STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVIII**, Number 1, 2013

ON THE USE OF ELO RATING FOR ADAPTIVE ASSESSMENT

MARGIT ANTAL

Abstract. In this paper, we present a new item response model for computerized adaptive testing: Item Response Theory combined with Elo rating. Adaptive test systems require a calibrated item bank and item calibration methods are usually based on Item Response Theory (IRT). However, these methods require item pretesting on large sample sizes, which is very expensive. Hence, this paper presents alternative methods for item difficulty calibration.

Results show that combining IRT with Elo rating is an alternative model for adaptive item sequencing, which offers not only estimations for abilities, but for item difficulties too. The new adaptive item sequencing model was compared with IRT on artificial data. Results show that the new method is able to estimate the ability of the examinee, although more items are required compared to IRT. Hence, this method is recommended for test systems where adaptation to the user knowledge level is a requirement, but the duration of the measurement is less important, i.e. practice systems.

1. Introduction

There are more and more adaptive e-learning systems trying to fulfill specific needs of learners. These systems can be adapted to user knowledge, user interests or user individual traits [2], to name a few specific elements of a user model. In this paper we study adaptive test systems where item sequencing is adapted to user knowledge. These systems are known as Computerized adaptive testing (CAT) systems. Many organizations test their examinees using CAT tools (e.g. TOEFL, GMAT, GRE). The CAT tools used in these examinations are not free software, therefore they cannot be deeply evaluated nor compared. However, Economides and Roupas [4] succeed in performing a rough comparison of these CAT systems based on demo versions of these applications. These CAT tools select items from a calibrated item pool. The item

Received by the editors: December 10, 2012.

2010 *Mathematics Subject Classification.* 97Q70.

1998 *CR Categories and Descriptors.* K.3 [**Computers and Education**]: Computer and Information Science Education – *Computer science education* .

Key words and phrases. Item Difficulty, Item Response Theory, Performance Assessment.

pool calibration is usually performed by the means of Item response theory (IRT, [12]).

There are a lot of free and commercial software designed for item calibration (BILOG-MG, MULTILOG, PARAM-3PL etc.) by means of IRT and based on item pre-test data. Unfortunately, quality item pre-testing is very costly, thus educational institutions cannot afford it. This includes both the cost of item prestest and maintenance of such a system as well. Moreover, the item parameter estimates obtained through item pretest are very sensitive to examinees. Stocking in paper [11] conducted a study in order to analyze optimal examinee quality for accurate item parameter estimation. She found that item calibration is very sensitive to examinee quality; hence inappropriate examinees can easily lead to incorrect item parameters. She also concluded that a broad distribution of abilities may provide more information than a bell-shaped distribution.

The first objective of this paper is to search for alternative item parameter estimation methods, which are comparable with IRT item calibration estimates. In this paper we restrict ourselves to the 1PL IRT model, which uses only one item parameter, namely item difficulty. Very few studies are to be found in this field. Wauters et. al. in paper [15] conducted a study in this direction and ranked six alternative item difficulty estimation methods by their correlates to the IRT based estimates. [14] is a previous version of this study. [9] applied the Elo rating for on the fly item difficulty estimation. There are a few studies about teachers or experts ability to estimate item difficulty, concluding that teachers are not very accurate in this task [6], [7], [15].

The second objective of this paper is to introduce a new adaptive item sequencing method based on Elo Rating System (ERS), which is able to estimate both the examinees' ability and item difficulties. In our ERS method we combined Elo rating based ability estimation with IRT based next item selection method. Even though the Elo rating for adaptive item sequencing was studied by other researchers too [1], [9], [15], a detailed comparison with IRT is missing. We compared our new item sequencing method to IRT in order to find out the proper test length (number of test items used) for accurate ability estimations.

This paper is structured in five major sections. The next section describes the ERS and the IRT briefly. Section 3 is devoted to the presentation of our item difficulty estimation methods and their comparison. Section 4 compares IRT and ERS based adaptive item sequencing. Section 5 outlines our conclusions.

