**Evaluation of Computer Adaptive Testing (CAT) in onlinelearning platforms: A performance engineering approach to CAT using simulation**

Bahram Saleh Sedghpour

[sedghpour@sru.ac.ir](mailto:sedghpour@sru.ac.ir)

School of Humanities, Shahid Rajaee

Teacher Training University

Tehran, Iran

Mohammad Reza Saleh Sedghpour

[msaleh@cs.umu.se](mailto:msaleh@cs.umu.se)

Department of Computing Science,

Umeå University

Umeå, Sweden

**Abstract:**

The social distancing in the age of *COVID-19 pandemic* by the governments has a significant impact on education industry. Almost all educational institutes migrated to employ online learning platforms as the main medium to educate their audiences. The problem with such migration becomes acute when we intend to perform the assessment. Converting the traditional paper-and-pencil testing to the online version may cause chaotic situations; in a way that we may no longer be able to certify the learner's ability. Computer adaptive testing aims to handle the issues and facilitate more accurate estimations of examinee's ability. Most of the institutes both in private and public sectors around the world are using computer adaptive testing. A computer adaptive testing features various settings with different possible configurations. However, there have been no systematic studies on the effects of these settings on test performance and robustness. Furthermore, the exact impact of various parameters for computer adaptive testings are poorly understood. This work presents a large set of experiments conducted to investigate these issues using a simulator. Our experiments reveal effective settings and configurations of computer adaptive testing. The findings presented will be useful to institutions seeking to design and implement computer adaptive testing more systematically and also open up new areas of research for academics in the area of computer adaptive testings.

**Keywords:** Item response theory, Computer adaptive testing, Simulation

**1- Introduction**

Recently, computer-based testing (CBT) attracted considerable attention both in research community and education industry and it is increasingly being implemented across the world. With the beginning of COVID-19 pandemic, almost all of the educational institutes migrated to online learning systems, the need for such testing is felt more than ever. CBT offers many advantages over traditional paper-and-pencil testing [14]. The CBT itself is not a new idea and there are several old studies which dealt with CBT and the application of latent trait models to item-bank construction, item selection, and computer adaptive testing (CAT) [2]. The general measurement profession had been working with CBT and, more specifically, with CAT since the early 70s. Creation of item response theory (IRT) and Rasch models led to advancements of CAT [6]. The dichotomous Rasch model presents that each test-taker is characterized by an ability level expressed as a number along an infinite linear scale of the relevant ability. The test-taker *n*'s ability is recognized as being units from that local origin. Similarly, each item is identified by a difficulty level also expressed as a number along the infinite scale of the relevant ability. The difficulty of item *i* is identified as being units from the local origin of the ability scale. This relationship between test-takers and items is expressed by the dichotomous Rasch model [12]:

where is the probability that test-taker n succeeds on item *i*, and is the probability of failure. The natural unit of the interval scale constructed by this model is termed the logit (log-odds unit). The logit distance along the unidimensional measurement scale between a test-taker expected to have 50% success on an item, (i.e., at the person at same position along the scale as the item,) and a test-taker expected to have 75% success on that same item is log(75% / 25%) = 1.1 logits. Based on this model, there are many studies which developed a CAT algorithm such as [1].

Despite the considerable interest in computer adaptive testing, simulation and their potential benefits for standard testing, there is a lack of simulation studies on various aspects that affects CAT such as item banks, examinee's ability and different parameter settings. Some of them may have little impact on standard test and could be ignored in some use cases. However, in other scenarios, improper parameter setting could invalidate a standardized test. To clarify the impact of test-takers ability and item banks as well as the effects of varying the parameter settings of CAT, this work seeks to answer three research questions:

**RQ1: How does item bank size relate to the number of examinees?**

**RQ2: How do different parameters of logistic model affect the ability of test?**

**RQ3: How does examinee's ability affect the performance of test?**

To answer these questions, we present a set of experiments performed using CAT simulator. In these experiments, the simulation is performed using CATsim[10] and a range of different CAT scenarios and parameter settings are tested to evaluate their impact on tests ability and illustrate the challenges of designing and implementing a CAT platform. The results obtained provide new insights into the practical use of CAT settings and how such settings interact to enhance the platform ability.

