Shana Moothedath
Department of Electrical Engineering, Indian Institute of Technology Bombay
Email: shana@ee.iitb.ac.in

Prasanna Chaporkar
Department of Electrical Engineering, Indian Institute of Technology Bombay
Email: chaporkar@ee.iitb.ac.in

Madhu N. Belur
Department of Electrical Engineering, Indian Institute of Technology Bombay
Email: belur@ee.iitb.ac.in

DOI: http://dx.doi.org/10.18820/2519593X/pie.v34i4.7
ISSN 0258-2236
e-ISSN 2519-593X
Perspectives in Education
2016 34(4): 99-115

A maximum likelihood based offline estimation of student capabilities and question difficulties with guessing

Abstract:

In recent years, the computerised adaptive test (CAT) has gained popularity over conventional exams in evaluating student capabilities with desired accuracy. However, the key limitation of CAT is that it requires a large pool of pre-calibrated questions. In the absence of such a pre-calibrated question bank, offline exams with uncalibrated questions have to be conducted. Many important large exams are offline, for example the Graduated Aptitude Test in Engineering (GATE) and Japanese University Entrance Examination (JUEE). In offline exams, marks are used as the indicator of the students' capabilities. In this work, our key contribution is to question whether marks obtained are indeed a good measure of students' capabilities. To this end, we propose an evaluation methodology that mimics the evaluation process of CAT. In our approach, based on the marks scored by students in various questions, we iteratively estimate question parameters such as difficulty, discrimination and the guessing factor as well as student parameters such as capability using the 3-parameter logistic ogive model. Our algorithm uses alternating maximisation to maximise the log likelihood estimate for the questions and students' parameters given the marks. We compare our approach with marks-based evaluation using simulations. The simulation results show that our approach out performs marks-based evaluation.

Keywords: 3-parameter logistic IRT model, alternating optimisation, offline exams, computerised adaptive test.

1. Introduction

The multiple choice exams are the most popular assessment scheme for large scale exams such as the computerised adaptive test (CAT), Graduate Record Examinations **(**GRE), Scholastic Aptitude Test (SAT) and so on. The important features of multiple-choice exams that make it more popular are that these exams are easy to evaluate and the evaluation criteria can be implemented uniformly without any bias. In subjective exams where students give descriptive answers for every question, the question of partial correctness comes into play, which may result in a biased evaluation. In addition, the time and effort required for evaluating a subjective exam is quite high. On the other hand, in the case of multiple-choice exams, there will be exactly one correct answer and the whole notion of a

"partially correct answer" disappears. However, the effect of guessing appears in multiple-choice exams. By way of example, consider an item with four options out of which exactly one is the correct answer and the remaining three are distractors. In this case, a student of extremely low ability who is unprepared for the exam has a 0.25 probability of answering it correctly through guessing. Thus, a guessed response even though it does not give any information about the actual capability of a candidate contributes to his/her test score and thus skews the assessment. Moreover, in situations with "partial knowledge" the guessing factor becomes more significant, since even without knowing the correct answer to an item if a candidate is successful in eliminating a few distractors of the item with his/her partial information about the item, his/her chances of getting it correct is greater. Thus, the probability that a candidate with partial knowledge about the item getting it correct through guessing is greater that particular item fails to distinguish a candidate with partial knowledge from a candidate with full information. On the other hand, a student who knows the basic method to solve an item can make minor errors, which can lead to the wrong choice of response and get zero credit for that item. Therefore, the effect of guessing depends on the nature of the item and is thus an item parameter.

CAT is one of the most popular evaluation schemes (Van der Linden, Wim & Glas, 2000). The important feature of CAT that popularises it, is the "adaptive" feature of conducting the exam. In CAT, the test item that a candidate is going to answer next depends on his/her responses to the previous questions. If a candidate answers a question of a certain difficulty level correctly, s/he will be given a question of slightly greater difficulty level. However, if the response is not correct, then the next question will be slightly less difficult. For the adaptive selection of items for all the test takers, CAT maintains an item pool that consists of a large number of items spanning a range of content levels and difficulty levels and every item is selected based on a selection algorithm. In this way, every candidate taking a CAT exam will undergo a self-tailored exam. Even though CAT exams are superior to other exams in various aspects, it does have some shortcomings as well (Way, Davis & Fitzpatric, 2006). The difficult task of conducting a CAT exam is the construction and maintenance of the item pool. The item pool is the prime requirement and it should contain questions in a wide range of difficulty levels so that the exam is good enough to estimate the capabilities of low and high capability candidates. The challenges associated with constructing and maintaining the item pool is:

1) Questions in the item pool should be pre-calibrated. For calibrating questions, extra test items are given as field tests in every exam. These are uncalibrated questions that are given in the exam which do not affect the test score of a candidate but whose difficulties are determined from the responses of the candidates whose capabilities are estimated from the pre-calibrated questions. The difficulty of an item is fixed only after taking a sufficient number of field tests. The problem associated here is that, as a number of students see these questions, the difficulty of the question is no longer the estimated one. Thus, the questions that the candidates see will not have the calculated difficulties when they are used for testing. Therefore, the entire process of calibrating questions and then estimating the capabilities of candidates using those calibrated questions will be erroneous in a cyclic manner.

2) The item pool should be periodically repopulated. Items that are frequently given for the exam will become known to the examinees. Consequently, the difficulty of the question is different from the calibrated value as time progresses. This will result in the wrong estimation of capabilities. To avoid this, the pool should be restored and fresh items should replace the

known ones. However, accurately finding the time point at which an item in the pool is to be replaced is not an easy task.

3) In exams with many disciplines, constructing and maintaining an item pool for each discipline is quite a difficult task. It is very expensive and requires plenty of effort to construct an item pool.