2. Theoretical background

2.1. **Elo Rating System.** The ERS was introduced for rating chess players by Arpad Elo in 1978 [5]. In this rating system each player has a rating, which represents their relative ability, thus a higher rating indicates a better player. Players are given an initial ability, which is continuously updated based on match results. The ERS system uses simple computations for updating players ratings.

If A, B are the two players with abilities (performances) Θ_A and Θ_B, formulas (1) and (2) are for ability estimations based on the match result. S_A is the match result for player A (0 loss, 0.5 - draw and 1 win) and $E(S_A)$ is the expected match result for player A in formula 3. S_B is the match result for player B, therefore $1 - S_A$ and $E(S_B)$ is the expected match result for player B (4). The K factor in formulas (1) and (2) weighs the performance change of a player during the matches and usually it is a constant.

$$\hat{\Theta}_A = \Theta_A + K(S_A - E(S_A)), \tag{1}$$

$$\hat{\Theta}_B = \Theta_B + K(S_B - E(S_B)), \tag{2}$$

$$E(S_A) = \frac{1}{1 + 10^{\frac{\Theta_B - \Theta_A}{400}}}, \tag{3}$$

$$E(S_B) = \frac{1}{1 + 10^{\frac{\Theta_A - \Theta_B}{400}}}, \tag{4}$$

In adaptive test systems one of the players is the examinee and the other one is the test item. Before answering the test item, the examinee has ability Θ and the test item has difficulty b, both being on the same scale, usually [-3, 3]. In this case a match is an answer given to the item, which can be correct or incorrect. We derive the new formulas for ability (5) and difficulty (6) estimates as follows

$$\hat{\Theta} = \Theta + K(S - E(S)) \tag{5}$$

$$\hat{b} = b - K(S - E(S)) \tag{6}$$

If the examinee gives a correct answer, then S is 1 and 0 otherwise. Thus for every correct answer the proficiency of the examinee increases and the difficulty of the answered item decreases and vice versa. In this study a constant value

0.4 [15] was used for K, although K could be chosen to reflect the uncertainty in ability estimates by making it a function of the number of answered items. For the expected match result $E(S)$ we used the logistic function given in formula (7).

$$E(S) = \frac{1}{1 + e^{-(\Theta - b)}} \tag{7}$$

2.2. **Item Response Teory.** IRT has its roots in Psychometrics and it defines a method for adaptive item selection. In this model each item is characterized by an Item Characteristic Curve (ICC, [10]), which shows the correct answer probability as a function of ability. ICC is usually modeled by a 3 parameter logistic function (8) also known as 3PL, where Θ represents student ability; a, b and c are item parameters representing discrimination, difficulty and guessing probability, and D is a scaling factor.

$$P(\Theta) = c + \frac{1 - c}{1 + e^{-Da(\Theta - b)}} \tag{8}$$

Replacing $c = 0$, $a = 1$ and $D = 1$ in formula (8) we obtain the 1PL model also known as the Rasch model, whose logistic function (9) is identical with (7).

$$P(\Theta) = \frac{1}{1 + e^{-(\Theta - b)}} \tag{9}$$

Items are administered based on their usefulness in ability estimation. For this purpose items are ranked on their item information values and the maximum one is chosen [10]. Item information for the 3PL model is given in formula (10).

$$I_i(\Theta) = \frac{P_i'(\Theta)^2}{P_i(\Theta)(1 - P_i(\Theta))} \tag{10}$$

The item information function for the 1PL model is

$$I_i(\Theta) = P_i(\Theta)(1 - P_i(\Theta)). \tag{11}$$

For new ability estimates we used the formulas (12) and (13) [10],

$$\Theta_{n+1} = \Theta_n + \frac{\sum\limits_{i=1}^{n} S_i(\Theta_n)}{\sum\limits_{i=1}^{n} I_i(\Theta_n)} \tag{12}$$

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

$$S_i(\Theta) = (u_i - P_i(\Theta))\frac{P_i'(\Theta)}{P_i(\Theta)(1 - P_i(\Theta))} \tag{13}$$

In formula (13) u_i is 1 if the answer for the ith item is correct and 0 otherwise. The standard error shows the accuracy of ability estimation, and after N administered items it can be computed by the following formula

$$SE(\Theta) = \frac{1}{\sqrt{\sum_{i=1}^{N} I_i(\Theta)}}. \tag{14}$$

3. Item difficulty estimation

Item parameters estimation is usually carried out prior to computerized adaptive testing and requires additional costs.