**2- Background**

In this section, we discuss previous studies relevant to the work presented herein. We divide these earlier studies into works concerning CAT and IRT studies.

**2-1- Computer adaptive testing (CAT)**

Unlike linear tests, in which items are sequentially presented to test takers and their ability estimated at the end of the test, in a computerized adaptive test (CAT), an examinees’ ability is updated after the response of each item. The updated knowledge of an examinee’s ability at each step of the test allows for the selection of more informative items during the test itself, which in turn reduce the standard error of estimation of their ability at a faster rate.

**2-1-1- The CAT lifecycle**

In general, a computerized adaptive test has a very well-defined lifecycle as show in Fig. 1. As the first step, the examinee’s initial ability is estimated, then an item is selected based on the current ability estimation, the ability is reestimated based on the answers to all items up until now as the next step and finally if a stopping criterion is met, the test will stop and if not the procedure will be repeated from second step.

**2-2- Item response theory (IRT)**

A test identifies the state of each student in the subject. The fact that a student achieves average performance in a subject does not indicate that the student has learned half of the content, possibly learned one topic very well, and almost nothing in other related topics. Various methods, processes, approaches, and tools have been developed to support teachers and students to improve the assessment process. Among the different techniques and theories associated with these studies, Item Response Theory has gained projection for its effectiveness. There are different types of logistic models in IRT; one, two and three parameters logistic models, a series of models in which examinees and items are represented by a set of numerical values (the models’ parameters) [7]. Various studies showed the benefit of using IRT for testing in different subjects such as cognitive ability testing [13], developing questionnaires [4], [15, 16] and language testing [9]. The list of IRT usage is not ended here.



Figure 1: CAT lifecycle.

The logistic models of Item Response Theory are unidimensional, which means that a given assessment instrument only measures a single ability (or dimension of knowledge). The instrument, in turn, is composed of items in which examinees manifest their latent traits when answering them [5]. In unidimensional IRT models, an examinee’s ability is represented as θ. Since the scale of θ is up to the individuals creating the instrument, it is common for the values to be around the normal distribution.

Under the logistic models of IRT, an item is represented by the following parameters:

1. a represents an item’s discrimination parameter, that is, how well it discriminates individuals who answer the item correctly (or, in an alternative interpretation, individuals who agree with the idea of the item) and those who don’t.
2. b represents an item’s difficulty parameter. This parameter, which is measured in the same scale as θ , shows at which point of the ability scale an item is more informative, that is, where it discriminates the individuals who agree and those who disagree with the item.
3. c represents an item’s pseudo-guessing parameter. This parameter denotes what is the probability of individuals with low ability values to still answer the item correctly.
4. d represents an item’s upper asymptote. This parameter denotes what is the probability of individuals with high ability values to still answer the item incorrectly.

For a set of items , when , the three-parameter logistic model is reduced to the two-parameter logistic model. Additionally, if all values of are equal, the two-parameter logistic model is reduced to the one-parameter logistic model. Finally, when , we have the Rasch model [11].

Under IRT, the probability of an examinee with a given value to answer item correctly, given the item parameters, is given by [[3,](#page12) 8]:

The information this item gives is computed as [3, 8]:

The sum of the information of all items in a test is called test information [3]:

The amount of error in the estimate of an examinee’s ability after a test is called the standard error of estimation [3] and it is given by:

Since the denominator in the calculation of the is , it is clear to see that the more items an examinee answers, the smaller gets.

Despite the recent academic interest in CATs, few studies have investigated the impact of different parameter settings. This study differs from previous works on CATs and IRT in that it is based on a systematic analysis of different test parameter settings using a simulator.

**3- Approach**

In this section we introduce our approach to data collection and analysis. We then discuss factors that could potentially reduce the validity of our experiments and the measures taken to ensure validity.

