Monitoring Learners' Proficiency: Weight Adaptation in the Elo Rating System

K. WAUTERS

Katholieke Universiteit Leuven, Belgium

P. DESMET

Katholieke Universiteit Leuven, Belgium

AND

W. VAN DEN NOORTGATE

Katholieke Universiteit Leuven, Belgium

Adaptive item sequencing is a well-established adaptation technique for personalizing learning environments and can be achieved through an intense reciprocity between the item difficulty level and the learner's proficiency. Consequently, the need to monitor learners' proficiency level is of great importance. On that account, researchers have brought forward the Elo rating system. While the Elo rating system has its origin in chess, it has proven its value during its short history in the educational setting. Elo's algorithm implies that the rating after an event is function of the pre-event rating, the weight given to the new observation and the difference between the new observed score and the expected score. It seems reasonable to adapt the weight of Elo's algorithm as function of the number of observations: the more previous observations we have, the more certain we are about the learner's proficiency estimate, and the less this estimate should be affected by a new observation. The aim of this paper is to search for weights as a function of the number of previous observations that results in optimally accurate proficiency estimates, making use of a real data set. Results indicate that the Elo algorithm with a logistic weight function better approximates the parameter estimates obtained with the item response theory than the Elo algorithm with a fixed weight.

Key Words and Phrases: IRT, proficiency and Elo rating

1. INTRODUCTION

The research on e-learning environments has long been focused on delivering information online without taking into account the characteristics of the particular learner, course material and/or the context. It is only recently that research attention is drawn to dynamic or adaptive e-learning environments. An adaptive learning environment creates a personalized learning opportunity by incorporating one or more adaptation techniques to meet the learners' needs and preferences (Brusilovsky 1999). One of those adaptation techniques is adaptive item sequencing, in which the sequencing of the learning material is adapted to learner-, item-, and/or context characteristics (Wauters, Desmet & Van den Noortgate 2010). Hence, adaptive item sequencing can be established by matching the difficulty of the item to the proficiency level of the learner. Recently, the interest in adaptive item sequencing has grown, as it is found that excessively difficult items can frustrate learners, while excessively easy items can cause learners to lack any sense of challenge (e.g. Pérez-Marín, Alfonseca & Rodriguez 2006, Leung & Li 2007). Learners prefer learning environments where the item selection procedure is adapted to their proficiency. This is already accomplished to a certain extent in computerized adaptive tests (CATs; Wainer 2000).

A prerequisite for efficient adaptive item sequencing is to be able to estimate the learner's proficiency level at an early stage of the learning process and follow the learning curve adequately. Hence, the problem of the proficiency estimation is twofold (Wauters et al. 2010). On the one hand, we need to estimate the learner's proficiency level when little information is provided, which is referred to as the cold start problem (Masthoff 2004). On the other hand, we

Authors' addresses: K. Wauters, ITEC/IBBT, Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Kortrijk, Belgium. E-mail: kelly.wauters@kuleuven-kortrijk.be; P. Desmet, ITEC/IBBT, Faculty of Arts, Katholieke Universiteit Leuven, Kortrijk, Belgium. E-mail: pietdesmet@kuleuven-kortrijk.be; W. Van den Noortgate, ITEC/IBBT, Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Kortrijk, Belgium. E-mail: wim.vandennoortgate@kuleuven-kortrijk.be.

