Role of Chance in Objective Type Competitive Examinations

PRAMOD SRIVASTAVA, SHOBHIT MAHAJAN

The format of multiple choice examinations is used to determine rankings in competitive examinations in various higher education institutions. A statistical experiment based on actual multiple choice question papers reveals the preponderance of chance in determining rankings based on performance. With so much of "chance" determining performance, it is questionable if multiple choice-based competitive examinations provide a good measure of the intellectual abilities of students.

Pramod Srivastava (*pramodk.srivastava@gmail.com*) retired as professor of Physics from the University of Delhi and Shobhit Mahajan (*shobhit.mahajan@gmail.com*) is with the Department of Physics and Astrophysics, University of Delhi, Delhi.

One of the distinguishing characteristics of the higher education scenario in recent times has been the dominance of competitive examinations. Immediately after class XII, it is impossible to get into any professional course on the basis of the school leaving marks alone. From medicine to engineering to fashion design – every single course and indeed institution demands that one sits for an open examination to get admission. And this continues till after one graduates for postgraduate studies in medicine, engineering, management, computer applications, etc. Finally, a whole range of government/public sector jobs are also available only after clearing competitive examinations, the coveted civil services being only the most well known.

There is of course a stated rationale behind this proliferation of competitive examinations. The logic goes as follows: given that the standards of various examinations (Central Board of Secondary Education (CBSE), state boards, universities, etc) are so different, there is no way that marks obtained in these examinations can be compared and used to assess the quality of the candidate. Hence, the need to carry out another assessment. Of course, it is assumed by each agency that its examination in some sense does indeed assess the true quality of the student, something which is far from obvious.

Without going into the merits and demerits of this logic, in this piece, we want to try and see if the performance in the entrance examinations does indeed measure the capabilities of the student. More specifically, we want to see if there is an element of chance in the performance.

Multiple Choice Examinations

It is a common sentiment that the passing in a competition is, apart from many other factors, a matter of chance. Here we attempt to give this statement a fair measure of objective meaning with regard to one kind of method of testing, namely, multiple choice objective examinations. Multiple choice objective examinations have reduced dependence on long, essay type questions and eliminated the uncertainty associated with the multiplicity of human examiners. This format makes the scripts simple to evaluate and, in fact, capable of being evaluated by a machine.

The elimination of the human examiner has led many to believe that the result of a multiple choice test is sacrosanct. Of course, there is no doubt that the subjectivity which might creep in because of a human examiner is eliminated with this format. However, to jump from this to say that the result of a multiple choice examination is a true representation of the inherent capabilities of the examinee is where the problem arises.

For instance, if the questions were replaced by a different set of questions which are of equal level of difficulty, would the marks scored by the examinees remain more or less the same? If not, then there is a hidden sampling error in the choice of questions. This error can, in principle, be analysed by making the examinees undergo many equivalent tests over a short span of time. Clearly, this is not a very feasible experiment with any reasonable sample of students.

Instead, to try and understand the inherent sampling errors in multiple choice tests, we adopt another procedure which mimics the actual experiment closely. Our experiment is based on the actual data from the entrance examination to a professional course which was taken by over 57,000 students. The actual test consisted of two papers, with 100 questions in each paper and the total marks of each student in the two papers were used to rank the students. Of course, since the marks were only out of 200 and the number of students was many times that, there was a lot of degeneracy – a large number of people had the same rank and there were gaps in the ranks whenever more than one person gets the same mark and hence rank.

With this original data set, we prepared another fictitious data set which comprised

10 question papers with 50 questions each. These 50 questions were taken randomly from the original 100 questions in one paper. (We repeated the whole exercise with the other paper with no substantial change in the results.) In this way now, we had 11 test papers for each candidate – one original one and 10 fictitious ones.

Since the fictitious papers were made from the original test paper, the marks of each candidate in each of the 10 fictitious papers were known. Thus each candidate had 11 marks and ranks – one original and 10 in the fictitious papers. In fact, there are several students associated with each original rank and each of these students has 10 other fictitious ranks.

We plot the original data on a graph, i e, make a plot of the ranks in the original data vs the corresponding marks. This curve is shown as the unbroken line in Figure 1.

At each rank, we now take the number of students who got that rank and find their ranks and marks in the 10 fictitious papers. Typically, there are around a 100 students at most ranks and remembering that each of these students has 10 other ranks, we get a large enough sample to warrant a statistical interpretation. We take this distribution and compute its mean and standard deviation, which gives us an estimate of the spread around the mean.

This information is plotted in Figure 1 as the broken line. The data points are the mean of the distribution at each rank and the error bars are the standard deviation.

To make the data more obvious, the same plot for the first 30 distinct ranks is shown in Figure 2.

