Acta Univ. Sapientiae, Informatica, **2,** 2 (2010) 168–183

Computerized adaptive testing: implementation issues

Margit Antal
Sapientia Hungarian University of Transylvania, Department of Mathematics and Informatics
Tîrgu Mureş
email: manyi@ms.sapientia.ro

Levente Erős
Sapientia Hungarian University of Transylvania, Department of Mathematics and Informatics
Tîrgu Mureş
email: ideges@gmail.com

Attila Imre
Sapientia Hungarian University of Transylvania, Department of Human Sciences
Tîrgu Mureş
email: imatex@ms.sapientia.ro

Abstract. One of the fastest evolving field among teaching and learning research is students' performance evaluation. Computer based testing systems are increasingly adopted by universities. However, the implementation and maintenance of such a system and its underlying item bank is a challenge for an inexperienced tutor. Therefore, this paper discusses the advantages and disadvantages of Computer Adaptive Test (CAT) systems compared to Computer Based Test systems. Furthermore, a few item selection strategies are compared in order to overcome the item exposure drawback of such systems. The paper also presents our CAT system along its development steps. Besides, an item difficulty estimation technique is presented based on data taken from our self-assessment system.

Computing Classification System 1998: K.3.1
Mathematics Subject Classification 2010: 97U50
Key words and phrases: computerized adaptive testing, item response theory

1 Introduction

One of the fastest evolving field among teaching and learning research is students' performance evaluation. Web-based educational systems with integrated computer based testing are the easiest way of performance evaluation, so they are increasingly adopted by universities [3, 4, 9]. With the rapid growth of computer communications technologies, online testing is becoming more and more common. Moreover, limitless opportunities of computers will cause the disappearance of Paper and Pencil (PP) tests. Computer administered tests present multiple advantages compared to PP tests. First of all, various multimedia can be attached to test items, which is almost impossible in PP tests. Secondly, test evaluation is instantaneous. Moreover, computerized self-assessment systems can offer various hints, which help students' exam preparation.

This paper is structured in more sections. Section 2 presents Item Response Theory (IRT) and discusses the advantages and disadvantages of adaptive test systems. Section 3 is dedicated to the implementation issues. The presentation of the item bank is followed by simulations for item selection strategies in order to overcome the item exposure drawback. Then the architecture of our web-based CAT system is presented, which is followed by a proposal for item difficulty estimation. Finally, we present further research directions and give our conclusions.

2 Item Response Theory

Computerized test systems reveal new testing opportunities. One of them is the adaptive item selection tailored to the examinee's ability level, which is estimated iteratively through the answered test items. Adaptive test administration consists in the following steps: (i) start from an initial ability level, (ii) selection of the most appropriate test item and (iii) based on the examinee's answer re-estimation of their ability level. The last two steps are repeated until some ending conditions are satisfied. Adaptive testing research started in 1952 when Lord made an important observation: ability scores are test independent whereas observed scores are test dependent [6]. The next milestone was in 1960 when George Rasch described a few item response models in his book [11]. One of the described models, the one-parameter logistic model, became known as the Rasch model. The next decades brought many new applications based on Item Response Theory.

In the following we present the three-parameter logistic model. The basic

component of this model is the item characteristic function:

$$P(\Theta) = c + \frac{(1-c)}{1+e^{-Da(\Theta-b)}}, \tag{1}$$

where Θ stands for the examinee's ability, whose theoretical range is from $-\infty$ to ∞, but practically the range -3 to $+3$ is used. The three parameters are: a, discrimination; b, difficulty; c, guessing. Discrimination determines how well an item differentiates students near an ability level. Difficulty shows the place of an item along the ability scale, and guessing represents the probability of guessing the correct answer of the item [2]. Therefore guessing for a true/false item is always 0.5. $P(\Theta)$ is the probability of a correct response to the item as a function of ability level [6]. D is a scaling factor and typically the value 1.7 is used.

