Acquiring Item Difficulty Estimates: a Collaborative Effort of Data and Judgment

K. WAUTERS

Katholieke Universiteit Leuven, Belgium

P. DESMET

Katholieke Universiteit Leuven, Belgium

AND

W. VAN DEN NOORTGATE

Katholieke Universiteit Leuven, Belgium

The evolution from static to dynamic electronic learning environments has stimulated the research on adaptive item sequencing. A prerequisite for adaptive item sequencing, in which the difficulty of the item is constantly matched to the knowledge level of the learner is to have items with a known difficulty level. The difficulty level can be estimated by means of the item response theory (IRT), as often done prior to computerized adaptive testing. However, the requirement of this calibration method is not easily met in many practical learning situations, for instance, due to the cost of prior calibration and due to continuous generation of new learning items. The aim of this paper is to search for alternative estimation methods and to review the accuracy of these methods as compared to IRT-based calibration. Using real data, six estimation methods are compared with IRT-based calibration: proportion correct, learner feedback, expert rating, paired comparison (learner), paired comparison (expert) and the Elo rating system. Results indicate that proportion correct has the strongest relation with IRT-based difficulty estimates, followed by learner feedback, the Elo rating system, expert rating and finally paired comparison.

Key Words and Phrases: IRT, proportion correct, learner feedback, expert rating, paired comparison, graded response model and Elo rating

1. INTRODUCTION

Most e-learning environments are static, in the sense that they provide for each learner the same information in the same structure using the same interface. One of the recent tendencies is that they become dynamic or adaptive. 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 curriculum/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, a feature which is already present to a certain extent in computerized adaptive tests (CATS; Wainer 2000).

A prerequisite for adaptive item sequencing is to have items with a known difficulty level. Therefore, an initial development of an item bank with items of which the difficulty level is known is needed. This item bank should be large enough to include at any time an item with a difficulty level within the optimal range that has not yet been presented to the learner. In CAT, the item response theory (IRT; Van der Linden & Hambleton 1997) is often used to

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

generate such a calibrated item bank. IRT is a psychometric approach that emphasizes the fact that the probability of a discrete outcome, such as the correctness of a response to an item, is function of qualities of the item and qualities of the person. Various IRT models exist, differing in degree of complexity, with the simplest IRT model stating that a person's response to an item depends on the person's proficiency level and the item's difficulty level. More complex IRT models include additional parameters, such as an item discrimination parameter and a guessing parameter. Obtaining a calibrated item bank with reliable item difficulty estimates by means of IRT requires administering the items to a large sample of persons in a non-adaptive manner. The sample size recommended in the literature varies between 50 and 1000 persons (e.g. Kim 2006, Linacre 1994, Tsutakawa & Johnson 1990). Because IRT has been a prevalent CAT approach for decades, it seems logical to apply IRT for adaptive item sequencing in learning environments that consist of simple items. However, the difference in data gathering procedure of learning and testing environments has implications for IRT application in learning environments. In many learning environments, the learners are free to select the item they want to make. This combined with the possibly vast amount of items provided within the learning environment leads to the finding that many exercises are only made by few learners (Wauters et al. 2010). Even though IRT can deal with structural incomplete datasets (Eggen 1993), the structure and huge amount of missing values found in the tracking and logging data of learning environments can easily lead to non-converging estimations of the IRT model parameters. In addition to this, the maximum likelihood estimation procedure implemented in IRT has the disadvantage of being computationally demanding.

Due to these impediments that go together with IRT based calibration, we are compelled to search for alternative estimation methods to estimate the difficulty level of items. Some researchers have brought up alternative estimation methods. However, the accuracy of some solutions were not compared to IRT based calibration and the various methods were not compared in a single setting. The purpose of this study is to review the accuracy of some alternative estimation methods as compared to IRT-based calibration in a single setting.