**3-1- Data collection**

To collect data for our study, we developed a tool that repeatedly configures the selected test scenarios in CATsim. To do this, we identified four test parameters in CATsim:

(1) Bank size

(2) Examinee population

(3) Item type representing different logisitc models

(4) Initializer representing the ability of examinees which can be fixed or random.

We deployed the tool on a bare-metal machine with 16GB of RAM, two Intel Xeon E5430 2.66 GHz CPUs with four cores and hyper-threading, and a 256 GB NVMe drive running Ubuntu 20.04 LTS. The Python version is 3.8.11. We repeated each experiment 10 times. In total, we collected data from 370 different experiments and taking over 180 hours to complete in total.

**3-2- Analysis**

Our initial goal in the analysis was to determine how strongly the selection of different parameter values impacted the performance of a test. To this end, we performed visual analysis of mean square error (MSE), bias and overlap of test items.

For each configuration of test in the CATsim, we plot mean square error (MSE), bias and overlap of each test to compare the results. We then iteratively discuss and classify the resulting plots.



Figure 2: Relation between bank size and number of examinee: The relation is described in terms of Bias and Overlap ratio for different item bank sizes; 100 items (blue lines), 1000 items (orange lines), and 10000 items (green lines).

**3-3- Threats to validity**

Despite careful research design, studies such as those presented here inevitably have limitations and factors that may reduce their validity. The most important of these limitations and factors for this work are summarized below.

**3-3-1- External validity**

This paper used only a specific version of each tool. While we argue that the chosen tool versions are representative of tools that would be used test beds, the findings presented herein cannot be directly generalized to other versions, particularly since performance may differ between versions even if the same configuration policies are used. To summarize, while we have no direct evidence that our results are applicable to the studied tools in all cases, we expect that similar results would be obtained if different variants of the chosen tools were used.

**3-3-2- Internal validity**

Internal validity is inevitably affected by the fact that some design decisions must be made when defining the configuration values to test. Empirical data analysis has not suggested that the whole stack's behavior would have been radically different if values other than those chosen were used, but this is clearly impossible to prove. Another internal validity threat is that we performed all experiments on bare-metal platforms. Some of the executions in our study could thus have been affected by our choice of hardware for data collection. However, we consider it unlikely that the general validity of our results would be threatened by the specific hardware chosen for the study. It seems unlikely that the general validity of our results would be impacted by the use of a bare-metal platform, however. Another internal validity threat is that there are infinite potential test scenarios, which could impact some of our experiments. However, we tested a wide range of different test scenarios to maximize the breadth of the conditions tested in our experiments and improve the robustness and reliability of our conclusions.



Figure 3: The impact of number of parameters in logistic model considering the number of examinee: The relation is described in terms of Mean Squared Error(MSE), Bias and Overlap rate for various numbers of parameters in the logistic model; 1 parameter (blue lines), 2 parameters (orange lines), 3 parameters (green lines) and 4 parameters (red lines).

**4- Results and Discussion**

To study the impact different test parameters (see section 3), we performed a series of experiments in which we compared the mean suqare error, bias and overlap rate observed with different test parameters. We repeated each experiment 10 times but since there were no significant differences between the results of replicate experiments we only discuss the results of individual experiments. We performed t-tests to evaluate the significance of differences between results obtained under different conditions, applying a significance threshold of p < .01.

**4-1- Item bank size and number of examinee**

To investigate the impact of item bank size and number of examinees, we conducted a series of experiments in which the test size was fixed to 20 items, three parameters selected as logistic model, the examinee’s abilities were randomly chosen, items were selected using maximum info method and the ability of examinees were estimated using numerical search method while varying the item bank size (100, 1000, 10000) and number of examinees related to the item bank size (from 1% to 100% of item bank size). The bias and the overlap rate of the experiments is shown in Fig. 2. Based on the results, the more larger the item bank is, the more steady behavior in terms of bias and overlap ratio is achieved. On the other hand, with increase of item bank size, the overlap ratio decreases. Another observation is that, when the item bank size are smaller and the number of examinees are less than 10% of item bank size, there are negative test bias.