Figure 1 shows item response function for an item having parameters $a = 1, \quad b = 0.5, \quad c = 0.1$. For a deeper understanding of the discrimination parameter, see Figure 2, which illustrates three different items with the same difficulty ($b = 0.5$) and guessing ($c = 0.1$) but different discrimination parameters. The steepest curve corresponds to the highest discrimination ($a = 2.8$), and in the middle of the curve the probability of correct answer changes very rapidly as ability increases [2].

The one- and two-parameter logistic models can be obtained from equation (1), for example setting $c = 0$ results in the two-parameter model, while setting $c = 0$ and $a = 1$ gives us the one-parameter model.

Compared to the classical test theory, it is easy to realize the benefits of the former, which is able to propose the most appropriate item, based on item statistics reported on the same scale as ability [6].

Another component of the IRT model is the item information function, which shows the contribution of a particular item to the assessment of ability [6]. Item information functions are usually bell shaped functions, and in this paper we used the following (recommended in [12]):

$$I_i(\Theta) = \frac{P_i'(\Theta)^2}{P_i(\Theta)(1-P_i(\Theta))}, \tag{2}$$

where $P_i(\Theta)$ is the probability of a correct response to item i computed by equation (1), $P_i'(\Theta)$ is the first derivative of $P_i(\Theta)$, and $I_i(\Theta)$ is the item information function for item i.

High discriminating power items are the best choice as shown in Figure 3, which illustrates the item information functions for the three items shown in

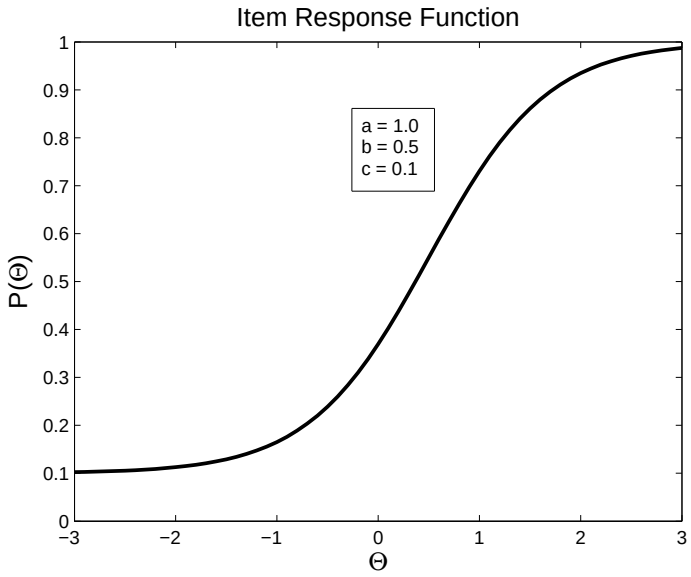

Figure 1: A three-parameter logistic model item characteristic function

Figure 2. All three functions are centered around the ability $\Theta = 0.5$, which is the same as the item difficulty.

Test information function I_{rr} is defined as the sum of item information functions. Two such functions are shown for a 20-item test selected by our adaptive test system: one for a high ability student (Figure 4) and another for a low ability student (Figure 5). The test shown in Figure 4 estimates students' ability near $\Theta = 2.0$, while the test in Figure 5 at $\Theta = -2.0$.

Test information function is also used for ability estimation error computation as shown in the following equation:

$$SE(\Theta) = \frac{1}{\sqrt{I_{rr}}}. \tag{3}$$

This error is associated with maximum likelihood ability estimation and is usually used for the stopping condition of adaptive testing.

For learner proficiency estimation Lord proposes an iterative approach [10], which is a modified version of the Newton-Raphson iterative method for solving equations. This approach starts with an initial ability estimate (usually a random value). After each item the ability is adjusted based on the response

