Normalization of marks in multi-session examinations

*Abhay G. Bhatt, Sourish Das and Rajeeva L. Karandikar**

When a test is conducted in several sessions using distinct question papers, normalization of scores is required to have a fair assessment of the candidates. Several selection tests nowadays are conducted in multiple sessions (using multiple choice questions). In this article we discuss various normalization schemes used in India when an examination involving multiple choice questions is conducted across various sessions. We illustrate through simulation, that the percentile-based normalization scheme outperforms all the other schemes.

Keywords: Multi-session examinations, multiple choice questions, normalization schemes, test scores.

IN recent times, examinations involving large number of candidates have been using multiple choice questions (MCQs). One reason for preferring these over traditional essay-type examinations is the enormous effort required to grade large number of answer books in a short time span in the latter case.

Till a few years ago, in most MCQ-type tests, the candidates marked the answer by pencil on specially formatted answer sheets. The sheets were read by a machine to capture the answers and results were prepared.

Over the last decade, some of these MCQ-type tests are administered via computers – candidates read the questions on a computer terminal and give their answers using mouse and keyboard, which are instantly captured in a database. This method has obvious advantages. It removes the need for printing large number of question papers and sending them to various centres, thus reducing the chances of foul play.

However, one limitation this brings in is that it puts an upper bound on the number of candidates that can appear in a test, as we need as many computer terminals as there are candidates. If the number of candidates is much larger than the number of computer terminals available for administering the test, the way out is to create two or more (as many as required) question papers. Candidates are divided in groups so that each group can be administered the test in one time slot and for each group, a distinct question paper is used.

The institution or the entity conducting the examination tries to ensure that the different question papers are of a same level of difficulty. In practice, however, this is difficult to achieve.

Abhay G. Bhatt is in the Indian Statistical Institute, New Delhi 110 016, India and Sourish Das and Rajeeva L. Karandikar are in the Chennai Mathematical Institute, Chennai 603 103, India.
*For correspondence. (e-mail: rlk@cmi.ac.in)

If a question could be used on multiple occasions, the difficulty level could be estimated statistically based on the earlier occasions when it was used. This is what is done in examinations such as GRE, TOEFEL, etc. where the questions are chosen out of a question bank and then suitable methods such as item response theory (IRT) are used to get the final score of each candidate[1–3]. However, this is not commonly done in India – in most examinations, a question once used in a test is not used again. This rules out use of IRT. Thus, the only way to assess the difficulty level of each question is using the opinion of experts. However this does not ensure that all question papers have the same level of difficulty as the perception of experts about difficulty levels is subject to judgemental errors.

The question then arises as to how can one compare the performance of two candidates who have appeared for the examination in two different shifts (and hence answered two different sets of questions).

This is being done by normalizing the marks of candidates in different shifts by putting them on a common scale in such a way that makes them amenable to comparisons. These normalized marks or scores are then used to rank the candidates for selection for admission or job, or for further screening.

For this, the candidates are assigned randomly to different sessions (or time slots) so that we can be assured that the talent that we are looking for is equally distributed across these sessions. Thus, if we see that the marks in one group are more than that in another group, we can conclude that this is mainly due to difference in difficulty levels. Thus to be fair to the candidates and to select the best candidates from among the applicants, some correction needs to be applied. This is achieved by normalizing the marks.

Various methods are used in practice for normalization of marks. These involve transformations of the raw

scores, or the actual marks secured by the candidates. These transformations are typically based on some statistical quantities – like mean, standard deviation, percentiles, etc. of the scores in that shift or that of a subset of this dataset.

The results of the transformation will give normalized scores which then can be used to rank the candidates across the different shifts.

We have seen data from various selection tests conducted across India in the recent past and it has been observed in few cases that the selected candidates were dominated by those appearing in one or two sessions while some sessions were highly under represented. The discrepancy was so much that the end-user agencies themselves were wondering as to how to defend the same if challenged in a court of law. This raises the question as to which is the right method of normalization.

Different normalization methods will give different rankings of the candidates. It is then necessary to know which normalization methods are reasonable and which are not so good, or, out of a set of proposed methods, which one is the best.