need to be able to follow the progression of the learner as learners are presumed to learn while they are working in the learning environment. As most learning environments are created to learn more than one skill, Bayesian networks are the dominant method of student modeling (Corbett & Anderson 1995). However, when no relationship exists between the skills to be learned, other methods can be applied. One method to model the learner's proficiency progress is by assessing while working in the learning environment, which can be done by means of progress testing. In progress testing, tests are frequently administered to allow for a quick intervention when atypical growth patterns are observed. In order to estimate the proficiency level of the learner more precisely, these tests can be made adaptive (Wainer 2000). Such CATs often make use of the item response theory (IRT; Van der Linden & Hambleton 1997) to estimate the proficiency level of the person, based of the results of prior calibration in which the item difficulty parameters are estimated. Hence, progress testing on the basis of CAT and IRT has the advantage that it makes more precise estimation of the person's proficiency level possible, but it requires costly prior calibration and it is computationally demanding. Furthermore, in order to be able to recover the atypical growth pattern, the proficiency assessment not only needs to be long enough to be accurate, but should also occur on a regular basis. Hence, progress testing is less appropriate for learning environments as it interrupts the learning process. Another approach, one that is well suited for learning environments, is the tracking of the learner's proficiency level. This can be achieved by updating the proficiency level of the learner after each item administration. One method that allows the tracking of the learner's proficiency level is the Elo rating system (Elo 1978). The Elo rating system was originally developed for rating chess performances and is recently implemented in the educational field (Brinkhuis & Maris 2010). When applied in chess, a chess player competes with his opponent, which results in a win, a loss or a draw. This data is known as paired comparison data, for which the Elo rating system was developed. In the educational field, we have a similar kind of paired comparison, where the learner is seen as a player and the item is seen as its opponent. In the formula of Brinkhuis and Maris (2010), the new proficiency rating after an item administration is function of the pre-administration rating, a weight given to the new observation and the difference between the actual score on the new observation and the expected score on the new observation. This expected score is calculated by means of the Rasch model. This formula implies that when the difference between the expected score and the observed score is high, the change in the estimate of the proficiency level is high. This algorithm enables continuous measurement, since the rating is updated after every event. This implies that the Elo rating system is order-sensitive. The formula for updating the proficiency level (the item difficulty is updated at the same time and in a similar way), is given by:

$$\theta_n = \theta_0 + W(X - X_e)$$

where θ_n is the new learner's proficiency level rating after the learner has answered an item, θ_0 is the pre-event rating, W is the weight given to the new observation, X is the actual observation (score 1 for correct, 0 for incorrect), and X_e is the expected observation which is estimated based on the Rasch model. Hence, the formula for updating the learner's proficiency level after a response (X=0 or1) becomes:

$$\theta_n = \theta_0 + W\left[X - \frac{\exp(\theta_0 - \beta_0)}{1 + \exp(\theta_0 - \beta_0)}\right],$$

where β_0 is the estimated item difficulty level before that item is answered by this specific person. Because the estimation of the proficiency level becomes more stable when more items are answered by this person, the weight given to the new observation should decrease when the rating of the learner's proficiency level is based on more observations. In this study we will focus on this weight adaptation by searching for a function that can describe the decrease in weight. More specifically, the aim of this paper is to compare the proficiency levels obtained by means of IRT estimation with those estimated on the basis of the Elo rating system with various weight functions. The focus is not on tracking the learner's proficiency level, but on promptly estimating the learner's proficiency level. In future research, the focus will go to monitoring the learner's proficiency level.

2. EXPERIMENT

2.1 Learners' proficiency Estimation Methods

2.1.1 *Item Response Theory.* To estimate the learners' proficiency level, the IRT model with a single item parameter proposed by Rasch (Van der Linden & Hambleton 1997) is used. The Rasch model models the probability of answering an item correctly as a logistic function of the difference between the person's proficiency level (θ) and the item difficulty level (β):

$$Prob(X_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}$$

In this study, the proficiency level of the learners is estimated with known item difficulty levels. The difficulty level of the items was estimated in advance by Selor, the selection agency of the Belgian government. Selor used examinee data to estimate the item difficulty parameters by means of IRT.

2.1.2 *Elo Rating.* In this study, the Elo rating systems implemented by Brinkhuis and Maris (2010) was used to estimate learners' proficiency level.

$$\theta_n = \theta_0 + W\left[X - \frac{\exp(\theta_0 - \beta_0)}{1 + \exp(\theta_0 - \beta_0)}\right]$$

The weight function should have two crucial characteristics: (1) it should slowly decrease when the number of items answered by that learner increases; (2) it should never become zero or negative. A function that satisfies these requirements is the logistic function. Hence, we propose a weight function of the logistic family in which three parameters will be varied.

$$W = \frac{W_0}{1 + a * \exp(b * N_{ip})},$$

where N_{ip} is the number of items answered by learner p before answering item i. The three parameters that will be varied in this equation are W_0, which is the starting weight, a, and b. The a parameter influences the slope of the logistic curve (figure1a), the b-parameter influences at what point the weight is reduced from W_0 to $W_0/2$ (figure 1b). A b-parameter set to 50 indicates that the weight given to the new observation is halved by the time the fiftieth item is presented.