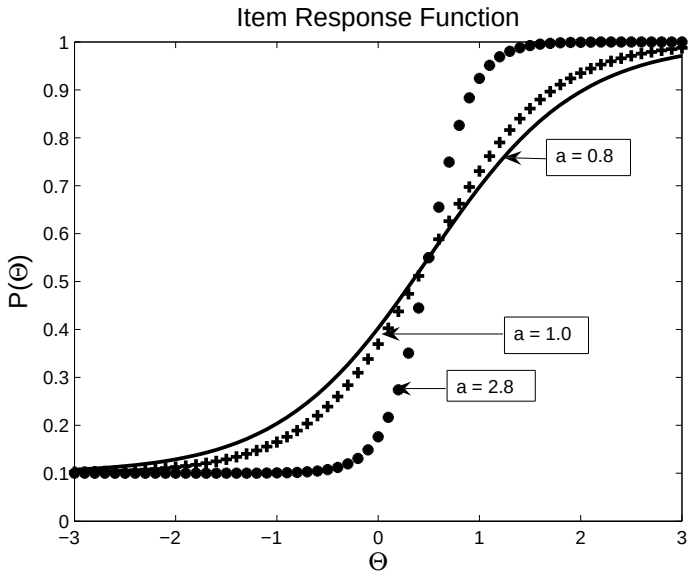

Figure 2: Item characteristic functions

given by the examinee. For example, after n questions the estimation is made according to the following equation:

$$\Theta_{n+1} = \Theta_n + \frac{\sum_{i=1}^{n} S_i(\Theta_n)}{\sum_{i=1}^{n} I_i(\Theta_n)} \tag{4}$$

where $S_i(\Theta)$ is computed using the following equation:

$$S_i(\Theta) = (u_i - P_i(\Theta)) \frac{P_i'(\Theta)}{P_i(\Theta)(1 - P_i(\Theta))}. \tag{5}$$

In equation (5) u_i represents the correctness of the ith answer, which is 0 for incorrect and 1 for correct answer. $P_i(\Theta)$ is the probability of correct answer for the ith item having the ability Θ (equation (2)), and $P_i'(\Theta)$ is its first derivative.

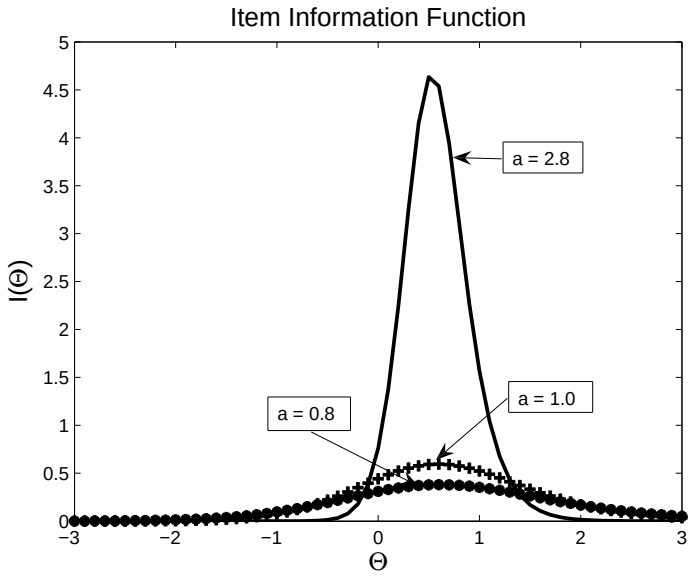

Figure 3: Item information functions

2.1 Advantages

In adaptive testing the best test item is selected at each step: the item having maximum information at the current estimate of the examinee's proficiency. The most important advantage of this method is that high ability level test-takers are not bored with easy test items, while low ability ones are not faced with difficult test items. A consequence of adapting the test to the examinee's ability level is that the same measurement precision can be realized with fewer test items.

2.2 Disadvantages

Along with the advantages offered by IRT, there are some drawbacks as well. The first drawback is the impossibility to estimate the ability in case of all correct or zero correct responses. These are the cases of either very high or very low ability students. In such cases the test item administration must be stopped after administering a minimum number of questions.