Normalization schemes

In the description below, we assume that there are k question papers (administered in k distinct time slots). We denote by G_i, the set of candidates who answered the ith question paper. We describe four methods of normalization in this section and compare them in the following section.

z-Score method

One of the most commonly used methods of normalization is to transform the score using mean and standard deviation. For every group G_i, the mean marks μ_i and standard deviation of marks σ_i among all the candidates in the group G_i are calculated.

The marks s of a student in the group G_i are transformed to $T_i(s)$ by the transformation

$$T_i(s) = \mu^* + (s - \mu_i)\frac{\sigma^*}{\sigma_i}, \tag{1}$$

where μ_i, σ_i are the mean and standard deviation of the marks of candidates in the ith group and $\mu^* = \max\{\mu_l : 1 \le l \le k\}$ and $\sigma^* = \max\{\sigma_l : 1 \le l \le k\}$.

Thus for a candidate with raw score S and belonging to the group G_i, his/her normalized score is $T_i(S)$.

The normalized scores of all candidates are taken together to generate the ranks or merit list. This formula has an advantage that the normalized score of each candidate is larger than or equal to his/her raw score. However, the normalized score can be higher than the maximum score.

Other choices of μ^* and σ^* are also used; for example, σ^* could be the standard deviation in the group with highest mean. It can be seen that the ranks (or the merit list) produced by different choices of μ^* and σ^* are the same. Indeed, denoting by

$$E_i(s) = \frac{(s - \mu_i)}{\sigma_i}, \tag{2}$$

it can be seen that the ranks produced by $\{T_i : 1 \le i \le k\}$ for any choice of μ^* and σ^* (with $\sigma^* > 0$) are the same as the ranks produced by $\{E_i : 1 \le i \le k\}$. Thus we call $E_i(S)$ as the standardized score of a candidate with score S from group G_i. We refer to this as the z-score method.

w-Score method

The standardization using eq. (1) is perhaps motivated by the belief that when there are a large number of candidates in each group, the distribution of marks in each group would be normal and the standardization via eq. (1) would transform them to the same distribution, namely standard normal distribution.

Looking at scores of several examinations with a large number of candidates we have seen that in most situations, the distribution of marks is far from normal – the deviation is maximum in the tails of the distribution. In cases where the examination is to be used for selection, the interest is in the candidates whose scores are in the top few per cent or the upper tail of the score distribution.

In view of this, another method considered is as follows: suppose the top 1% candidates are to be selected. Then let

$$F_i(s) = \frac{(s - \xi_i)}{\theta_i}, \tag{3}$$

where ξ_i, θ_i are the mean and standard deviation of the marks of top 1% candidates in the ith group. This yields the standardized scores of candidates – the score of a candidate in the ith group with score S is $F_i(S)$. We call this the w-score method.

g-Score method

Another method currently being used in India, including for GATE and CAT, is the following (subsequently called the g-score method). Here the normalized score is given by

$$M_i(s) = \alpha + (s - \alpha_i)\frac{(\beta - \alpha)}{(\beta_i - \alpha_i)}, \tag{4}$$

where α is the sum of mean and standard deviation of all candidates, α_i the sum of mean and standard deviation of

all candidates in the group G_i, β the mean of the top 0.1% of all candidates and β_i is the mean of the top 0.1% of all candidates in group G_i.

Let us note that here if we take the standardized score as

$$N_i(s) = \frac{(s - \alpha_i)}{(\beta_i - \alpha_i)}, \tag{5}$$

then the rankings produced by the standardized scores given by eq. (5) and normalized scores given by eq. (4) are the same. We call this the *g*-score method.

p-Score method

Here, instead of the transformation by mean and standard deviation, the percentile score in each session is taken as the standardized score. This has an advantage as it does not assume any specific form of the distribution of marks. It does not even require that the distributions across the groups be the same.

It should be noted that when the data have ties (which is invariably the case when we have data on scores of a large number of candidates), the ranks are not uniquely defined and each statistical software has its own default method. Thus the method to resolve the ties has to be specified by the end-user. In the context of normalization, it makes sense to assign equal score to the toppers in all the shifts. This is achieved by defining the standardized score $P_i(s)$ corresponding to a score s of a candidate in the group G_i as follows:

$$P_i(s) = \frac{\gamma_i(s)}{\lambda_i}, \tag{6}$$

where $\gamma_i(s)$ is the number of candidates in the ith shift scoring less than or equal to s marks, and λ_i the total number of candidates in the ith shift who appeared for the examination. We call this the *p*-score method.

