incorrect we can be sure that the candidate does not know the answer and in case the answer is correct, we are not sure if the candidate actually knows the answer or the outcome is due to a random guess. That is why whenever we talk of multiple choice tests, the issue of negative marks for an incorrect answer always crops up. Are there negative marks? If so what is the negative marking scheme? The discussion on negative marks often throws up differing views among experts. While some feel that there should be no negative marks as one should not take away credit that has been earned, some others argue that there should be nominal negative marks. Yet others argue that it does not matter: it is the same rule for everyone.

Even among those who feel that there should be negative marks, there is confusion as to the quantum of negative marks for an incorrect answer. Some argue that if every question has n alternatives, the correct negative mark for an incorrect answer should be $1/n$. The common interpretation of correct seems to be that a candidate choosing an answer randomly should not get any advantage on the average. In other words, if a candidate ticks all answers in a test randomly, the expected score of such a candidate should be 0. Simple calculation[1] shows that for this to happen the correct negative mark for an incorrect answer should be $1/(n-1)$. It is easy to show that if the negative score for an incorrect answer is $1/(n-1)$, the expected score of a candidate remains the same as the score based on his knowledge. The expected advantage from random guessing being zero does not guarantee that it has no impact on selection.

An important question that needs to be answered is: how many candidates who should not have been selected get selected because of random guessing. In other words, we need to examine how many candidates gatecrashed into the list of selected candidates. We will discuss this in the next section.

Another factor that has a big impact on the outcome is the difficulties that arise when there are incorrect or ambiguous questions. Often the solution of such a problem is to award marks to all candidates. This has an impact on the final selection. However, we have not factored this here. After all, this can be avoided if the administrators of the test are careful.

Is the impact of random guessing marginal?

Let us analyse the impact of random guessing on the ranks of the candidates and the subsequent selection of the candidates. Let us consider a situation where there are 200,000 candidates and the test is to select up to 1000 candidates (for admission to a course or selection for a job). It is common in India to have selections of such magnitudes, such as in the admission in engineering colleges or in the recruitment of large technology companies. The test consists of 200 questions. The candidate with serial number i knows answers to X_i questions. We will call X_i as the true score of the ith candidate, as it is the score based on his/her knowledge (X_i lies between 0 and 200).

The candidate may guess the answers to the questions for which he/she does not know the answers, getting credit for the ones he/she got right by chance, and possibly getting negative marks for the ones where he/she got the wrong answer. Let Z_i denote the observed score of the ith candidate.

Ideally we should have selected the top 1000 students based on their true scores, i.e. X_is; but true scores are not observable, only Z_is are observable and hence we would select the top 1000 students based on their observed scores.

The author is in the Chennai Mathematical Institute, Plot-H1, SIPCOT IT Park, Padur PO, Siruseri 603 103, India.
e-mail: rlk@cmi.ac.in

Let L denote the number of (lucky) candidates that have been selected, but would not be selected if we had been able to observe $\{X_i: 1 \le i \le 200{,}000\}$. In other words, L is the number of candidates who ideally should not have been selected, but got selected because they were lucky and got ahead of others whose true score was higher than their own. Is L large or small? A large (as a percentage of 1000) value of L would suggest that random guessing has significant impact on the final selection.

In order to get an idea about the order of magnitude of L, we undertook a simulation exercise with reasonable assumptions about the distributions of underlying random variables (explained below). We considered different schemes of negative marking: $N = 0$, $N = 0.25$ and $N = 1/3$. In order to analyse the impact, we also need to model the behaviour of the candidates with regard to random guessing. We assume that $P\%$ candidates resort to random guessing on questions where they do not know the (correct) answer.

Table 1 gives the average number of candidates that have been selected on the basis of observed ranks, who would not have made it if we could observe the true ranks or true scores. The results are based on 10,000 simulations of the underlying random variables. All results have been rounded to the nearest integer for better comprehension.

We see that random guessing has a significant impact. If there is no penalty for an incorrect answer ($N = 0$) and over 20% candidates resort to guessing, on the average over 200 of the 1000 candidates selected are gatecrashers. When the negative score for incorrect answers is 0.25 and more than 40% candidates are resorting to guessing, we would be selecting over 100 candidates on the average out of 1000 who should not have been selected. Even when $N = 1/3$ (when a candidate cannot change his/her expected score by random guessing), on the average over 100 candidates are gatecrashing if over 80% candidates resort to guessing.

Only when the negative score is 1/3 (something that is in the control of the examination organizers) and when only 10% candidates guess (examination organizers cannot control this proportion), the average number of those who gatecrashed reduces to about 17 and if 20% guess, the number is around 33.

We have seen that a large percentage of candidates can gatecrash the selected list via random guessing (except perhaps when $N = 0.5$ and $P \le 30$). Let us explore as to what the gap is between the cut-off based on true scores and the true score of the weakest candidate making it to the list. Let G denote the difference between the true cut-off and the true score of the candidate selected with the smallest true score. If G is small, we may ignore the effect of random guessing, but a higher value of G should raise an alarm because it means candidates much weaker than other better available candidates have been selected.