4) One other issue with CAT exams is the option to go back in the exam. More clearly, a candidate can return to a previous question and reattempt it at any point in time in the exam. While there is variation across various adaptive tests on whether or not to allow modifying past attempts, there is disagreement about to what extend this will help in estimating the parameters (Way *et al*., 2006). Since this feature of reattempting items is not incorporated in our analyses, we will not pursue it in this work. Apart from all these difficulties, security issues are also a major concern in CAT exams.

Charles Spearman came up with the first theory of psychometric test analysis known as the classical test theory (CTT) in 1906. Sixty years later, Lord and Novick reformulated CTT using a modern mathematical statistical approach (Lord, Novick & Birnbaum, 1968). The main shortcoming of CTT is that it does not consider the item properties. To be precise, in multiple-choice exams where the total score is considered as the measure of the candidate's capability, the items that a candidate answered correctly does not play any role in deciding his/her capability. In such a situation, answering an easy question correctly and a very difficult question correctly fetches him/her the same credit that does not seem to be appropriate. Binet and Simon (1916) introduced the item based test theory known as the item response theory (IRT) in 1916, where the item parameters such as the difficulty of the question are also considered in the assessment. This paper uses an IRT model for all the analyses conducted.

In this paper, we are focusing on offline exams. By the term "offline", we mean the exams in which the scores of the test takers are not available after the end of the test. In CAT exams, by the end of the test, each examinee gets to know his test score. However, in offline exams the test score is disclosed to the public as well as the test takers after a certain timeframe. In these exams, test scores of a candidate not only depends on his/her sole performance but also on the general nature of the exam. In addition, here the questions are not pre-calibrated.

The main point that we are focusing on in this work is that in offline exams when total marks are used as the input measure for estimating the capabilities of the students, then score comparison across disciplines, years and sessions is not justified. Scores need to be compared across disciplines when students with scores in different disciplines apply for a common programme. For example, a student with a score in computer science engineering can apply for a programme in electrical engineering and vice versa. Similarly, many interdisciplinary courses consider scores from various disciplines while applying. Therefore, score comparison across disciplines becomes vital. Score comparison across years becomes relevant in those exams that have a validity of more than a year. In such exams, a candidate can apply for a programme while his/her score is valid. In such a case, it is imperative to compare scores across years. The third scenario is a multiple session exam, where students take exams in different batches answering different question papers and are finally ranked in a single rank list. For example, in cases of large-scale offline exams such as GATE, students take tests in different test centres for the same discipline by answering different question papers and are finally ranked in a common rank list. Here question papers are different for different batches and therefore comparison of scores cannot be justified if total marks are used as the only deciding parameter.

Summary of contribution

We propose a maximum likelihood based alternating optimisation algorithm for the three-parameter logistic model for estimating the student parameter, capability and the question parameters, difficulty, discrimination and guessing. In our previous work, we proposed an alternating optimisation based estimation of student capabilities and question difficulties (Moothedath, 2016) for the two-parameter Rasch model. The effect of guessing is not considered in that work. In this paper, the effect of guessing is included and experimental results are demonstrated to compare the proposed maximum likelihood based algorithm with the mark-based method. However, the exams considered in this work are not adaptive and negative marking is not considered here.

Organisation of the paper

Section 2 details the model employed in this work for estimating the student parameter, capability and the question parameters, difficulty, discrimination and guessing. The details of the maximum likelihood estimation are given in section 3 and the likelihood function of the concerned problem is formulated here. Section 4 summarises the pseudocode for the proposed scheme. For verifying the performance of the proposed maximum likelihood based scheme, we conducted a few experiments. The details of the experiments conducted and the metrics that are used for the comparison is given in section 5. Simulation results corresponding to these experiments are given in section 6. Section 7 and section 8 comprises of concluding remarks and references respectively.

2. Model

This section discusses the model used in this paper for assessment. We employed Birnbaum's three-parameter model for all the analyses done in this work. Birnbaum proposed an item characteristic curve which (Baker, 1985) gives the probability of j^{th} student answering i^{th} question correctly.

$$P_i(c_j) = g_i + (1-g_i)\,\frac{exp\{a_i(c_j-d_i)\}}{1+exp\{a_i(c_j-d_i)\}} \tag{1}$$

where c_j denotes the capability of the j^{th} student and d_i, a_i and g_i denotes the difficulty, discrimination and guessing factor of the i^{th} question respectively. The guessing factor is the likelihood that a student of extremely low ability answers the item correctly. The parameters of the model are as follows: (1) capability c_j, (2) difficulty d_i, (3) discrimination a_i, (4) guessing factor g_i. Henceforth, *i* stands for the question index and *j* denotes student index.

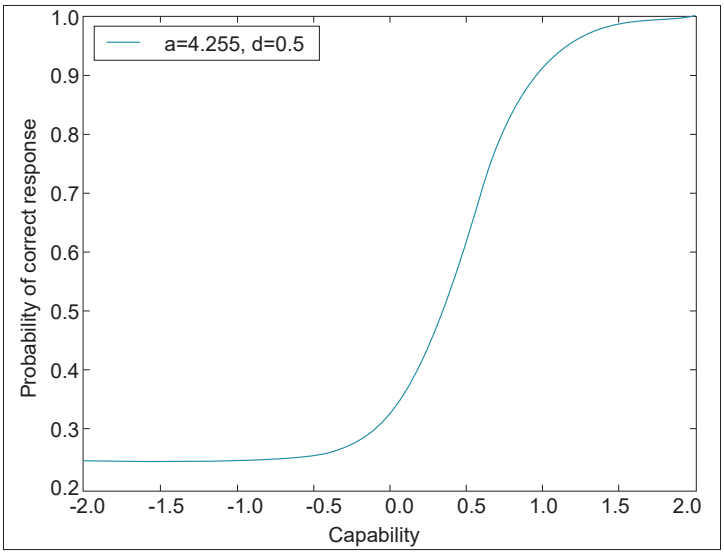

Figure 1: ICC for correct response with d = 0.5, a = 4.255 and g = 0.25

Figure 1 is the item characteristic curve (ICC) which shows the variation of probability of answering a question of difficulty 0.5 correctly when 0.25 guessing factor is involved. The plot shows that an unprepared candidate who has no knowledge about the item can answer it correctly with 0.25 probability. In addition, as capability increases the probability of answering correctly also increases and finally saturates to 1.