We have thus described four methods of transforming raw scores of candidates across the k groups $\{G_i : 1 \le i \le k\}$ to standardized scores via the transformations $\{E_i, : 1 \le i \le k\}$ (z-score), $\{F_i, : 1 \le i \le k\}$ (w-score), $\{N_i, : 1 \le i \le k\}$ (g-score) and $\{P_i, : 1 \le i \le k\}$ (p-score).

The standardized scores can then be transformed to a suitable scale to bring it say, in the same range as the raw scores, or between 0 and 100. The transformation of standardized scores to normalized scores is via one fixed increasing function so that the ranks based on standardized scores are the same as those based on normalized scores. This has a psychological aspect – candidates are upset if the normalized score is less than their raw score, but are happy if it is more. However, if only ranks matter, then only standardized score matters and the final transformation to convert to normalized score is not important.

The choice of this transformation is important if the normalized score is used, over and above the ranks, for any decision making, say, when it is combined with a score in the interview to generate the final ranking.

Comparisons via simulation

Since the aim of normalization (or standardization) is to correct for difference in difficulty levels of two examinations, let us consider the ideal case, when the two question papers are of the same difficulty level.

So if we assume that all candidates actually answered the same question papers but have been randomly tagged as group 1, group 2, etc.

Then if we are selecting say p% of the candidates, then roughly p% candidates from each group should make it to the selected list. So the difference between selected proportions across the groups is an indication of the distortion the normalization method is introducing.

If $x_1, x_2, \ldots, x_k$ are the proportions of candidates in each group that are selected, and if $\tilde{p} = p/100$, then the quantity

$$\frac{(\max\{x_i : 1 \le i \le k\} - \min\{x_i : 1 \le i \le k\})}{\tilde{p}},$$

denotes deviation, and higher the deviation, the worse we are from ideal selection criterion. We express this as a percentage

$$D = \frac{(\max\{x_i : 1 \le i \le k\} - \min\{x_i : 1 \le i \le k\})}{\tilde{p}} \times 100,$$

Table 1. Two groups – score distribution: normal

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	6.07	4.58	2.45	5.12	8.74
w-Score	5.31	3.99	2.12	4.50	7.70
g-Score	17.03	12.69	6.90	14.54	24.52
p-Score	0.20	0.11	0.10	0.20	0.30

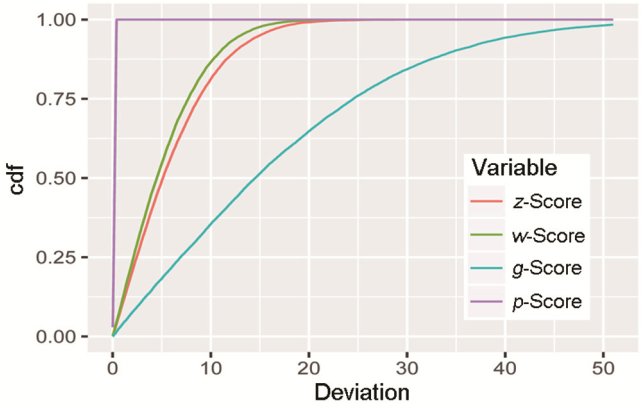

Figure 1. Two groups – score distribution: normal.

Consider the case of two groups, each of 25,000 candidates and the target is to select 1% candidates. If instead of selecting 250 from each group, we select 225 and 275 respectively, from the two groups, the deviation becomes 20%.

The deviation thus measures the gap between maximum and minimum across groups as a percentage of the target number from each group.

Table 2. Two groups – score distribution: Laplace

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	5.79	4.32	2.34	4.94	8.34
w-Score	6.83	5.19	2.72	5.76	9.81
g-Score	18.89	14.12	7.63	15.98	27.29
p-Score	0.20	0.12	0.10	0.20	0.30

Table 3. Two groups – score distribution: uniform

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	24.96	18.82	10.06	21.06	36.19
w-Score	2.59	1.94	1.03	2.22	3.76
g-Score	6.79	5.14	2.70	5.70	9.82
p-Score	0.20	0.12	0.10	0.20	0.30