Not only the cost is an impediment but also the continuous addition of new test items, requiring continuous calibration. Stocking [11] observed that item calibration is very sensitive to examinee abilities, therefore non-compliant subjects can lead to incorrect item parameters. A few studies [8], [9], [15] tried to offer alternative and more lightweight solutions to item parameters estimation; moreover, [15] compared 6 alternative estimation methods with IRT-based calibration and found that Proportion Correct method has the strongest correlation with IRT-based difficulty estimates. Proportion Correct method is a simple approach, which estimates the difficulty level of items by dividing the number of incorrect answers by the number of total answers for the given item. In order to grasp the differences between various item difficulty estimation methods we conducted an experiment, which is partially similar to the one conducted by Wauters et al, 2011. We compared three item difficulty estimation methods: IRT, ERS and Proportion Correct using real test data.

3.1. **Participants.** Students from Computer Science, Automation and Applied Informatics and Informatics participated in this study. Data were collected during real examination sessions at Object-oriented programming. There

Table 1. Pearson correlation matrix of the item difficulty estimates for the OOP1 dataset

	IRT	ERS	Prop. Correct
IRT	1	0.942	0.994
ERS		1	0.937
Prop. Correct			1

Table 2. Pearson correlation matrix of the item difficulty estimates for the OOP2 dataset

	IRT	ERS	Prop. Correct
IRT	1	0.930	0.991
ERS		1	0.931
Prop. Correct			1

are two datasets: OOP1, collected in 2011 from 74 students and and OOP2, collected in 2012 from 63 students.

3.2. **Material and procedure.** The tests were administered using Moodle [16] and both tests consisted of 30 items (single choice, multiple choice, fill in). Students were familiarized with the Moodle test environment as they had already completed a pretest session. A test is considered reliable if repeated administration of the test give the same result. Test reliability was measured using Cronbachs alpha coefficient and we obtained 0.857 and 0.873 for datasets OOP1 and OOP2, respectively.

IRT model parameter calibration was conducted in R [17] using the ltm package. For Proportion Correct we used formula (15),

$$\hat{b_i} = 1 - \frac{n_i}{N_i} \tag{15}$$

where N_i and n_i are the total number of answers, the number of correct answers to the ith item respectively. Finally, for the ERS formulas (5), (6) and (7) were used.

3.3. **Results.** Figure 1 shows item difficulty estimations obtained by IRT, Proportion Correct and ERS methods for items in the OOP1 dataset. IRT estimation was conducted using the *rasch* function of the *ltm* package [17]. Using formula (15), the values obtained by the Proportion Correct method were scaled to the [-3,3] interval in order to be comparable to IRT estimations.

For ERS estimation each item difficulty was initialised by 0 and we used formulas (5), (6) and (7) in the estimation process.

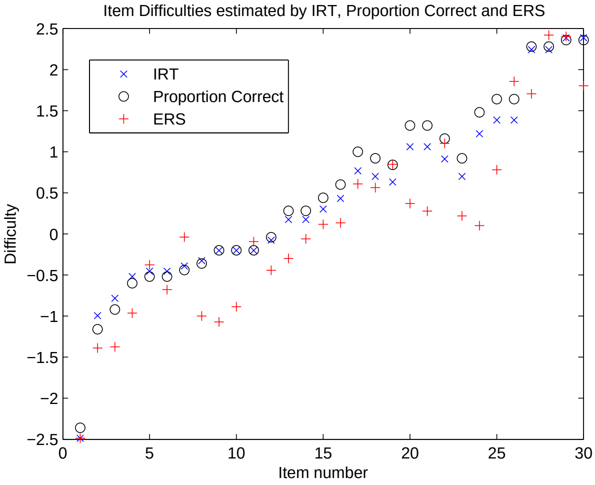

Figure 1. Item difficulty estimations by IRT, Proportion Correct and ERS methods (OOP1 dataset)

Tables 1 and 2 show Pearson correlations between the estimated item difficulty parameters using different estimation methods.