For the simulation model described here, Table 2 gives the results of the average gap. Once again all results are rounded to the nearest integer.

The gap is largest – 16, when there are no negative marks and when only 10% candidates guess. Even under most scenarios the gap is 10 or more on the average.

Having seen that the average gap is large, let us examine as to how weak could the weakest candidates be among those selected. Let T denote the true rank of the weakest candidate who has been selected. Once again high value of T (relative to 1000) suggests weakness of the multiple choice test-based selection.

Table 3 shows the average of T for different combinations of N and P based on 10,000 simulations rounded to the nearest integer.

Except for $N = 1/3$ and $P = 10$, we see that when we select 1000 candidates, on the average candidates with rank above 3000 are making it to the list. For several scenarios, the average T is 3500 and more.

This means the test fails to select better candidates even though there are on the average 2000 or more candidates who are better than those that the test is selecting. And the number is much higher under several scenarios.

Model for simulation

Suppose there are 200,000 candidates and the test is to select up to 1000 candidates. The test consists of 200

Table 1. Average L: number of candidates who should not have been selected, but have been selected

Percentage candidates guessing	Negative marks for an incorrect answer		
	0	0.25	1/3
10	136	27	17
20	198	51	33
30	225	75	47
40	256	98	59
50	216	113	72
60	232	117	84
70	181	115	97
80	197	116	109
90	152	124	122

Table 2. Average G: gap between true cut-off and true score of weakest candidate selected

Percentage candidates guessing	Negative marks for an incorrect answer		
	0	0.25	1/3
10	16	9	8
20	15	10	9
30	15	11	9
40	14	11	9
50	13	11	10
60	13	11	10
70	12	11	10
80	12	10	10
90	11	10	10

questions. Recall our notation: the candidate with serial number i knows answers to X_i questions.

Out of the $200 - X_i$ questions, if the candidate decides to guess, he/she guesses the answer by randomly choosing one out of the four options in the remaining $W_i = 200 - X_i$ questions.

We model X_i, W_i as follows: Let X_i be the integer approximation to Y_i, where Y_i has normal distribution with mean 125 and standard deviation 20. We would like to remark that the distribution of true scores around the true cut-off is all that counts (for the quantities we are monitoring in this article) and thus if we select, say 0.5% as in this study, then the distribution of scores of the top 3–5% candidates alone matters and the rest does not. So Gaussian assumption is not critical to this study.

We assume that a candidate resorts to guessing with probability P: writing $H_i = 1$, if the ith candidate guesses and $H_i = 0$ otherwise, with distribution of H_i being Bernoulli with success probability P. We also assume that H_i and X_i are independent.

Let A_i denote the number of questions a candidate got correct out of W_i by random guessing. Then (conditional on W_i) A_i is binomial with $n = W_i$ and $p = 0.25$.

If N represents the negative marks for an incorrect answer, the (observed) score of the ith candidate Z_i is given by

$$Z_i = X_i + A_i + N * H_i * (W_i - A_i).$$

Table 3. Average T: true rank of the weakest candidate selected

Percentage candidates guessing	Negative marks for an incorrect answer		
	0	0.25	1/3
10	7392	3277	2750
20	6902	3647	3099
30	6363	3911	3283
40	6163	4076	3415
50	5348	4123	3516
60	5285	4050	3609
70	4607	3929	3684
80	4635	3843	3762
90	4129	3854	3837

Table 4. Fifth percentile of L: number of candidates who should not have been selected, but have been selected

Percentage candidates guessing	Negative marks for an incorrect answer		
	0	0.25	1/3
10	114	17	10
20	167	35	20
30	178	54	29
40	198	69	38
50	172	75	47
60	168	79	57
70	137	74	67
80	147	79	77
90	97	87	87

We can verify that for $N = 1/3$, conditional expectation of Z_i given X_i equals X_i, i.e.

$$E(Z_i \mid X_i) = X_i.$$

Also, the random vectors (X_i, W_i, H_i, A_i), $1 \le i \le 200{,}000$, are independent.

We simulate the random variables described above and compute the score Z_i for $1 \le i \le 200{,}000$.

We only observe the scores of candidates Z_i and we can only rank and select candidates based on their score Z_i. Let $\mathcal{F}$ be the set consisting of the serial number of students selected based on the scores Z_i. Since there can be ties (several candidates having the same score), we may have to choose a few more or a few less. To be precise, let us assume that we select not more than 1000 candidates, so that if the number of candidates with score greater than or equal to 177 is 983 while there are 32 candidates with score 176, we select only 983.

Since X_i denotes the number of questions the ith candidate knows, ideally we would have liked to rank the candidates on $\{X_i\}$ and select up to 1000 ranks. Let $\mathcal{G}$ be the set consisting of the serial number of students who should have been selected. Let S denote the cut-off (unobserved) based on true scores, i.e.

$$S = \min\{X_i : i \in \mathcal{G}\},$$

and let R_i denote the (true) rank of the ith candidate based on true scores.