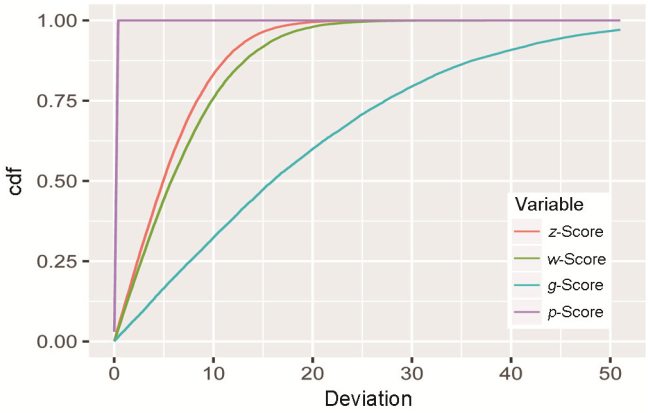

Figure 2. Two groups – score distribution: Laplace.

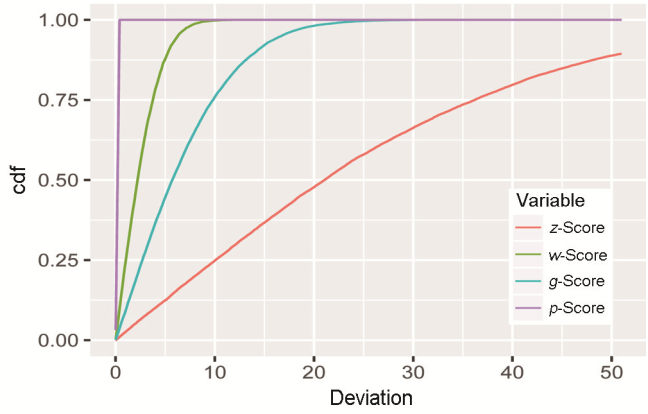

Figure 3. Two groups – score distribution: uniform.

Taking the number of groups to be two and 25,000 candidates from each group, we will simulate scores of the candidates from a normal distribution and for each of the four methods of normalization, we compute the deviation d. We will repeat this 20,000 times. Thus for each of the four methods we obtain the distribution of the deviation. Table 1 gives the mean, standard deviation and the three quantiles for each of the four methods. Figure 1

Table 4. Two groups – score distribution: $t(3)$

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	11.85	14.34	3.90	8.34	14.94
w-Score	27.52	22.44	10.28	22.17	39.06
g-Score	41.13	30.62	16.65	35.14	59.55
p-Score	0.20	0.12	0.10	0.20	0.30

Table 5. Five groups – score distribution: normal

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	12.59	4.66	9.24	12.24	15.52
w-Score	11.02	4.13	8.02	10.69	13.60
g-Score	34.76	12.87	25.45	33.66	42.87
p-Score	0.32	0.07	0.28	0.34	0.37

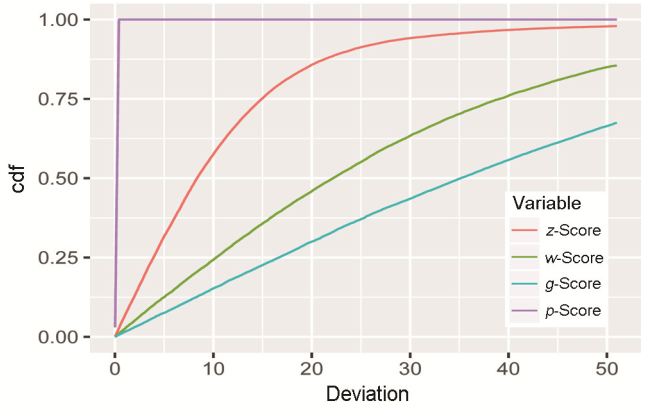

Figure 4. Two groups – score distribution: $t(3)$.

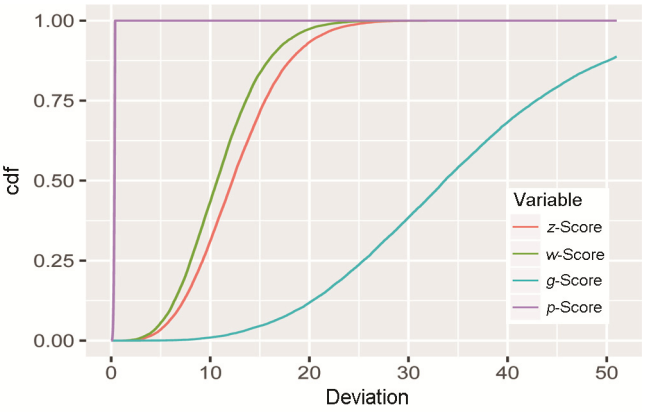

Figure 5. Five groups – score distribution: normal.

gives the (estimated) cumulative distribution function (cdf) of the deviation D.