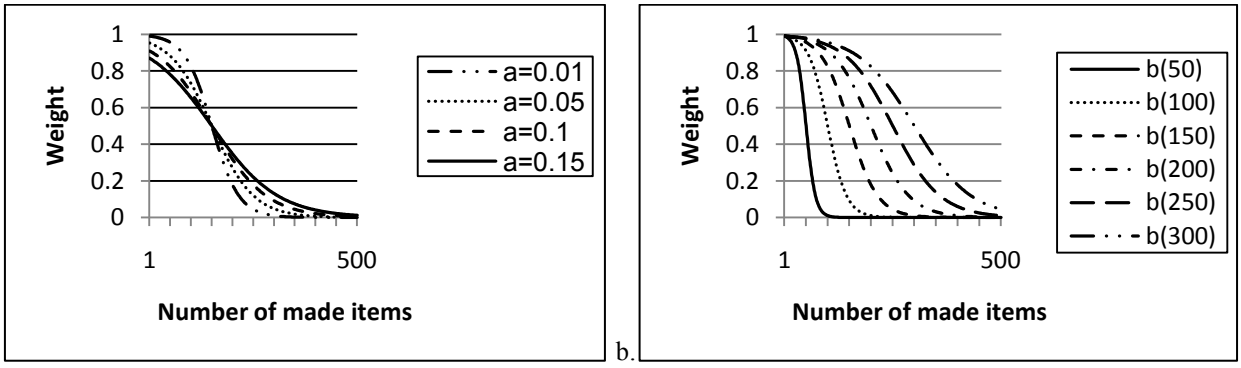

Fig. 1. Logistic weight function with W_0=1, b=150 (i.e. the weight is halved by the time the 150th item is presented) and several a-values (1a); Logistic weight function with W_0=1, a=.01 and several b-values (1b).

We hypothesize that the Elo rating formula will be more efficient when the starting weight is high, because when little information is known about the learner's proficiency level all the information needs to be extracted from the new observations. Furthermore, we hypothesize that the weight should decrease as the number of prior observations increases. When the estimation of the learner's proficiency level is based on many observations, the estimate is fairly reliable and the new observation should not give much weight to the updating of the proficiency level.

2.2 Method

2.2.1 Participants. Students from ten educational programs in the Flemish part of Belgium (1st and 2nd Bachelor Linguistics and Literature – K.U.Leuven; 1st, 2nd and 3th Bachelor Teacher-Training for primary education – Katho Tielt; 1st and 2nd Bachelor Teacher-Training for secondary education – Katho Reno; 1st and 2nd Bachelor of Applied Linguistics – HUB and Lessius; and 1st Bachelor Educational Science – K.U.Leuven) were contacted to participate in the experiment. Two hundred thirty two students completed the whole experiment.

2.2.1 Material and Procedure. The study consisted of two sessions, each taking approximately half an hour. The learning material consisted of items on French verb conjugation, supposedly measuring one single skill. The instructions, comprising information on the login procedure for the learning environment and on the proceedings of the experimental study were sent to the participants by email. When students were logged on, they were given an informed consent. Subsequently, they completed 25 items. After two weeks, students completed the next 25 items. The difficulty of both tests of 25 items can be considered equal as the composing items are of equal difficulty level. IRT estimation and the Elo rating system both combine the answers of the learner on the two tests, resulting in 50 items.

2.3 Results

Learners score significantly higher on the posttest (M=66.64, SD=17.66) compared to the pretest (M=63.86, SD=18.15), t(231)=-3.09, $p<.001$. For each of the two tests, a logistic regression analysis with the raw score as dependent variable and the item order and item difficulty parameter as independent variables is performed to assess the within-test gains. Difficulty significantly predicted the odds of a correct response (β=-0.78, t(1)=705.29, $p<.001$ for the pretest, and β=-0.67, t(1)=541.57, $p<.001$ for the posttest). However, no significant effect was found of item order, indicating no within-test learning gains.