Each of the quantities L, G, T described above measures the extent of mismatch between $\mathcal{F}$ and $\mathcal{G}$. These quantities can be described as follows:

$$L = \#(\mathcal{F} \cap \mathcal{G}^c)$$

$$G = S - \min_{i \in \mathcal{F}} X_i$$

$$T = \max\{R_i : i \in \mathcal{G}\}.$$

For the model described above, we have given average values of L, G, T in the previous section for various choices of N and P.

It is well known that average alone does not describe a distribution. For example, the average can be high because the random variable in question takes a large value with a small probability, while with overwhelming probability it takes small values. So we give below the 5th percentile of L, G, T in each of the scenarios below.

Table 4 shows that if $N = 1/3$ and $P = 90$ so that 90% candidates resort to guessing, then with 95% probability we will end up selecting 87 or more candidates (about 9%) who should not have been selected.

Table 5 shows that under several scenarios considered, the gap G is 8 or more with 95% probability.

Table 6 shows that we are selecting candidates with (true) rank over 2000 with 95% probability under most of the scenarios. Selecting a candidate with (true) rank of

2000 means that we are leaving out 1000 candidates who are better than the selected candidate. This shows the weakness of the selection scheme.

Even with $N = 1/3$ and $P = 60$ or $P = 70$, we would be selecting candidates with rank about 2500 or more with 95% probability.

A better alternative

One possibility is to increase the number of alternatives in each question from which the candidate can choose the correct answer. Increasing the number of alternatives to five from four changes the situation marginally. And anyone who has set questions in a multiple choice test knows that setting credible alternatives in a question is not easy. So going beyond five seems rather difficult.

One simple way to expand the possible set of solutions is to have questions that may have one or more correct answer(s) and to get credit the candidate should select all the correct answers and not select any incorrect answer. Then a question with four alternatives is turned into a question with 15 alternatives. Here is an example of such a question:

Which of the following are prime numbers?

(A) 63
(B) 37
(C) 91
(D) 83

Table 5. Fifth percentile of G: gap between true cut-off and true score of weakest candidate selected

Percentage candidates guessing	Negative marks for an incorrect answer		
	0	0.25	1/3
10	13	6	5
20	13	7	6
30	12	8	6
40	12	8	7
50	10	8	7
60	10	8	7
70	9	8	7
80	9	8	8
90	8	8	8

Table 6. Fifth percentile of T: true rank of the weakest candidate selected

Percentage candidates guessing	Negative marks for an incorrect answer		
	0	0.25	1/3
10	5286	2049	1748
20	4820	2542	2008
30	4315	2626	2253
40	4299	2890	2293
50	3784	2827	2363
60	3722	2646	2563
70	3317	2624	2584
80	3321	2622	2607
90	2916	2651	2644

Since 37 and 83 are prime numbers and 63 and 91 are not, (B) and (D) are correct options, whereas (A) and (C) are incorrect. Thus to get credit, a candidate must tick the two alternatives (B) and (D), and not tick (A) or (C).

Such tests have been discussed in the literature[1] and have been in use. In the proposed scheme, there is no partial credit or negative marks. So the candidate gets one mark if he/she ticks all the correct options and does not tick any incorrect answer; otherwise he/she gets zero marks for that question.

It is easy to see in the above example that there are 15 possible choices ($4C1 + 4C2 + 4C3 + 4C4 = 4 + 6 + 4 + 1 = 15$). With 15 alternatives, the impact of random guessing is negligible.

It is important to give the instruction correctly so as to avoid the problem that occurred in a major examination recently (http://education.gaeatimes.com/2010/05/26/iit-kharagpur-professor-underlines-mistake-3897/). Such questions have been tried in various tests where there is a subsequent round of interview and the scores in the test seem to have much better correlation with the performance than a traditional multiple choice test.

Since such questions are likely to be more substantive and would require analysis, more time should be given to candidates. That is, a reasonably good candidate should have enough time to answer all the questions within the time limit. Also, the pattern, instructions and some examples should be made available to the candidates before the test. It will also eliminate the possibility that a subsection of candidates might get unfair advantage by having prior knowledge about the type of test.

Conclusion

We have considered a situation where we are to select the top 1000 out of 200,000 students based on a multiple choice test with four alternatives to each question and with exactly one correct answer. If the negative score for an incorrect answer is $N = 1/3$, then the expected score of a candidate does not change by random guessing.

However, simulation reveals that the impact on the set of selected candidates is significant. With 95% probability, we would be selecting candidates whose true rank could be as high as 2500.

Of course, if we stick to traditional question–answer tests where the candidate has to write down the solution, then that would be the best. However, if for practical reasons one has to resort to a multiple choice test that can be evaluated via a computer, then a better alternative is to have questions that have one or more correct answers and then to postulate that to get credit a candidate must select all correct answers and not select any incorrect answer.

1. Bush, M., Alternative marking schemes for on-line multiple choice tests. In 7th Annual Conference on the Teaching of Computing, Belfast; available at http://www.caacentre.ac.uk/dldocs/bushmark.pdf

Received 10 July 2010; revised accepted 31 August 2010