We can see that the p-score method performs the best, with mean of the deviation being only 0.2, the z-score and w-score are comparable with mean of the deviation above 5 and the g-score performs very poorly, with mean being over 17. Indeed, we can see from the cdf that the graph for p-score rises to 1 sharply, while the other graphs rise gradually – this means that for p-score, the deviation is less than 1% with a very high probability (over 99%).

We now come to the case when the underlying distribution is Laplace. Table 2 and Figure 2 give results when the score distribution is Laplace. We see that the p-score still performs well and the g-score is the worst.

Next we take score distribution to be uniform. Table 3 and Figure 3 show the results. This time we see that while the p-score method is the best and very good, the z-score performs rather poorly.

Table 4 and Figure 4 show results for the case where the underlying score distribution is t with three degrees of

Table 6. Five groups – score distribution: Laplace

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	11.84	4.39	8.69	11.54	14.61
w-Score	14.29	5.57	10.28	13.66	17.63
g-Score	39.14	14.51	28.60	38.06	48.36
p-Score	0.32	0.06	0.28	0.34	0.37

Table 7. Five groups – score distribution: uniform

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	51.00	18.99	37.14	49.39	62.89
w-Score	5.33	1.99	3.89	5.17	6.58
g-Score	14.08	5.21	10.30	13.68	17.44
p-Score	0.32	0.07	0.28	0.34	0.37

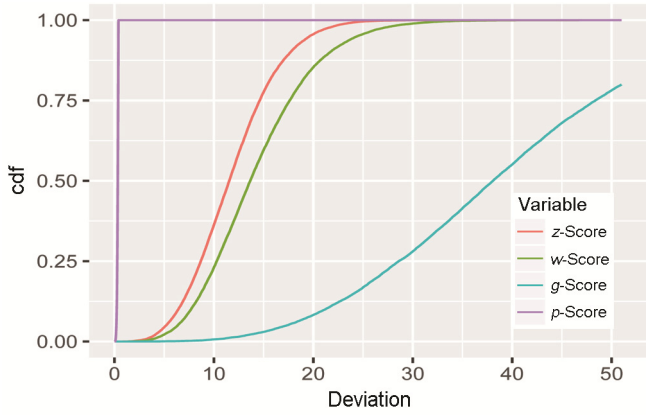

Figure 6. Five groups – score distribution: Laplace.

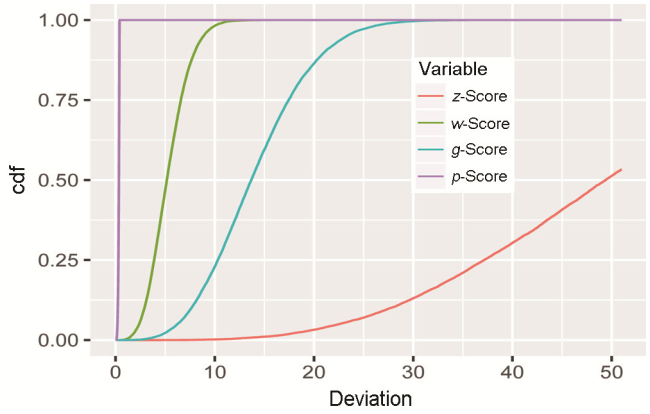

Figure 7. Five groups – score distribution: uniform.

Table 8. Five groups – score distribution: $t(3)$

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	24.00	15.37	14.68	20.24	28.25
w-Score	72.06	47.97	39.83	58.04	88.80
g-Score	84.69	31.72	61.88	81.81	104.20
p-Score	0.32	0.07	0.28	0.34	0.37

Table 9. 75 groups – score distribution: normal

	Mean	SD	$Q1$	$Q2$	$Q3$
z-Score	25.86	3.32	23.66	25.66	27.92
w-Score	23.22	3.42	20.83	22.93	25.21
g-Score	71.56	9.66	64.96	70.43	77.69
p-Score	0.39	0.01	0.39	0.40	0.40