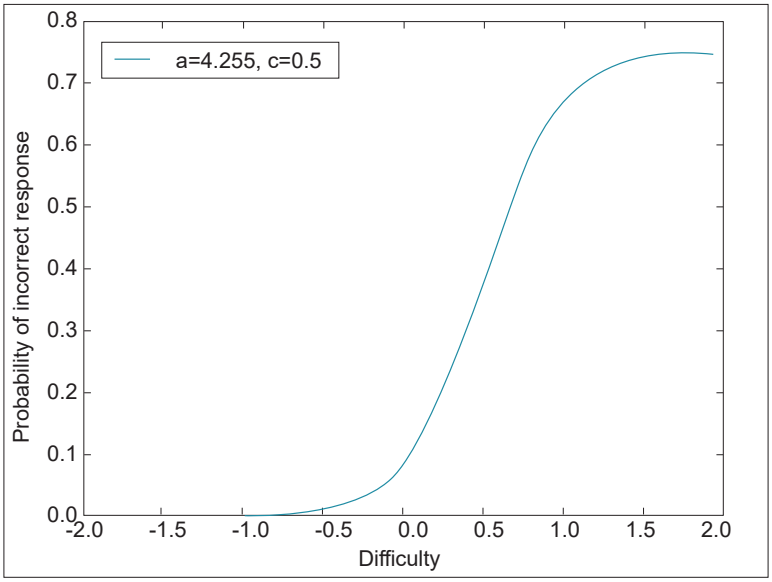

Figure 2: ICC for incorrect response with c = 0.5, a = 4.255 and g = 0.25

Figure 2 shows the item characteristic curve that shows the variation of probability of answering incorrectly as a function of question difficulty. The plot shows that for quite a low difficulty

question the probability of answering incorrectly is very low and as the difficulty increases the probability of answering incorrectly saturates to a value lower than 1. Thus, even when the question difficulty is very high when compared to the capability, there is still the probability of answering it correctly because of the guessing factor involved.

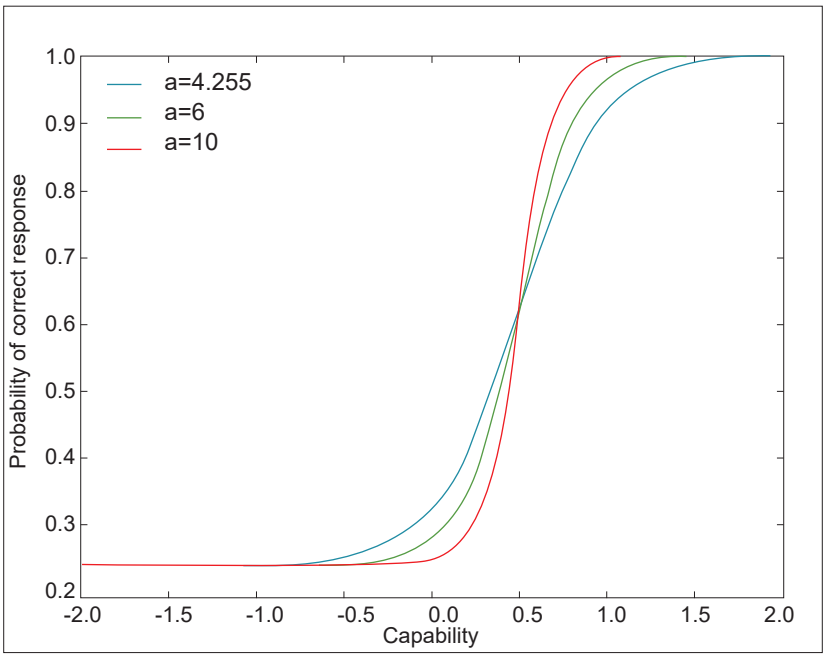

Figure 3: ICC for correct response with d = 0.5 and g = 0.25

Figure 3 shows the variation of probability of answering correctly for questions for different discrimination values. It is clear from the plot that as the discrimination value increases the plot becomes more and more steep. The steeper the curve, the better the item is, as it can differentiate candidates of diverse capabilities. Thus, it is always advisable to include questions of large discrimination values in the test. However, constructing items with large discrimination levels is very difficult.

3. Maximum likelihood estimation

Maximum likelihood estimation is a technique for estimating the parameters of a statistical model, given the observations. It estimates the parameter values that maximise the likelihood of making the observations, given the parameters. The likelihood of a set of parameters, θ, given the response X is given by,

$$L(\theta|X) = P(X|\theta) \tag{2}$$

The objective of this paper is to estimate the capabilities of the test takers and the difficulty, discrimination and guessing of the items of the test, given the responses. The responses of the candidates are a dichotomous data set denoted as R, the response matrix. The matrix R has students as the rows and questions forming the columns of the matrix. Let nS denote

the number of students and nQ denote the number of questions. Thus, the matrix R is a nS × nQ matrix. Negative marking is not considered in this paper and so every entry in R is either 0 or 1. One corresponds to a correct response of the student and 0 corresponds to an incorrect response or an unattended question. Thus, the problem can be formulated as, given the response matrix R of an exam, we need to estimate the capability vector, $C = [c_1,c_2,...,c_{ns}]$, the difficulty vector, $D = [d_1,d_2,...,d_{nQ}]$, the discrimination factor, $A = [a_1,a_2,...,a_{nQ}]$ and the guessing vector $G = [g_1,g_2,...,g_{nQ}]$. The likelihood function for the above problem is:

$$L(C,D,A) = Prob(R|C,D,A) \tag{3}$$

The likelihood function of a student depends on his/her response to all the questions. While for a question, the likelihood function depends on the responses to that particular question by all the students. Therefore, the likelihood function of the entire test is the product of the likelihood functions of each student for all questions under the assumption that all examinees are independent. For $L(C,D,A) = Prob(R|C,D,A)$, the logistic ogive model, the logistic function is given by

$$P_{ij} := g_i + (1-g_i)\frac{exp\{a_i(c_j-d_i)\}}{1+exp\{a_i(c_j-d_i)\}} \tag{4}$$

$$Q_{ij} := (1-g_i)\frac{1}{1+exp\{a_i(c_j-d_i)\}} \tag{5}$$

where c_j denotes the capability of the j^{th} student, d_i, a_i and g_i denotes the difficulty, discrimination and guessing factor of the i^{th} question respectively.

Then the global likelihood function of the exam is formulated as

$$Prob(R|C,D,A := \Pi_{j=1}^{nS}\Pi_{i=1}^{nQ} \; P_{ij}^{m_{ij}}(1-P_{ij})^{1-m_{ij}}, \tag{6}$$

where nS is the number of students taking the test, nQ is the number of questions in the test and m_{ij} is the entry in the response matrix R corresponding to the $(i,j)^{th}$ location. If student i made item j correct, then m_{ij} =1, else it is 0. Using the logarithm, we get the log likelihood function as

$$L(C,D,A) := logProb(R|C,D,A) = \sum_{j=1}^{nS}\sum_{i=1}^{nQ} m_{ij} logP_{ij} + (1-m_{ij})log(1-P_{ij}). \tag{7}$$

4. Proposed algorithm

We propose a maximum likelihood based alternating optimisation algorithm for solving this. Alternating optimisation, otherwise called the Gauss Siedel optimisation method, is a technique for optimising functions involving a large number of variables by partitioning the set of variables into different blocks. In every step, optimisation is done in one block of variables keeping the other sets fixed and this is done sequentially. We want to maximise the likelihood of the exam given by equation (6). However, maximising (6) is equivalent to maximising (7), since the log is a monotonically increasing function. Thus the objective function here is the log likelihood function given by equation (7) and the variables over which optimisation is carried out, the *C*, *D, A* and G vectors. The pseudocode for the proposed algorithm is given below.

Algorithm

Input: Raw marks matrix

Output: Student capability vector *C*, question difficulty vector *D* and question discrimination vector *A*.

1: Initialise *D,A,G*.

2: **while** error norms of estimated levels in previous iteration ≥ tolerance value do

3: **for** each student *j* do

4: Using D, A and G find $c\epsilon[0,1]$ such that *L* is maximum.

5: $c_j := \underset{c}{argmax} L(\acute{C},D,A,G)$

6: **end for**

7: **for** each question i do

8: Using C, A and G find $d\epsilon[0,1]$ such that L is maximum.

9: $d_i := \underset{d}{argmax} L(C,\acute{D},A,G)$

10: **end for**

11: **for** each question i do

12: Using C, D and G find $a\epsilon[0,6]$ such that L is maximum.

13: $a_i := \underset{a}{argmax} L(C,D,\acute{A},G)$

14: **end for**

15: **for** each question i do

16: Using C, D and A find $g\epsilon[0,1]$ such that L is maximum.

17: $g_i := \underset{g}{argmax} L(C,D,A,\acute{G})$

18: **end for**

19: go to step 3

20: **End while**

5. Comparison metrics and variables

We conducted a few experiments to verify the performance of the proposed maximum likelihood (ML) based method with the conventional raw marks (RM) based method. For this we compare the raw marks rank list (RM rank list) and the maximum likelihood rank list (ML rank list) with the actual capability rank list (AC rank list). The AC rank list is the ordered list of students arranged in the decreasing order of their actual capability (AC) levels. The ML rank list is formed at the end of the estimation process by arranging candidates in the descending order of the estimated ML capability values. Similarly, candidates are arranged in descending order of total marks to form the RM rank list. For x% cut-off bound, the ML cut-off (RM cut-off) is the capability of the (x/nS) × 100^{th} candidate in the ML rank list (RM rank list). The experiments conducted are: (1) fixed number of students and varied number of questions, (2) fixed number of questions and varied number of students and (3) multiple session exam

where students take exams in batches answering different question papers but finally fall into a common rank list. The parameters used for comparing the rank lists and drawing conclusions are: (i) number of false-positives, (ii) number of desired students qualified and (iii) number of qualified students.

False-positives are the non-deserving set of candidates that enter the rank list within the cut-off bound after the assessment. In the ML rank list (RM rank list), these students' actual capability level is below the cut-off capability but they lie within the cut-off bound in the ML rank list (RM rank list). These candidates qualified for the exam but actually were not supposed to. Thus, it is always advisable to have a lower number of false-positives in the exam so that truly deserving candidates qualify for the exam.

Deserving candidates qualified in the ML (RM) scheme are those students that are present within the cut-off bound in the ML rank list (RM rank list) and AC rank list. Therefore, desired candidates are the population that corresponds to the set of students who are the actual deserving ones. It is advisable to have a greater number of deserving candidates in the rank list. The number of qualified students refers to the number of students who qualified for the exam. All students whose capabilities is greater than or equal to the cut-off capability in the respective rank lists is qualified in that particular rank list. That is, in the ML rank list (RM rank list) those students whose ML (RM) capability values are greater than or equal to the ML (RM) cut-off capability is qualified in the ML rank list (RM rank list).