Proportion Correct has the strongest correlation with IRT calibration. However, all estimation methods produced difficulty estimates, which highly correlated with the other estimation methods.

4. Ability estimation using adaptive item sequencing

A computerized adaptive test is usually administered using the following steps:

(1) Administer a few items of average difficulty. Score the responses and estimate the person's initial ability level.
(2) Select the item that provides the most information using the person's current ability estimate and score the response.
(3) Re-estimate the person's ability level
(4) If stopping criterion is met, then stop the test, otherwise go to step 2.

The ability estimates obtained by IRT-based adaptive item sequencing are compared with ERS-based one using simulated data.

The only difference between the two methods is the re-estimation formula used for the person's ability, namely (12) for IRT and (5) for ERS. However, the ERS-based sequencing offers a formula for item difficulty estimation (6),

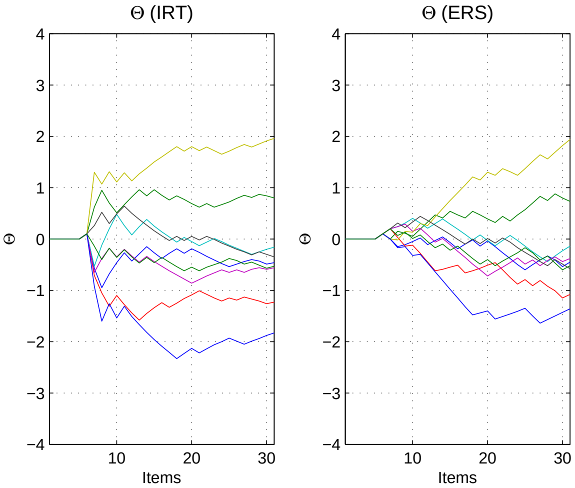

Figure 2. Ability estimations by IRT and ERS methods for the first 10 artificial examinees using a 30 item length test

which was not used in this experiment. The next item in both methods is the one having the maximum item information (10) for the person's ability.

There were 1000 artificial examinees simulated using an ideal item bank with 200 items having difficulties uniformly distributed in the [-3, 3] interval. In order to get comparable results, we generated the answer patterns for the examinees using a random number generator with uniform distribution. The first five items were randomly chosen and the estimation process was started after the 5th administered item. For each simulated examinee the same number of items were administered and the same predefined response pattern was used for both methods.

Figure 2 presents the results for the first 10 examinees, whereas the first two subfigures of figure 3 show ability values estimated during adaptive sequencing of 30 items using IRT and ERS methods for the 6th examinee. The third subfigure 3 of figure shows the answers given by the examinee. In this experiment we did not use stopping criteria, consequently all the 30 items were administered.

In the second experiment we compared IRT- and ERS-based adaptive sequencing methods from the viewpoint of ability estimates convergence. In order to achieve it we used various fixed length tests and compared the ability estimates obtained by the two methods at the end of tests. For test lengths we set the following values one by one: 10, 15, 20, 25 and 30. For each test length we repeated the experiment described above and obtained the final

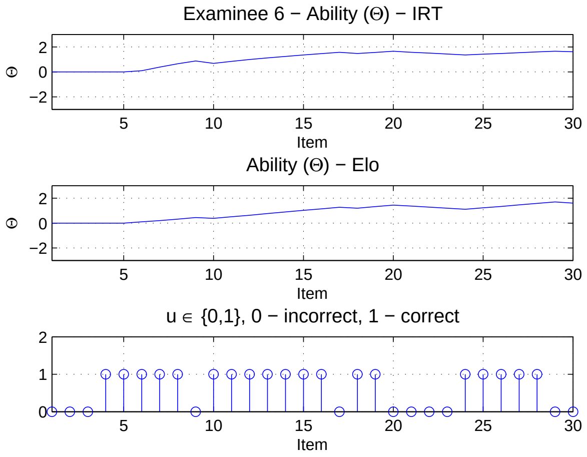

FIGURE 3. Ability estimations by IRT and ERS methods for a given examinee

TABLE 3. Influence of test length on ability estimations with IRT and ERS methods