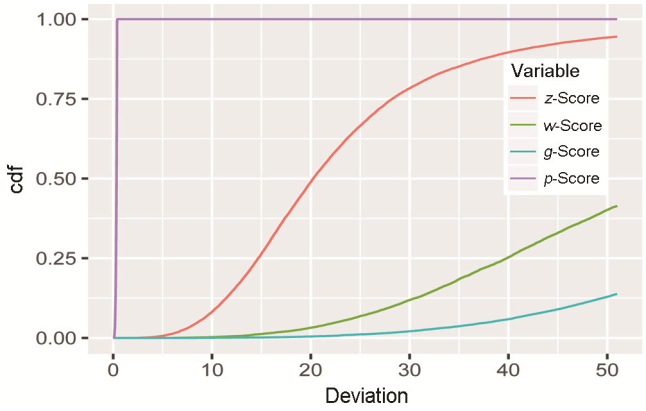

Figure 8. Five groups – score distribution: $t(3)$.

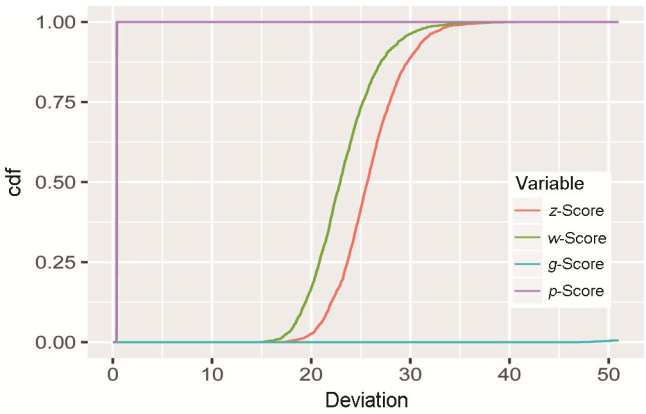

Figure 9. 75 groups – score distribution: normal.

freedom. This time we see that while the *p*-score method is the best and very good, the other methods perform rather poorly.

The discrepancies increase when the number of parallel sessions increases. We illustrate the same via simulation when we have five groups. The results are given in Tables 4–8 and Figures 4–8.

We can see that while the *p*-score method continues to do well, for all the other methods, deviations are more than what they were in case of two groups. When using the *g*-score method, if the score distribution is $t(3)$, then with 50% probability we are likely to see a deviation of over 80%. This means, while we are targeting 250 candidates from each group, the gap between minimum and maximum number of candidates across groups is over 200.

We will like to add that for a given question paper we do not know before hand as to what would be the distribution of scores. Thus, it makes sense to use a method that makes no assumption about the underlying distribution. The *p*-score is such a method.

The distortion only increases if the number of groups increases. We are giving here the statistics of deviation when there are 75 groups – only when the distribution is normal (Table 9 and Figure 9). In other cases, the methods other than the *p*-score method do much worse.

Final recommendation

We have argued that the standardization step should use percentile score (in each group). The topper in each group would have score 1.

If this score is to be used further, one could transform it suitably. One possibility is to multiply the standardized score by 100 to yield a score between 0 and 100.

Yet another possibility is to convert the scores to a range that is different from the original (raw) scores so that no one considers that their scores were reduced in normalization. Also, one could avoid fractional scores. Here is a suggested transformation that will map the scores to an integer in range 300–800

$$W_i(s) = 300 + \frac{\gamma_i(s) * 500}{\lambda_i}, \tag{7}$$

where $\gamma_i(s)$ is the number of candidates in the ith shift scoring less than or equal to s marks, and λ_i is the total number of candidates in the ith shift who appeared for the examination.

The final normalized score $X_i(S)$ of a candidate in ith group with score S is defined as the smallest integer greater than or equal to $W_i(S)$. The topper in each group will have a score of 800.

1. Baker, F., The *Basics of Item Response Theory*, ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, MD, USA, 2001.
2. Fox, J.-P., *Bayesian Item Response Modeling: Theory and Applications*, Springer, 2010.
3. Takane, Y. and de Leeuw, On the relationship between item response theory and factor analysis of discretized variables *J. Psychometr.*, 1987, **52**, 393–408.

Received 17 February 2019; revised accepted 14 August 2019

doi: 10.18520/cs/v118/i1/34-39