6. Simulation results

In this section, we discuss the simulation results showing the comparison of the proposed method with the conventional marks based scheme. We used PYTHON as the programming platform for all the analyses. The exams were simulated using candidates of randomly generated known capability values answering questions of randomly generated known difficulties. Then, we used the proposed algorithm for estimating their capability vector C, difficulty vector D, discrimination vector A and guessing vector G from the response matrix R.

Tables and figures in this section demonstrate the simulation results of the conducted experiments. ML here stands for the maximum likelihood based assessment result and RM stands for the raw marks based assessment result. Table I and table II corresponds to the experiment where we fixed the number of questions and varied the number of students for the 10% and 30% cut-off bound respectively. The simulation results affirms that the number of false-positives is less in the proposed scheme when compared to the conventional raw marks based scheme. In addition, the number of candidates qualified is greater in the RM scheme. This is because a greater number of students obtain the same score and thus the number of candidates qualified will be greater than the specified cut-off bound. This results tie in the RM scheme, which need to be resolved. However, the ML scheme does not result in many cases of a tie as this method not only takes into consideration the total score of the candidates but also considers which of the questions they got right. The third parameter, number of desired candidates qualified, is greater for the RM case over the ML case. This is because of the large number of qualified candidates here. We verified that if we allow the same number of candidates to qualify in both the schemes then ML gives the greater number of desired candidates as well.

Table 1: Comparison for nQ = 30 and nS = 200, 500, 1000 and 2000 and cut-off bound = 10%

Metrics and parameters	nQ = 30									
	nS = 200		nS = 500		nS = 1000		nS = 2000		nS = 5000	
	ML	RM	ML	RM	ML	RM	ML	RM	ML	RM
No. of candidates qualified	21	24	51	63	101	157	201	228	501	791
No. of false-positives	13	16	31	40	57	100	124	143	295	522
No. of deserving candidates qualified	8	9	20	23	44	57	77	85	206	269

Table 2: Comparison for nQ = 30 and nS = 200, 500, 1000 and 2000 and cut-off bound = 30%

Metrics and parameters	nQ = 30									
	nS = 200		nS = 500		nS = 1000		nS = 2000		nS = 5000	
	ML	RM	ML	RM	ML	RM	ML	RM	ML	RM
No. of candidates qualified	61	69	151	181	301	359	602	662	1501	1774
No. of false-positives	21	26	57	78	88	130	196	245	438	637
No. of deserving candidates qualified	40	43	94	103	213	229	406	417	1063	1137

Figure 4 and figure 6 indicate the variation of the number of false-positives for the different values of the number of students for the proposed ML scheme and the conventional RM scheme for 10% and 30% cut-off bound respectively. The plot shows that the number of false-positives is less in the proposed scheme when compared to RM scheme.

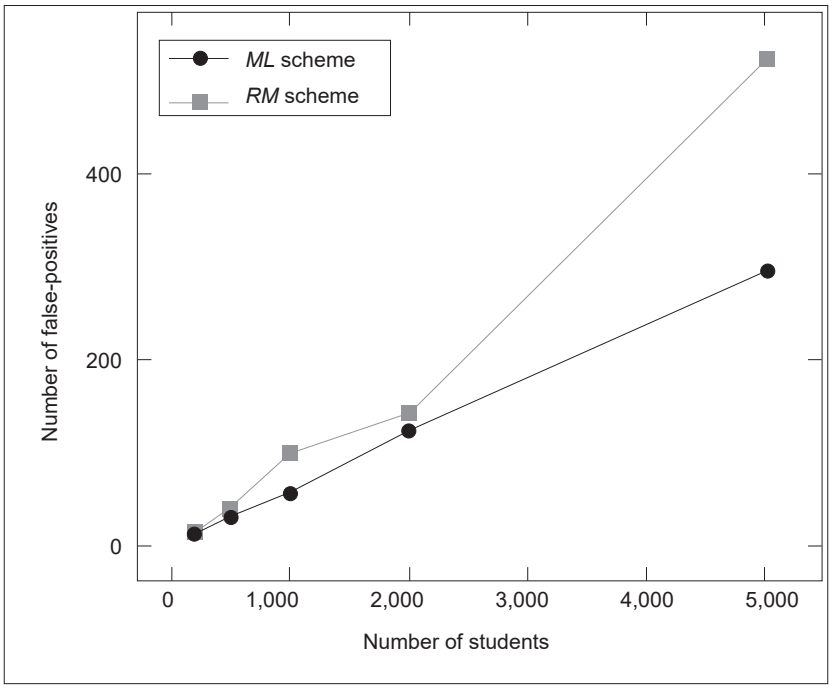

Figure 4: Number of false-positives for nQ = 30 and different nS for 10% cut-off

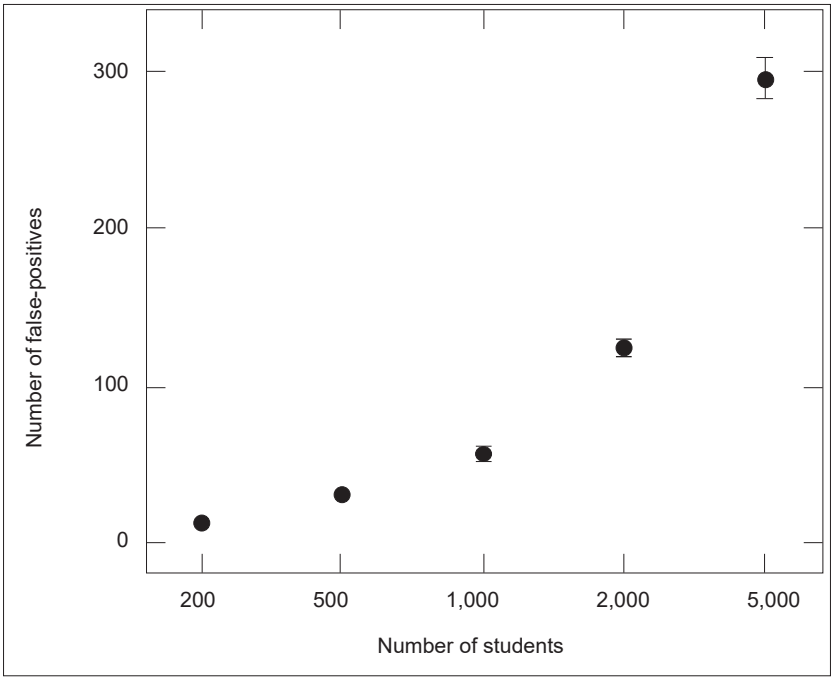

Figure 5: Demonstration of 90% band of false-positives for nQ = 30 and different nS for 10% cut-off