To further study the samples, we choose an arbitrary rank in the original paper. We take a rank near 1,000, assuming that in a typical examination, rank 1,000 would be the qualifying rank. At rank 1,060 (68 marks in the original Paper 1), we take the 164 students who have got this rank. For these 164 students, we consider their ranks in each of the 10 fictitious papers. Thus we have in all 1,640 ranks in the fictitious papers and find their mean and the standard deviation of this distribution. The distribution is shown in Figure 3.

Figure 1: Ranking Noise (Paper 1) with Errors

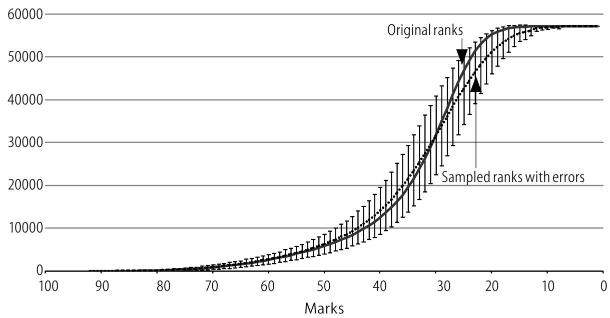


Figure 2: Ranking Noise (Paper 1) with Errors

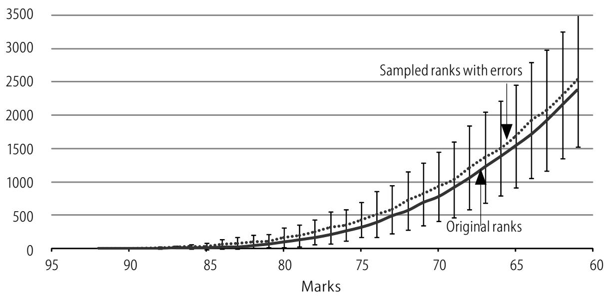


Figure 3: Histogram of Ranks in 10 Sample Papers for Students Having 68 Marks in Original Paper

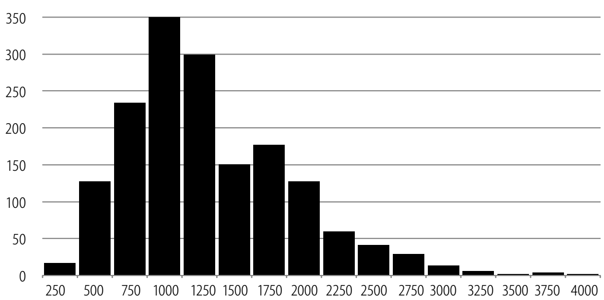
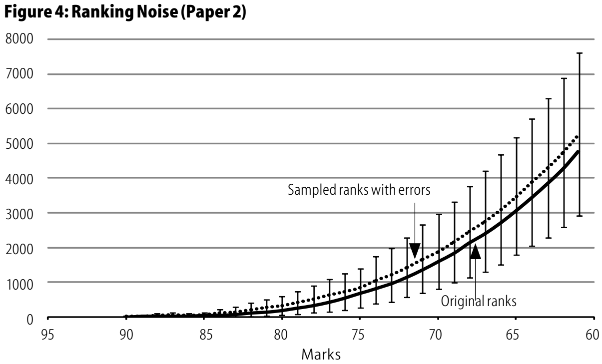


Figure 4: Ranking Noise (Paper 2)



It is interesting to see the same graphs for Paper 2 which we exhibit as Figure 4. The second paper shows a trend similar to the first paper. In detail however the number of candidates at each rank is more than in the Paper 1. The curve therefore rises more steeply and the standard errors are also a little higher.

The detailed graph of the sample in this case is shown for the rank 968 (73 marks in Paper 2) as Figure 5 (p 23). The number of candidates at this rank in this paper is 185. This histograms presented is therefore for 1,850 ranks. The mode of the distribution is near 1000 and the quartiles are near 550 and 1,575.

We have performed this analysis with a set of 10 sample papers. To see if the trend is dependent on the sample size (number of papers) we repeated the analysis for a sample of 50 papers and we find a very similar histogram. This is shown in Figure 6 (p 23).

From the data, we calculate the number of "students" who are in a given range of marks. For Paper 1, we see that at rank 1,060 in the original paper, there were 164 students. If we take the distribution of their marks in the 10 fictitious papers, we see that roughly 50% of the students lie between rank 758 and 1,521, a relatively large spread in ranks. Similarly, for Paper 2, the rank we have chosen is 968 and in this case the spread for 50% is between ranks 646 and 1,639. It is remarkable that the two papers show a relatively similar spread.