The second drawback is that the basic IRT algorithm is not aware of the test content, the question selection strategy does not take into consideration

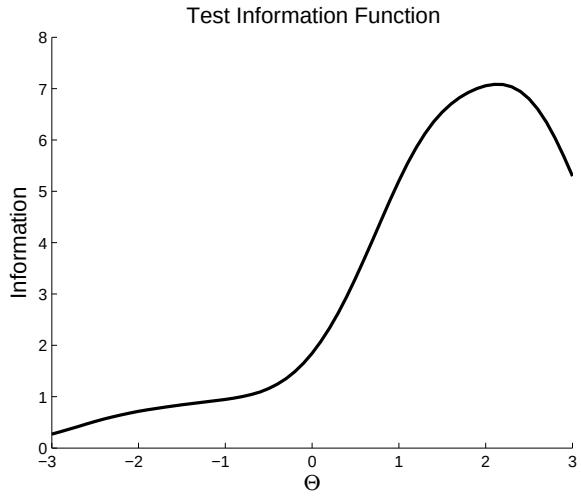

Figure 4: Test information function for a 20-item test generated for high ability students

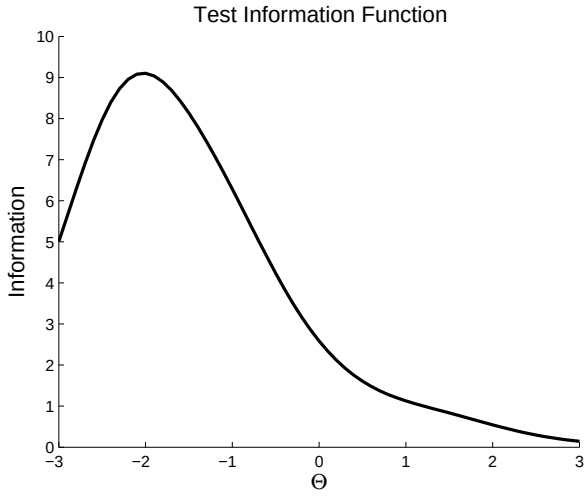

Figure 5: Test information function for a 20-item test generated for low ability students

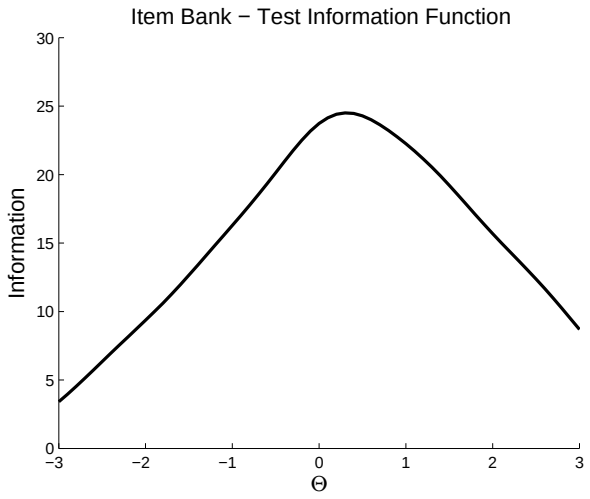

Figure 6: Test information function for all test items

to which topic a question belongs. However, sometimes this may be a requirement for generating tests assessing certain topics in a given curriculum. Huang proposed a content-balanced adaptive testing algorithm [7]. Another solution to the content balancing problem is the testlet approach proposed by Wainer and Kiely [15]. A testlet is a group of items from a single curriculum topic, which is developed as a unit. If an adaptive algorithm selects a testlet, then all the items belonging to that testlet will be presented to the examinee.

The third drawback, which is also the major one, is that IRT algorithms require serious item calibration. Despite the fact that the first calibration method was proposed by Alan Birnbaum in 1968 and has been implemented in computer programs such as BICAL (Wright and Mead, 1976) and LOGIST (Wingersky, Barton and Lord, 1982), the technique needs real measurement data in order to accurately estimate the parameters of the items. However, real measurement data are not always available for small educational institutions.