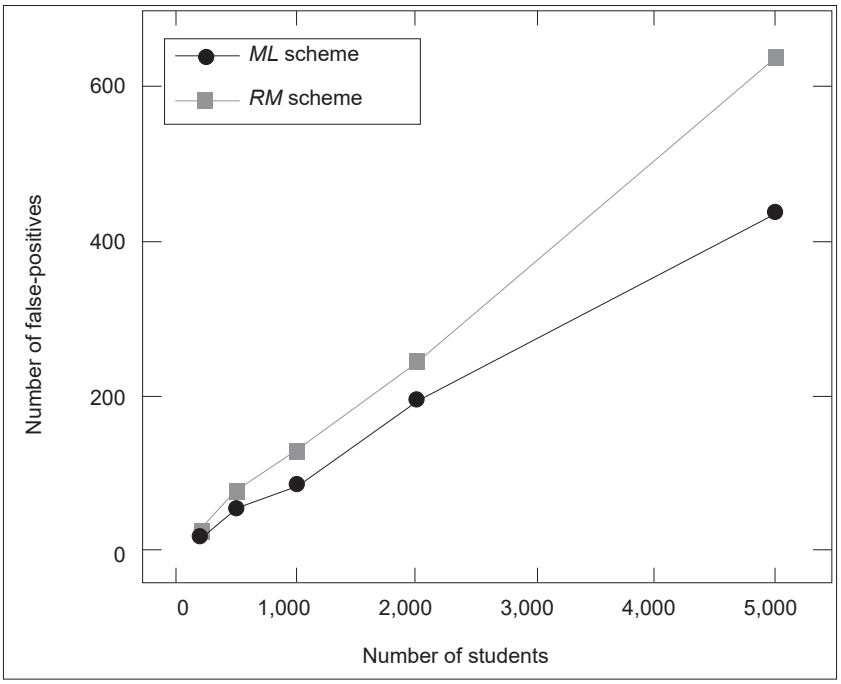

Figure 6: Number of false-positives for nQ = 30 and different nS for 30% cut-off

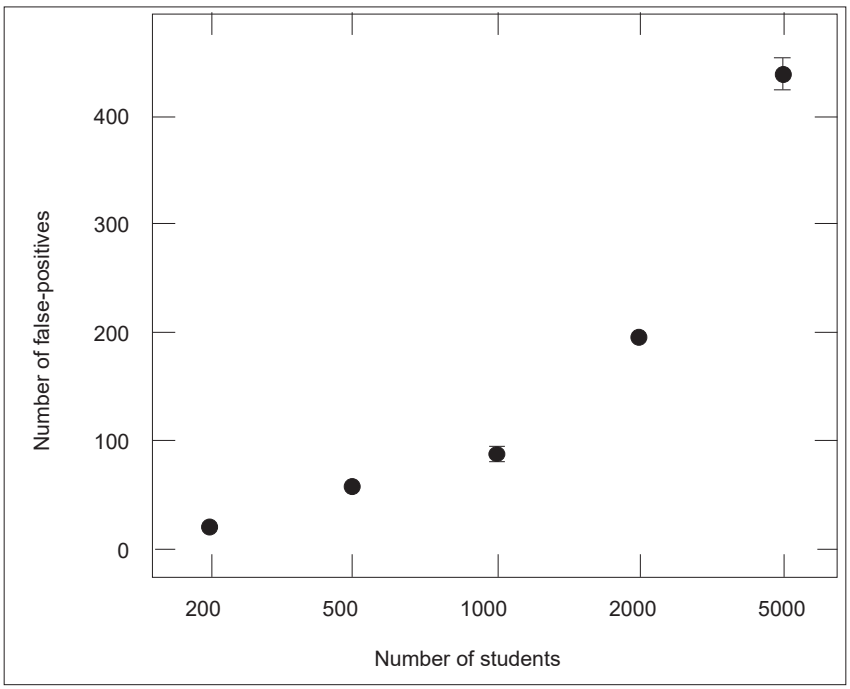

Figure 7: Demonstration of 90% band of gate-crashers for nS = 2000 and different nQ for 30% cut-off

Figure 5 and figure7 demonstrate the 90% band of false-positives for fixed nQ and varied nS for the 10% and 30% cut-off respectively. All the experiments here are done for 50 different exams and all the values in the tables and all the data points in the figures correspond to their average. Thus, this plot is drawn to see the variation of the number of false-positives in the ML scheme for 90% of the exams. This is to verify the spread of the number of false-positives for 90% of the exams conducted. The plot below shows a very narrow band indicating that for 90% of the exams over which this experiment is averaged, the number of false-positives vary in a narrow range.

Table III and table IV show the results corresponding to the experiment where the number of students is fixed and the number of questions is varied. This experiment is conducted to check the performance of the proposed method for exams of different length, more clearly, exams with a different number of items. The results confirm that the proposed method out performs the conventional raw marks based scheme in filtering out the most deserving candidates as the number of false positives is much less in the ML scheme over RM scheme. In addition, the number of ties created is also less in ML method.