The Pearson correlation between the estimated learners' proficiency level on the basis of IRT and the estimated learners' proficiency level on the basis of the Elo rating system was the criterion used to evaluate the efficacy of the logistic weight function. Detailed correlation results for the learners' proficiency level estimates are shown in table I.

Table I. Pearson correlation matrix of the learners' proficiency level estimates for the different parameter values of the logistic weight function included in the Elo rating system.

		W_0				
a	**b**	**1**	**0.8**	**0.6**	**0.4**	**0.2**
0	**0**	.79	.82	.85	.88	.92
0.01	**50**	.85	.87	.89	.91	.94
0.01	**100**	.80	.83	.86	.89	.92
0.01	**150**	.80	.82	.85	.89	. 92
0.01	**200**	.80	.82	.85	.88	.92
0.01	**250**	.79	.82	.85	.88	.92
0.01	**300**	.79	.82	.85	.88	.92
0.05	**50**	.86	.88	.90	.92	. 94
0.05	**100**	. 81	. 84	. 87	.90	.93
0.05	**150**	.81	.83	.86	.89	.93
0.05	**200**	.80	.83	.86	.89	.92
0.05	**250**	.80	.83	.86	.89	.92
0.05	**300**	.80	. 83	. 85	. 89	.92
0.10	**50**	.86	.88	.90	.92	.94
0.10	**100**	.82	.85	.87	.90	.93
0.10	**150**	.81	.84	.87	.90	.93
0.10	**200**	.81	.84	.86	.89	. 93
0.10	**250**	. 81	. 83	.86	.89	.93
0.10	**300**	.81	.83	.86	.89	.93
0.15	**50**	.86	.88	.90	.92	.94
0.15	**100**	.83	.85	.88	.90	.93
0.15	**150**	. 82	. 84	. 87	.90	.93
0.15	**200**	.82	.84	.87	.90	.93
0.15	**250**	.82	.84	.87	.90	.93
0.15	**300**	.81	.84	.86	.89	.93

Note: all correlations are statistically significant at the .001 significance level.

The Pearson correlation coefficient between IRT-based proficiency estimates and the proficiency estimates obtained by means of the Elo rating system shows that the Elo rating system with an initial weight of 0.2 (last column) outperformed the Elo rating system with an initial weight of 0.4 (t(229)=-10.93, $p<.001$) or higher. The results further indicate that the correlation between IRT-based proficiency estimates and the proficiency estimates obtained by means of the Elo rating system with a logistic weight function (a>0) is higher than the correlation between IRT-based proficiency estimates and the proficiency parameter estimates obtained through the Elo rating system with a fixed weight (a=0; for example, the difference between the correlation coefficients in case W_0=0.2 and a=0, and the correlation coefficient in case W_0=0.2, a=0.01 and the b-parameter is set so the weight is halved by the time the fiftieth item is presented to the learner, i.e. b=50 in table I is significant, t(229)=-6.90, $p<.001$). Furthermore, results indicate that the Pearson correlation coefficient is the highest when the a-parameter has a value above 0.01 and the b-parameter is set so the weight is halved by the time the fiftieth item is presented to the learner, i.e. b=50 in table I.

3. DISCUSSION

Monitoring the learner's progress is an important component in the creation of an adaptive learning environment by means of adaptive item sequencing. Because IRT is computationally demanding and the requirement of prior item difficulty estimation is costly, the Elo rating system has been put forward as an alternative by Brinkhuis and Maris (2010). The Elo rating system allows for dynamic parameter evaluation without prior item difficulty estimation. As it is already shown that the Elo rating system has its value in the educational field for tracking the learner's proficiency level, this article further builds on the Elo rating system and extends the formula of Brinkhuis and Maris (2010) by including a logistic weight function that describes the decrease in weight as the number of observations increases. Based on the response data of two hundred thirty two learners on fifty items, the proficiency level was estimated by means of IRT, the Elo rating system with a fixed weight value, and the Elo rating system with a logistic weight function in which three parameters were varied: (1) the initial weight, (2) the a-parameter, and (3) the b-parameter.