Test length	$\mu(\Theta_{IRT} - \Theta_{ERS})$	$\sigma(\Theta_{IRT} - \Theta_{ERS})$	$\mu(SE)$	$\sigma(SE)$
10	0.68	0.52	0.46	0.034
15	0.49	0.41	0.35	0.033
20	0.31	0.27	0.30	0.011
25	0.23	0.20	0.26	0.009
30	0.18	0.14	0.23	0.007

ability estimates for IRT and ERS methods after the end of the test. Table 3 summarizes the results of the experiment: the first column contains the test length, the second and the third columns contain the mean and the standard deviation of $\Theta_{IRT} - \Theta_{ERS}$ for the 1000 examinees. Columns 4 and 5 show the mean standard error and its standard deviation computed in the case of IRT estimate. It can be seen that IRT estimates the examinees ability faster than ERS. For the IRT method the standard error is considered a measure of the accuracy of the ability estimate and this value falls below 0.4 for a test containing approximately 15 items. For a 30 item length test the estimated abilities by the two methods are very close, the average difference is 0.18 with a standard deviation of 0.14.

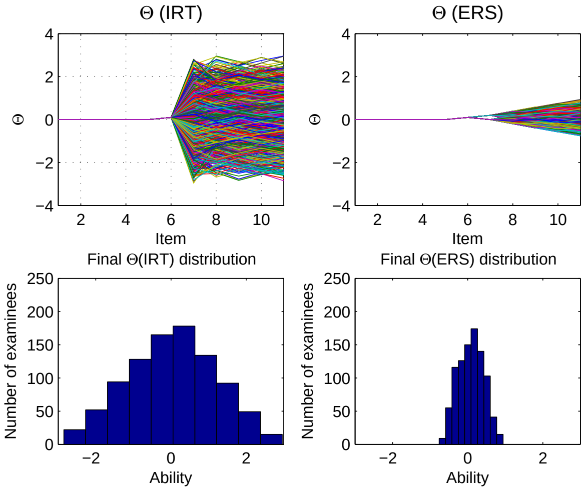

FIGURE 4. Ability estimations by IRT and ERS methods for 1000 artificial examinees using a 10 item length test

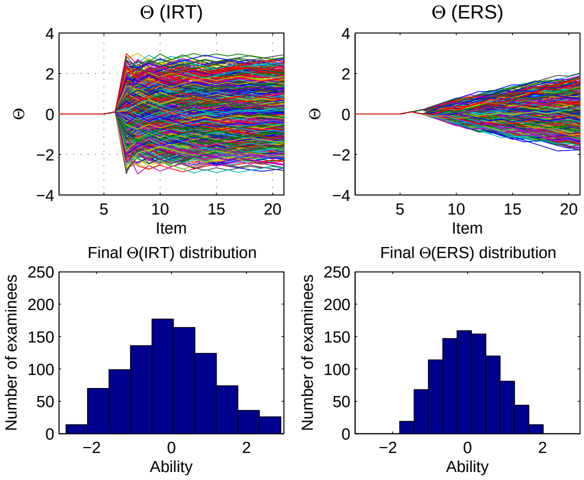

FIGURE 5. Ability estimations by IRT and ERS methods for 1000 artificial examinees using a 20 item length test

Figures 4, 5, 6 show the results of adaptive item sequencing simulations for 10, 20 and 30 items. On the top left figure Θ is estimated by IRT using formula (12), while on the top right figure by ERS using formula (5). Bottom figures

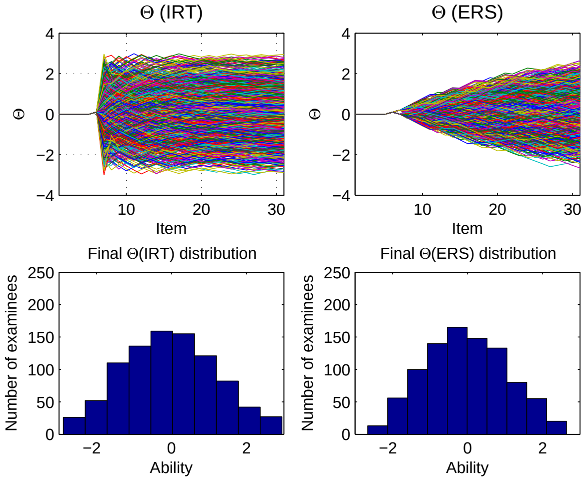

FIGURE 6. Ability estimations by IRT and ERS methods for 1000 artificial examinees using a 30 item length test

show the histograms of ability estimates. These figures show the influence of the test length on ability estimations.