Table 3: Comparison for nS = 2000 and nQ = 20, 30, 50 and 70 and cut-off bound = 10%

Metrics and parameters	nS = 2000							
	nQ = 20		nQ = 30		nQ = 50		nQ = 70	
	ML	RM	ML	RM	ML	RM	ML	RM
No. of candidates qualified	201	354	201	354	201	240	201	229
No. of false-positives	129	255	124	143	108	140	108	134
No. of deserving candidates qualified	72	97	77	85	93	100	93	96

Table 4: Comparison for nS = 2000 and nQ = 20, 30, 50 and 70 and cut-off bound = 30%

Metrics and parameters	nS = 2000							
	nQ = 20		nQ = 30		nQ = 50		nQ = 70	
	ML	RM	ML	RM	ML	RM	ML	RM
No. of candidates qualified	601	625	602	662	601	647	601	689
No. of false-positives	236	258	196	245	153	195	157	223
No. of deserving candidates qualified	365	367	406	417	448	452	444	466

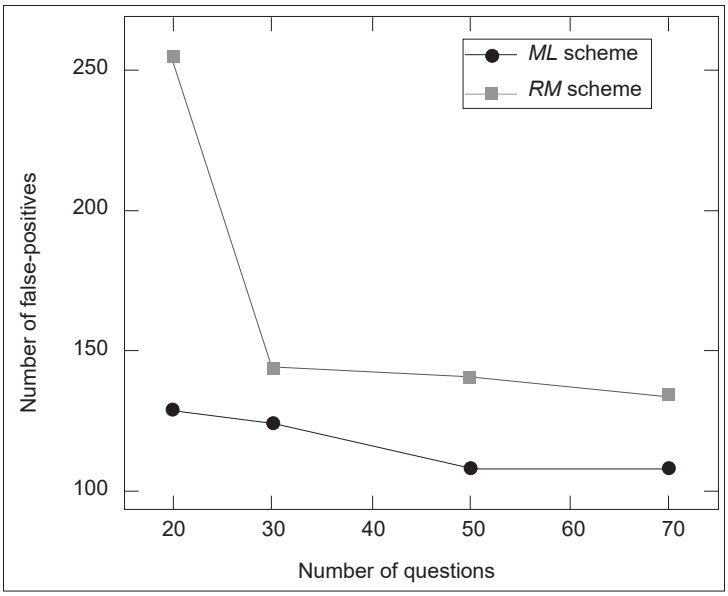

Figure 8: Number of gate-crashers for nS= 2000 and different nQ for 10% cut-off

Figure 8 and figure 10 show the variation of the number of false-positives for different values of the number of questions for the proposed ML scheme and the conventional RM scheme for the 10% and 30% cut-off bound respectively. The plot shows that the number of false-positives is less in the proposed scheme when compared to RM scheme.

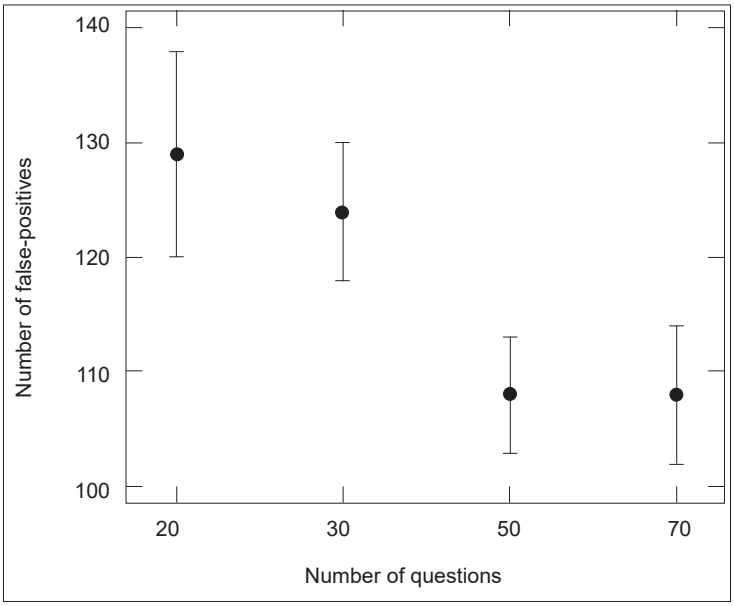

Figure 9: Demonstration of 90% band of gate-crashers for nS= 2000 and different nQ for 10% cut-off

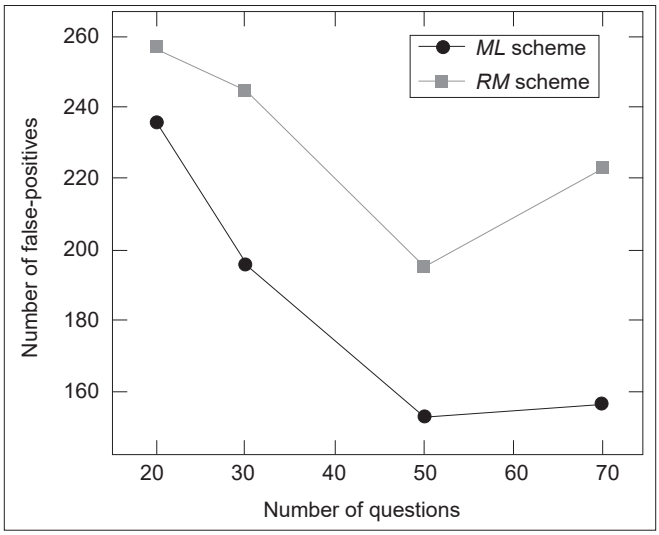

Figure 10: Number of gate-crashers for nS= 2000 and different nQ for 30% cut-off

Figure 9 and figure 11 demonstrate the 90% band of false-positives for fixed nS and varied nQ for the 10% and 30% cut-off respectively. All the experiments here are done for 50 different exams and all the values in the tables and all the data points in the figures correspond to their average. Thus, this plot is drawn to see the variation of the number of false-positives in the ML scheme for 90% of the exams. This is to verify the spread of the number of false-positives for 90% of the exams conducted. The plot below shows a very narrow band indicating that for 90% of the exams over which this experiment is averaged, the number of false-positives vary in a narrow range.