The findings indicate that, at least for data as the ones collected, the weight given to the new observation should be small at the beginning. This means that, even though the estimation of the learner's proficiency level is based on a small amount of observations and therefore unreliable, the outcome of a new observation only gives small weight to the update of the proficiency level. This result is inconsistent with our hypothesis and could be due to the simultaneous estimation of the item difficulty parameter and the learner's proficiency parameter in the Elo rating system. In this study, both the item difficulty level and the learner's proficiency level were estimated. Therefore, not only the estimation of the learner's proficiency level was unreliable at the beginning, but also the estimation of the item difficulty level and hence, of the expected outcome of the new observation. In such a situation, the small weight given to a new observation could be advocated.

Results also indicate that the weight should decrease as the number of observations increases. This finding can be deduced from the lower correlation of IRT with the Elo rating estimates obtained with a fixed weight compared to the correlation of IRT with the Elo rating estimates with a decreasing weight. This result is in line with our hypotheses arguing that as the estimation of the learner's proficiency level is based on more observations, the estimation becomes more stable and reliable. Furthermore, the results indicate that this decrease in weight should be steep as the best correlation with IRT-based estimation is found when the weight is halved by the time the fiftieth item is presented to the learner. This could also be explained by the simultaneous estimation of the item difficulty parameter and the learner's proficiency parameter in the Elo rating system. As the expected score on a new observation is unreliable due to the unreliable item difficulty estimate, the information provided by this new observation is small and little weight should be given to this new observation.

Even though this study gives an indication of the parameter values resulting in a good estimation of the learner's proficiency level, it should be considered that these results depend on this particular data set. We recognize that results might depend on the learning rate of the student, the skill difficulty level, the variance in item difficulty, etc.

Future research will review the learner's proficiency parameter estimates obtained on the basis of the Elo rating system with starting item difficulty parameter values estimated by means of prior IRT-based calibration. This should allow us to evaluate the logistic weight function within the Elo rating system for estimating the learner's proficiency level when items difficulty parameters, and hence the expected score on a new observation are more reliable. Another research question will address the relationship of the weight function with other variables, such as time between two learning sessions and hint usage.

REFERENCES

CORBETT, A., AND ANDERSON, J. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

BRINKHUIS, M.J.S., AND MARIS, G. 2010. Adaptive Estimation: How to Hit a Moving Target. Report No.2010-1, Measurement and Research Department, Cito, Arnhem

BRUSILOVSKY, P. 1999. Adaptive and Intelligent Technologies for Web-Based Education. *Künstliche Intelligenz*, 4, 19-25.

ELO, A.E. 1978. *The Rating of Chess Players, Past and Present*, B.T. Batsford Ltd., London.

LEUNG, E.W.C., AND LI, Q. 2007. An Experimental Study of a Personalized Learning Environment Through Open-Source Software Tools. *IEEE Transaction on Education*, 50, 331-337.

MASTHOFF, J. 2004. Group modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. *User Modeling and User-Adapted Interaction*, 14, 37-85.

PÉREZ-MARÍN, D., ALFONSECA, E., AND RODRIGUEZ, P. 2006. On the Dynamic Adaptation of Computer Assisted Assessment of Free-Text Answers. *Adaptive Hypermedia and Adaptive Web-Based Systems, Proceedings*, 4018, 374-377.

VAN DER LINDEN, W.J., AND HAMBLETON, R.K. 1997. *Handbook of Modern Item Response Theory*, Springer, New York.

WAINER, H. 200. *Computerized Adaptive Testing: a Primer*, Erlbaum, London.

WAUTERS, K., DESMET, P., AND VAN DEN NOORTGATE, W. 2010. Adaptive Item-Based Learning Environments Based on the Item Response Theory: Possibilities and Challenges. *Journal of Computer Assisted Learning,* 26(6), 549-562.