5. Conclusions

In this paper we presented alternative methods for item difficulty estimation. Using real test data we compared two alternative item difficulty estimation methods to IRT-based estimation. The high correlations between difficulty estimates obtained by IRT, ERS and Proportion Correct methods indicate that any of these can be used in real adaptive test systems. We also presented a new model for computerized adaptive testing: an IRT-based adaptive item sequencing with ERS-based ability estimate. We found that the new model is able to obtain a reliable examinee ability estimate using a test of at least 30 items. This result is in concordance with the finding obtained by van der Maas and Wagenmarkers regarding Elo rating used for ranking chess players, as they concluded that 25 games are needed to obtain a reliable Elo rating for a chess player [13]. Moreover, using this model we can obtain both the ability estimates of the examinees and item difficulty estimates of the test items. One limitation of this study is the size of population used for item difficulty estimation. Therefore we are planning to repeat these estimations as we gather more data.

6. Acknowledgements

This research has been supported by Sapientia Institute for Research Programmes.

References

[1] Brinkhuis, M. J. S., Maris, G. *Dynamic Parameter Estimation in Student Monitoring Systems.* CITO-report (2009).

[2] Brusilovsky, P., Millan, E. *User models for adaptive hypermedia and adaptive educational systems.* In: P. Brusilovsky, A. Kobsa and W. Neidl (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321, Berlin Heidelberg New York: Springer-Verlag, pp. 3-53 (2007).

[3] Cronbach, L.J. *Coefficient Alpha and the internal structure of tests.* Psychometrika 16(3), pp. 297-334 (1951).

[4] Economides, A.A., Roupas, C. *Evaluation of computer adaptive testing systems.* International Journal of Web-Based Learning and Teaching Technologies 2(1), pp. 70-87 (2007).

[5] Elo, A. E., *The rating of chess players, past and present.* B.T. Batsford, Ltd., London (1978).

[6] Impara, J. C., Plake, B. S. *Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method.* Journal of Educational Measurement, 35(1), pp. 69-81 (1998).

[7] Kibble, J. D., Johnson, T. *Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations?* Advances in Physiology Education 35, pp. 396-401 (2011).

[8] Kingsbury, G. *Adaptive Item Calibration: A Process for Estimating Item Parameters Within a Computerized Adaptive Test.* In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing (2009).

[9] Klinkenberg, S., Straatemeier, M., van der Maas, H. L. J. *Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation.* Computers & Education 57(2), pp. 1813-1824 (2011).

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

[11] Stocking,M. L. *Specifying optimum examinees for item parameter estimation in Item Response Theory.* Psychometrika 55(3), pp. 461-475 (1990).

[12] Van der Linden, W. J., Hambleton, R. K. *Handbook of modern item response theory.* New York, Springer (1997).

[13] Van der Maas, H. L., Wagenmakers, E.-J. *A psychometric analysis of chess expertise* American Journal of Psychology 118(1), pp. 29-60 (2005).

[14] Wauters, K., Desmet, P., Van Den Noortgate, W. *Acquiring Item Difficulty Estimations: a Collaborative Effort of Data and Judgment.* Educational Data Mining EDM pp. 121-128 (2011).

[15] Wauters, K., Desmet, P., Van Den Noortgate, W. *Item difficulty estimation: An auspicious collaboration between data and judgement.* Computers & Education 58(4), pp. 1183-1193 (2012).

[16] http://moodle.org/

[17] http://www.r-project.org/

Sapientia - Hungarian University of Transylvania, Faculty of Technical and Human Sciences, 540053 Tirgu Mures, 540485, Aleea Sighisoarei 1C., Romania

E-mail address: manyi@ms.sapientia.ro