Figure 4: The impact of examinee’s ability with normal distribution considering the number of examinee: The relation is described in terms of Overlap ratio for different standard deviations (columns) and different averages (rows)

**4-2- Number of parameters in logistic model**

To study the impact of number of parameters in logistic model considering number of examinees, we performed a series of experiments in which the test size was fixed to 20 items out of 1000 items, the examinee’s abilities were randomly chosen, items were selected using maximum info method and the ability of examinees were estimated using numerical search method while varying the number of logistic model parameters (1, 2, 3 and 4) and number of examinees related to the item bank size (from 1% to 100% of item bank size). The mean squared error, bias and the overlap rate of the experiments is shown in Fig. 3. With considering the meaning of each parameter as discussed in section 2.2, based on the results, although the minimum overlap rate belongs to the model with 1 parameter, the highest mean squared error belongs to same model as well. The model with 2 parameters had the minimum mean squared error while it behaves like other model with different parameters in terms of overlap rate and bias.

**4-3- Impact of examinee’s ability**

To analyze the impact of examinee’s ability considering the number of examinees in relation to item bank size, we conducted a series of experiments in which the test size was fixed to 20 items out of 1000 items, items were selected using maximum info method, the ability of examinees were estimated using numerical search method, and 4 parameters in logistic model were used while varying the distribution configuration of examinee’s ability; such as normal distribution with different values for standard deviation (0.1, 1, 2 and 5) and different values for average (-10, -2, 0, 2, 10), uniform distribution with set of minimum and maximum tuples ((-1, 1), (5, 10), (-10,-5),(200, 250) and (-250, -200)), fixed ability with set of defined abilities ( -200, -10, 0, 10 and 200) and number of examinees related to the item bank size (from 1% to 100% of item bank size). The mean squared error of items and bias for normal distribution configurations didn’t show any specific changes, but there are some slight changes in overlap rate which is shown in Fig. 4.. Based on the results, when there is higher standard deviation (more different abilities), the overlap rate is decreasing slightly, which means administered items are less likely to be shared between examaninee’s. On the other hand, the more average ability diverges to zero, the more overlap there is, which means if the average ability is more far to zero, then the administered items are more likely to be shared between examinees’.

Conversely, in uniform distribution, the more distance between the defined range and zero ability, the higher overlap rate is observed and the bias and mean squared error are not affected by the ability differences which is shown in Fig. 5. Finally, if the examinee’s ability defined at the same point, the more or less the abilities are, the higher overlap rate is observed as shown in Fig. 6, the bias and mean squared error are not affected by the ability differences.



Figure 5: The impact of examinee’s ability with uniform distribution considering the number of examinee: The relation is described in terms of Mean Squared Error(MSE), Bias and Overlap rate for different minimum and maximum value for the uniform distribution in tuple format; (-1, -1) (blue lines), (5, 10) (orange line), (-10,-5) (green line), (200,250) (red line) and (-250, -200) (magenta line)

**4-4- Discussion**

RQ1 asks "How does item bank size relate to the number of examinees?"

The impact of item bank size and number of examinees are presented in Section 4.1, which discusses different item bank sizes and different number of examinees.

Our experiments show that a large item bank is able to estimate the ability of examinee steadily with lower overlap rate.

RQ2 asks "How do different parameters of logistic model affect the ability of test?"

In section 4.2, we presented the impact of different parameters in logistic model and how they interact with the test performance.

Our experiments show that with two parameters, the CAT will achieve minimum mean squared error and considerable overlap rate and bias which does not differ from other selection. On the other hand, the minimum overlap rate is achieved with 1 parameter, while getting highest mean squared error. The 3 and 4 parameters can enhance the situation when the audience of CAT consists persons with highest or lowest ability.



Figure 6: The impact of examinee’s ability with fixed ability considering the number of examinee: The relation is described in terms of Mean Squared Error(MSE), Bias and Overlap rate for different minimum and maximum value for the fixed abilities; -200 (blue lines), -10 (orange lines), 0 (green lines), 10 (red lines) and 200 (magenta lines)

Finally, RQ3 asks"How does examinee’s ability affect the performance of test?"