The fourth drawback is that several items from the item bank will be overexposed, while other test items will not be used at all. This requires item exposure control strategies. A good review of these strategies can be found in [5], discussing the strengths and weaknesses of each strategy. Stocking [13] made one of the first overviews of item exposure control strategies and classified them in two groups: (i) methods using a random component along the

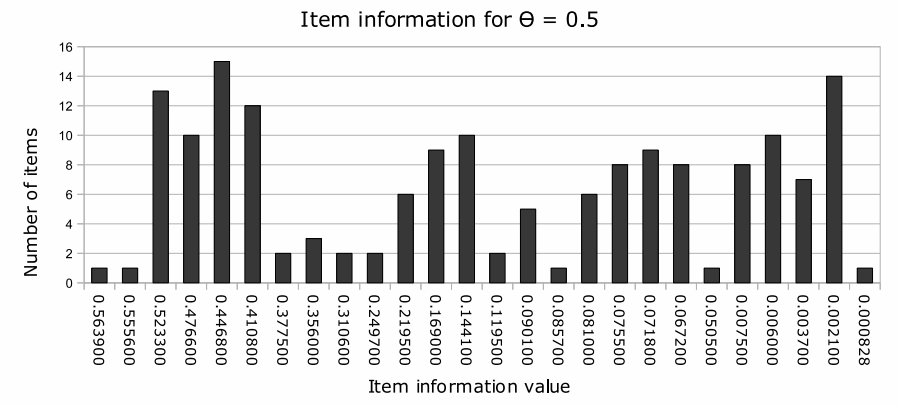

Figure 7: Item information clusters and their size

item selection method and (ii) methods using a parameter for each item to control its exposure. Randomization strategies control the frequency of item administration by selecting the next item from a group of items (e.g. out of the 5 best items). The second item exposure control strategy uses an exposure control parameter. In case of an item selection—due to its maximum information for the examinee's ability level—, the item will be administered only if its exposure control parameter allows it.

3 CAT implementation

3.1 The item bank

We have used our own item bank from our traditional computer based test system "Intelligent" [1]. The item bank parameters (a - discrimination, b - difficulty, c - pseudo guessing) were initialized by the tutor. We used 5 levels of difficulty from very easy to very difficult, which were scaled to the [-3,3] interval. The guessing parameter of an item was initialized by the ratio of the number of possible correct answers to the total number of possible answers. For example, it is 0.1 for an item having two correct answers out of five possible answers. Discrimination is difficult to set even for a tutor, therefore we used $a = 1$ for each item.

3.2 Simulations

In our implementation we have tried to overcome the disadvantages of IRT. We started to administer items adaptively only after the first five items. Ability (Θ) was initialized based on the number of correct answers given to these five items, which had been selected to include all levels of difficulty.

We used randomization strategies to overcome item exposure. Two randomization strategies were simulated. In the first one we selected the top ten items, i.e. the ten items having the highest item information. However, this is better than choosing the single best item, thus one must pay attention to the selection of the top ten items. There may be more items having the same item information for a given ability, therefore it is not always the best strategy choosing the first best item from a set of items with the same item information. To overcome this problem, in the second randomization strategy we computed the item information for all items that were not presented to the examinee and clustered the items having the same item information. The top ten items were selected using the items from these clusters. If the best cluster had less than ten items, the remainder items were selected from the next best cluster. If the best cluster had more than ten items, the ten items were selected randomly from the best cluster. For example, Figure 7 shows the 26 clusters of item information values constructed from 171 items for the ability of $\Theta = 0.5$. The best 10 items were selected by taking the items from the first two clusters (each having exactly 1 item) and selecting randomly another 8 items out of 13 from the third cluster.

Figure 8 shows the results from a simulation where we used an item bank with 171 items (test information function is shown in Figure 6 for all the 171 items), and we simulated 100 examinees using three item selection strategies: (i) best item (ii) random selection from the 10 best items (iii) random selection from the 10 best items and clustering. The three series in figure 8 are the frequencies of items obtained from the 100 simulated adaptive tests. Tests were terminated either when the number of administered items had exceeded 30 or the ability estimate had fallen outside the ability range. The latter were necessary for very high and very low ability students, where adaptive selection could not be used [2]. The examinee's answers were simulated by using a uniform random number generator, where the probability of correct answer was set to be equal to the probability of incorrect answer.