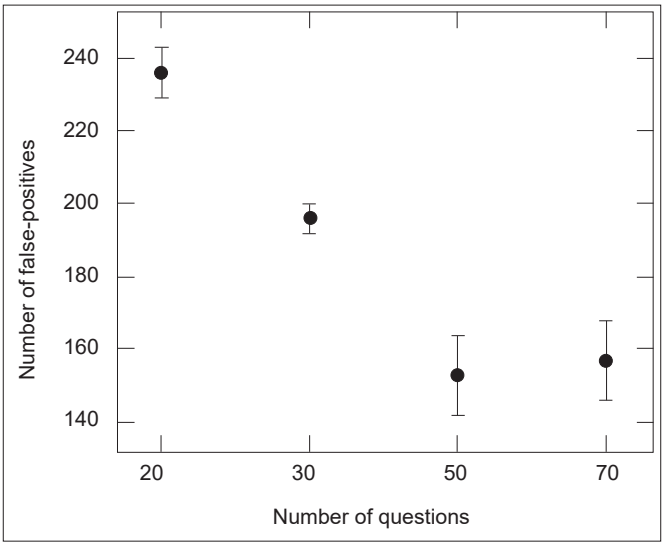

Figure 11: Demonstration of 90% band of gate-crashers for nS = 2000 and different nQ for 30% cut-off

Table V and table VI show the experimental results corresponding to a multiple session exam for the 10% and 30% cut-off. Here, students take exams in four different sessions answering different question papers and finally their scores are normalised so that they are ranked in a single rank list. The normalisation of scores of different sessions is done using the formula below.

$$\acute{m}_{ij} = \frac{m^g_t - m^g_q)}{m_{ti} - m_{iq})} (m_{ij} - m_{iq}) + m^g_q, \quad (8)$$

where m_{ij} is the actual marks obtained by the jth candidate in the ith session, m^g_t is the average marks of the toppers in all sessions, m^g_q is the mean of marks of all students in all sessions, m_{ti} is the top marks of the ith session and m_{iq} is the average marks of all the students in the ith session.

Table 5: Comparison for the multiple session exam for nS = 2000, nQ = 30 done in four sessions with cut-off bound = 10%

Metrics and parameters	nS = 2000 , nQ = 30							
	Session 1		Session 2		Session 3		Session 4	
	ML	RM	ML	RM	ML	RM	ML	RM
No. of candidates qualified	201	259	201	239	201	352	201	324
No. of false-positives	124	174	126	160	131	254	126	224
No. of deserving candidates qualified	77	86	75	79	70	98	75	100

Table 6: Comparison for multiple session exam for nS = 2000, nQ = 30 done in four sessions with cut-off bound = 30%

Metrics and parameters	nS = 2000 , nQ = 30							
	Session 1		Session 2		Session 3		Session 4	
	ML	RM	ML	RM	ML	RM	ML	RM
No. of candidates qualified	601	694	601	688	601	741	601	806
No. of false-positives	206	285	196	267	233	338	205	352
No. of deserving candidates qualified	395	409	405	421	368	403	396	454

7. Conclusion

We proposed a maximum likelihood based alternating maximisation algorithm for estimating student capabilities and question difficulties, discrimination and guessing of an offline exam. The model employed in this paper is the 3-parameter logistic ogive model, which is a well-researched item response model. Experimental tests confirm the improved performance of the proposed scheme over the conventional marks based scheme. Student capabilities were estimated and maximum likelihood estimated capability based rank list (MLC rank list) is compared with the raw marks based rank list (RM rank list). The number of false-positives in the top 10% and 30% is compared for both the rank lists with the actual capability based rank list (AC rank list) and it was found that the number of false-positives in the ML based method

is less for all the experiments. The experiments include varying the number of students and questions and importantly, the multiple session exam with students taking tests in different test centres, answering different question papers for the same discipline. The proposed method is implementable at institutional level as well as for estimating the ability levels of the students instead of using marks as the sole criteria.

References

Baker, F.B. 1985. *The basics of item response theory,* 2nd ed. Wisconsin: Heinemann.

Binet, A. & Simon, T. 1916. *The development of intelligence in children: The Binet-Simon Scale* (No. 11). New Jersey: Williams & Wilkins Company. https://doi.org/10.1037/11069-000

Indian Institute of Science. 2016. Graduate Aptitude Test in Engineering 2016 Information Brochure. India.

Lord, F.M., Novick, M.R. & Birnbaum, A. 1968. *Statistical theories of mental test scores*. Boston: Addison-Wesley.

Moothedath, S. 2016. A maximum likelihood based offline estimation of student capabilities and question difficulties. *Paper presented at the meeting of International Association for Educational Assessment*, Cape Town.

Sands, W.A., Waters, B.K. & McBride, J.R. 1997. *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association. https://doi.org/10.1037/10244-000

Van der Linden, Wim, J. & Glas, C.A. (Eds.) 2000. *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

Way, W.D., Davis, L.L. & Fitzpatrick, S. 2006. *Practical questions in introducing computerized adaptive testing for K-12 assessments.* Upper Saddle River, NJ: Pearson Educational Measurement.

Wu, L.E.E.T. 1993. Japanese University Entrance Examination Problems in Mathematics. *American Educator: The Professional Journal of the American Federation of Teachers*, 17(1), 25-35.