We investigated the impact of examinee’s ability in Section 4.3, in which we first considered normal distribution with various settings and then uniform distribution with different configuration and lastly the fixed abilities for the examinee’s.

Our experiments show that different examinee’s ability does not play a role in both test bias or mean squared error of items. One of the common observation in all situation is that when we have higher/lower abilities in comparison with zero, we will get slightly higher overlap rate.

**5- Outlook**

This paper studies different aspects of computer adaptive testing to provide guidance on the practical use of this method and to show how they can increase test performance. The computer adaptive testing landscape is rapidly evolving. Because test performance depends on many factors, implementing computer adaptive testing well can be challenging. The use of a computer adaptive testings enables outstanding estimation of examinee’s ability without imposing any particular implementation costs in long term, which suggests that it may be beneficial to develop such methods for online learning platforms. We therefore propose to build on the results presented herein by studying deeper for different stopping criterion and item selection methods.

**6- References**

[1] Wright BD. 1988. Practical Adaptive Testing CAT Algorithm. , 24 pages.

[2] Micheline Chalhoub-Deville. 2001. Language testing and technology: Past and future. Language learning & technology 5, 2 (2001), 95–98.

[3] RJ de Ayala. 2009. The theory and practice of item response theory. (2009).

[4] Maria Orlando Edelen and Bryce B Reeve. 2007. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of life research 16, 1 (2007), 5–18.

[5] Susan E Embretson and Steven P Reise. 2013. Item response theory. Psychology Press.

[6] Richard C Gershon. 2005. Computer adaptive testing. Journal of applied measurement 6, 1 (2005), 109–127.

[7] Frederic M Lord and Melvin R Novick. 2008. Statistical theories of mental test scores. IAP.

[8] David Magis. 2013. A Note on the Item Information Function of the Four-Parameter Logistic Model. Applied Psychological Measurement 37, 4 (2013), 304–315. https://doi.org/10.1177/0146621613475471 arXiv:https://doi.org/10.1177/0146621613475471

[9] TF McNamara. 1991. Test dimensionality: IRT analysis of an ESP listening test1. Language testing 8, 2 (1991), 139–159.

[10] Douglas De Rizzo Meneghetti and Plinio Thomaz Aquino Junior. 2018. Application and Simulation of Computerized Adaptive Tests Through the Package catsim. arXiv:1707.03012 [stat.AP]

[11] G. Rasch. 1966. AN ITEM ANALYSIS WHICH TAKES INDIVIDUAL DIFFERENCES INTO ACCOUNT. Brit. J. Math. Statist. Psych. 19, 1 (1966), 49–57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8317.1966.tb00354.x

[12] Georg Rasch. 1993. Probabilistic models for some intelligence and attainment tests. ERIC.

[13] Steven P Reise, Han Du, Emily F Wong, Anne S Hubbard, and Mark G Haviland. 2021. Matching IRT models to patient-reported outcomes constructs: The graded response and log-logistic models for scaling depression. psychometrika 86, 3 (2021), 800–824.

[14] Michael Russell, Amie Goldberg, and Kathleen O’connor. 2003. Computer-based testing and validity: A look back into the future. Assessment in education: principles, policy & practice 10, 3 (2003), 279–293.

[15] Maomi Ueno and Yoshimitsu Miyazawa. 2017. IRT-based adaptive hints to scaffold learning in programming. IEEE Transactions on Learning Technologies 11, 4 (2017), 415–428.

[16] Yehiry Lucelly Pulido Vega, Gloria Milena Fernandez Nieto, Silvia Margatira Baldiris, and Juan Carlos Guevara Guevara Bolaños. 2012. Application of item response theory (IRT) for the generation of adaptive assessments in an introductory course on object-oriented programming. In 2012 Frontiers in Education Conference Proceedings. IEEE, 1–4.