In order to be able to compare these item exposure control strategies, we computed the standard deviance of the frequency series shown in Figure 8. The standard deviance is $\sigma = 17.68$ for the first series not using any item

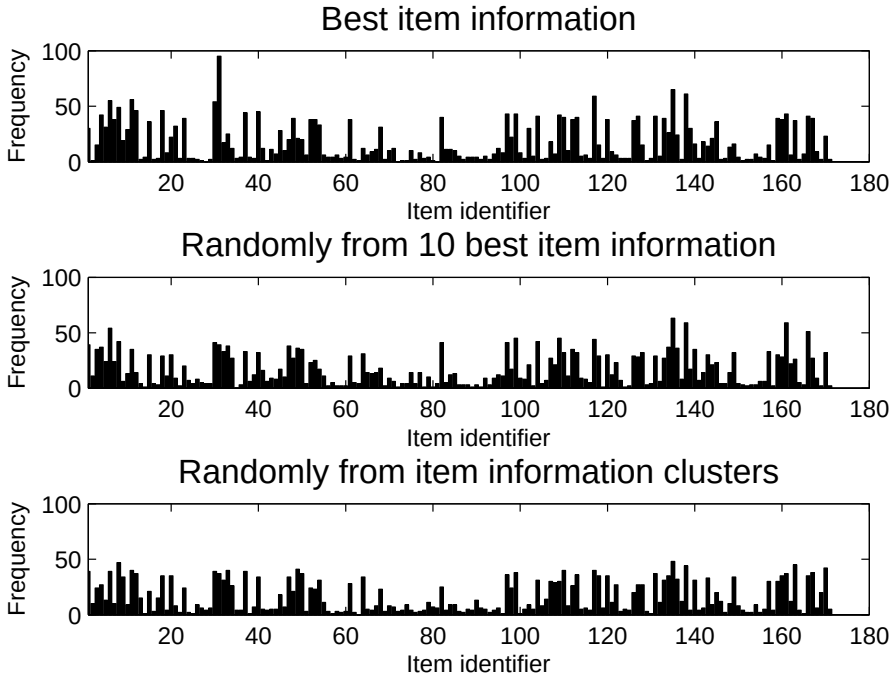

Figure 8: Item exposure with or without randomization control strategies

exposure control, it is $\sigma = 14.77$ for the second one, whereas for the third one is $\sigma = 14.13$. It is obvious that the third series is the best from the viewpoint of item exposure. Consequently, we will adopt this strategy in our distributed CAT implementation.

3.3 Distributed CAT

After the Matlab simulations we implemented our CAT system as a distributed application, using Java technologies on the server side and Adobe Flex on the client side. The general architecture of our system is shown in Figure 9. The administrative part is responsible for item bank maintenance, test scheduling, test results statistics and test termination criteria settings. In the near future we are planning to add an item calibration module.