Recall that these are students who got exactly the same marks (rank) in the original test. Furthermore, the fictitious papers that we have made are sampled from the original paper itself. But, even with this, there is indeed a spread in the ranks. What this implies is that the exact rank that a student gets in such an examination is crucially dependent on the sample of questions that are chosen from the larger set of questions. In our case, this was 50 questions from a

Figure 5: Histogram of Ranks in 10 Sample Papers for Students Having 73 Marks in Original Paper 2

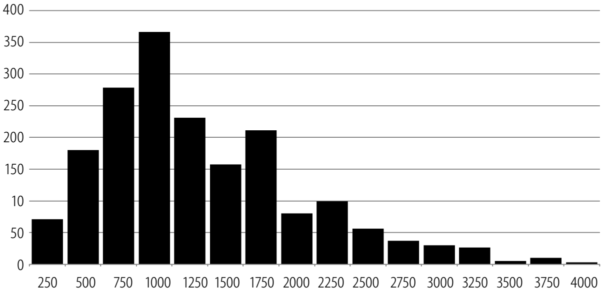


Figure 6: Histogram of Ranks in 50 Sample Papers for Students Having 68 Marks in Original Paper 1

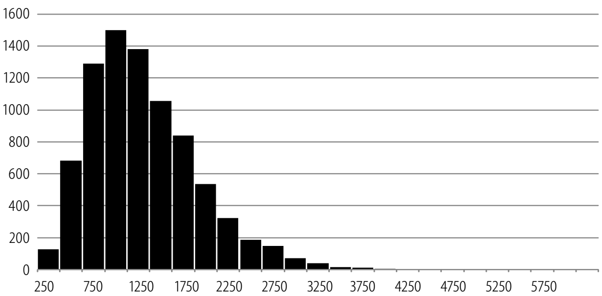
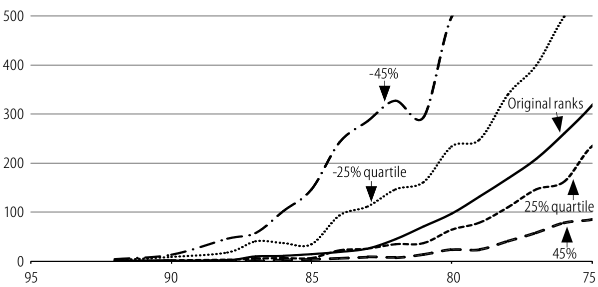


Figure 7: 50% and 90% Limits in the Sampled Ranks Compared with Original Rank



sample of 100, but in practice, the examiner is choosing the questions from a potentially very large set of possible questions of similar difficulty.

The Matter of Chance

This result is of profound consequence to the dynamics of multiple choice entrance examinations. What we see is that the same student who gets rank 1,000 (and thus, say, is admitted into the course) could as well have got rank 1,500 if the paper had been of similar difficulty but with, maybe, different questions. Clearly, at that rank he might not have made it into the course at all. We have done this analysis assuming a rank of 1,000 as the cut-off for admission, but clearly the conclusion will hold for any other rank as well. There will be significant number of students who could have obtained a lower or a higher rank – that they have got this rank is to a large extent a matter of pure chance.

Of course, this spread gets smaller as we go to higher ranks. In Figure 7 we show the spread for the first few distinct ranks. Here we see that the spread in ranks is much smaller in absolute terms, as is expected. However, this graph raises another critical issue – that of the significance of your absolute rank in determining the stream of your choice. In most entrance examinations (engineering entrance, postgraduate medical entrance, etc), the higher your rank, the more choice you have in terms of particular subject/stream. Thus the top 50 or 100 ranks in the Indian Institute of Technology-Joint Entrance Examination (IIT-JEE) can pick any stream in any IIT that they might want and this choice decreases as one goes down in ranks.

From Figure 7, we see that there is a considerable spread in the ranks, even at very high ranks. The multiple choice question papers are not able to distinguish between high ranks with any kind of precision and hence to use very rigid, rank-based criteria to determine options in terms of streams may not be a very fair method.

What we have thus shown is that in a multiple choice examination, there is a considerable element of chance in determining a candidate's rank. The origin of the chance factor that we have investigated is the sampling of the questions from the larger universe of similar questions. The spread in ranks because of the sampling process is actually fairly large, especially near the ranks where the cut-offs for admission are normally put. At higher ranks, this spread narrows down but there is still a spread which also negates the thesis that there is an inherent, qualitative difference in the capabilities of the person getting a very high rank and, say, one that is 50 ranks lower.

Competitive examinations started in China, hundreds of years ago to choose an elite, mandarin class from amongst the populace. In our own times, the fetish of these examinations has reached alarming proportions with students and parents. Playing on this fear of the parents and students, a huge coaching industry has mushroomed. The stigma attached to not "making" it into a professional course of one's choice is something which possibly scars the student for life. The underlying assumption being that one is not intellectually good enough since these examinations are ostensibly "true" indicators of one's academic abilities. Our analysis shows that this is far from true and that chance plays an enormous role in determining success in these kinds of examinations. Maybe, it is time for educational planners to rethink the inordinate emphasis placed on competitive examinations as the sole measure of intellectual abilities of students.