The test part is used by examinees, where questions are administered ac-

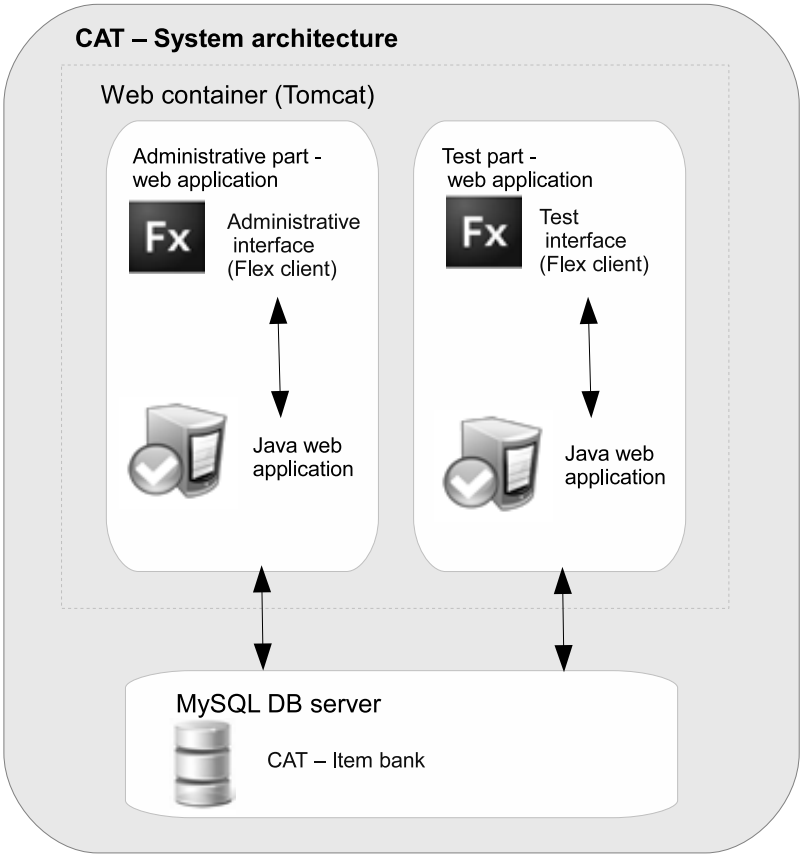

Figure 9: CAT-architecture

cording to settings. After having finished the test, the examinee may view both their test results and knowledge report.

3.4 Item difficulty estimation

Due to the lack of measurement data necessary for item calibration, we were not able to calibrate our item bank. However, 165 out of 171 items of our item bank were used in our self-assessment test system "Intelligent" in the previous term. Based on the data collected from this system, we propose a method for difficulty parameter estimation. Although there were no restrictions in using the self-assessment system, i.e. users could have answered an item several

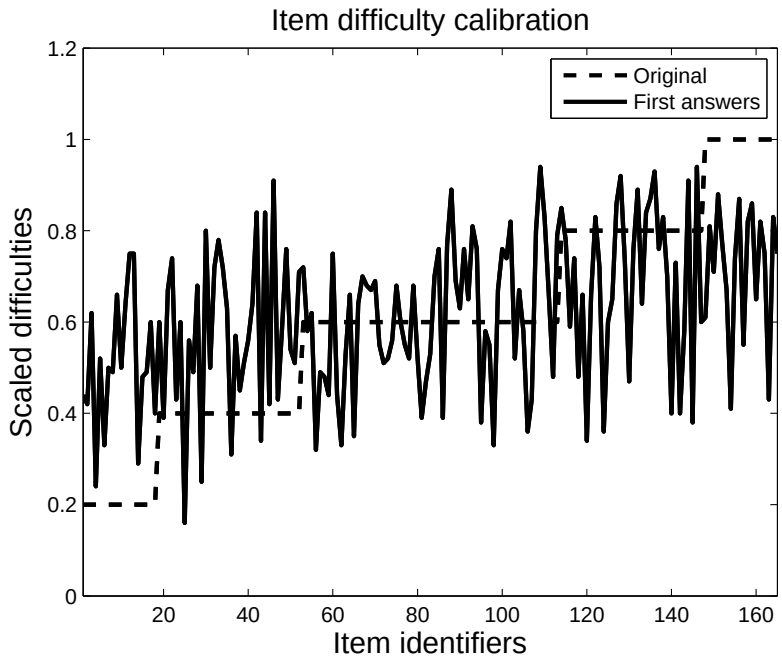

Figure 10: Item difficulty calibration

times, we consider that the first answer of each user could be relevant to the difficulty of the item.

Figure 10 shows the original item difficulty (set by the tutor) and the difficulty estimated by the first answer of each user. The original data series uses 5 difficulty levels scaled to the $[0, 1]$ interval. The elements of the "first answers" series were computed by the equation: $\frac{\text{all\_incorrect\_answers}}{\text{all\_answers}}$. We computed the mean difficulty for both series, and we obtained 0.60 for the original one and 0.62 for the estimated one. Conspicuous differences were found at the *very easy* and *very difficult* item difficulties.

4 Further research

At present we are working on the parameter estimation part of our CAT system. Although there are several item parameter calibration programs, this task must be taken very seriously because it influences measurement precision

directly. Item parameter estimation error is an active research topic, especially for fixed computer tests. For adaptive testing, this problem has been addressed by paper [8].

Researchers have empirically observed that examinees suitable for item difficulty estimations are almost useless when estimating item discrimination. Stocking [14] analytically derived the relationship between the examinee's ability and the accuracy of maximum likelihood item parameter estimation. She concluded that high ability examinees contribute more to difficulty estimation of difficult and very difficult items and less on easy and very easy items. She also concluded that only low ability examinees contribute to the estimation of guessing parameter and examinees, who are informative regarding item difficulty estimation, are not good for item discrimination estimation. Consequently, her results seem to be useful in our item calibration module.

5 Conclusions

In this paper we have described a computer adaptive test system based on Item Response Theory along its implementation issues. Our CAT system was implemented after one year experience with a computer based self-assessment system, which proved useful in configuring the parameters of the items. We started with the presentation of the exact formulas used by our working CAT system, followed by some simulations for item selection strategies in order to control item overexposure. We also presented the exact way of item parameter configuration based on the data taken from the self-assessment system.

Although we do not present measurements on a working CAT system, the implementation details presented in this paper could be useful for small institutions planning to introduce such a system for educational measurements on a small scale.

In the near future we would like to add an item calibration module to the administrative part of the system, taking into account the limited possibilities of small educational institutes.

References

[1] M. Antal, Sz. Koncz, Adaptive knowledge testing systems, *Proc. SZAMOKT XIX*, October 8–11, 2009, Tîrgu Mureş, Romania, pp. 159–164. ⇒176

[2] F. B. Baker, *The Basics of Item Response Theory*, ERIC Clearinghouse on Assessment and Evaluation, College Park, University of Maryland, 2001. ⇒170, 177

[3] M. Barla, M. Bielikova, A. B. Ezzeddinne, T. Kramar, M. Simko, O. Vozar, On the impact of adaptive test question selection for learning efficiency, *Computers & Educations*, **55,** 2 (2010) 846–857. ⇒169

[4] A. Baylari, G. A. Montazer, Design a personalized e-learning system based on item response theory and artificial neural network approach, *Expert Systems with Applications*, **36,** 4 (2009) 8013–8021. ⇒169

[5] E. Georgiadou, E. Triantafillou, A. Economides, A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005, *Journal of Technology, Learning, and Assessment*, **5,** 8 (2007) 33 p. ⇒175

[6] R. K. Hambleton, R. W. Jones, Comparison of classical test theory and item response theory and their applications to test development, *ITEMS – Instructional Topics in Educational Measurement*, 253–262. ⇒169, 170

[7] S. Huang, A content-balanced adaptive testing algorithm for computer-based training systems, in: *Intelligent Tutoring Systems* (eds. C. Frasson, G. Gauthier, A. Lesgold), Springer, Berlin, 1996, pp. 306–314. ⇒175

[8] W. J. Linden, C. A. W. Glas, *Capitalization on item calibration error in computer adaptive testing*, LSAC Research Report 98-04, 2006, 14 p. ⇒181

[9] M. Lilley, T. Barker, C. Britton, The development and evaluation of a software prototype for computer-adaptive testing, *Computers & Education*, **43,** 1–2 (2004) 109–123. ⇒169

[10] F. M. Lord, *Application of item response theory to practical testing problems*, Lawrence Erlbaum Publishers, NJ, 1980. ⇒171

[11] G. Rasch, *Probabilistic models for some intelligence and attainment tests*, Danish Institute for Educational Research, Copenhagen, 1960. ⇒169

[12] L. M. Rudner, *An online, interactive, computer adaptive testing tutorial*, 1998, http://echo.edres.org:8080/scripts/cat/catdemo.htm ⇒170

[13] M. L. Stocking, *Controlling item exposure rates in a realistic adaptive testing paradigm*, Technical Report RR 3-2, Educational Testing Service, Princeton, NJ, 1993. ⇒175

[14] M. L. Stocking, Specifying optimum examinees for item parameter estimation in item response theory, *Psychometrika*, **55,** 3 (1990) 461–475. ⇒181

[15] H. Wainer, G. L. Kiely, Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement*, **24,** 3 (1987) 185–201. ⇒175

Received: August 23, 2010 • Revised: November 3, 2010