2. EXPERIMENT

2.1 Related Work

2.1.1 *Item Response Theory.* To estimate the item difficulty, 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), called the item characteristic function:

$$\frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}$$

The IRT-based estimation of the difficulty level will be estimated on the basis of the learners' data obtained in this study. In addition to that, IRT-based calibration conducted on preliminary examinee data by Selor, the selection agency of the Belgian government, serves as true difficulty parameter values.

2.1.2 *Proportion Correct.* A simple approach to estimate the difficulty level of items is to calculate the proportion of correct answers by dividing the number of learners who have answered the item correctly by the number of learners who have answered the item. To obtain the item difficulty parameter, the proportion correct scores has to be converted as follows:

$$\beta_i = \log\left[\frac{1 - \frac{n_i}{N_i}}{\frac{n_i}{N_i}}\right],$$

where β_i denotes the item difficulty level of item i, n_i represents the number of learners who have answered item i correctly, and N_i represents the number of learners who have answered item i.

The advantage of this approach is that the item difficulty can be calculated online due to the easy formula which does not require many computational resources. Furthermore, the item difficulty can be updated after each administration. The lower the proportion of students who have answered the item correctly, the more difficult the item is. Johns, Mahadevan and Woolf (2006) have compared the item difficulty level obtained by IRT estimation with the percentage of students who have answered the item incorrectly, and found a high correlation (r=0.68).

2.1.3 *Learner Feedback.* Some researchers have applied learner's feedback in order to provide adaptive sequencing of courseware in e-learning environments (e.g. Chen, Lee & Chen 2005, Chen, Liu & Chang 2006, Chen & Duh 2008). After a learner has studied a particular course material, he is asked to answer two simple questions: "Do you understand the content of the recommended course material?" and "How do you think about the difficulty of the course materials?". After a learner has given feedback on a 5-point Likert scale, scores are aggregated with those

of other learners who previously answered this question by taking the average of the scores. The new difficulty level of the course material is based on a weighted linear combination of the course difficulty as defined by course experts and the course difficulty determined from collaborative feedback of the learners. The difficulty parameters slowly approach a steady value as the number of learners increases.

In this study the procedure of Chen et al. (2005) for adjusting the difficulties of the items is slightly altered. The course difficulty as defined by course experts is not taken into account. Instead, the difficulty estimates are solely based on the collaborative feedback of the learners. After an item is presented, the learner is asked a feedback question "How difficult did you find the presented item?". The learner answers on a 5-point Likert scale (Likert, 1932), ranging from -2 ("very easy") over -1 ("easy"), 0 ("moderate"), 1 ("difficult") to 2 ("very difficult"). The item difficulty based on learner feedback is then given by the arithmetic mean of the scores.

2.1.4 *Paired Comparison.* Another method, already used in CAT, to estimate the difficulty level of new items is paired comparison (Ozaki & Toyoda 2006, 2009). In order to prevent content leaking, experts are asked to assess the difficulty of items through one-to-one comparison or one-to-many comparison. In this method, items for which the difficulty parameter has to be estimated, are compared with multiple items, of which the item difficulty parameter is known. The underlying thought that prompts this item difficulty estimation approach is Thurstone's paired comparison model. While Thurstone (1994) modelled the preference judgment for object i over object j, Ozaki and Toyoda (2006, 2009) modelled the difficulty judgment of item i over item j.

In this study a similar procedure of the one employed by Ozaki and Toyoda (2009) is adopted to estimate the difficulty level by means of paired comparison. After an item is presented, the learner has to judge where the presented item should be located in a series of 11 items ordered by difficulty level from easy to difficult. This means that the raters have to make a one-to-many comparison with 11 items of which the item difficulty parameter is known. The probability that item i is more difficult than item 1, according to N raters is expressed as:

$$P_1(\beta_i) = \frac{1}{1 + \exp[-1(\beta_i - b_1)]},$$

Where β_i is the difficulty of item i judged by the raters, b_1 is the difficulty parameter of item 1 as estimated by the preliminary IRT analysis, conducted by Selor.

In this study 11 items are presented simultaneously and the raters have to select one out of 12 categories: $i<1$, $1<i<2$,…, $10<i<11$, $11<i$. Because the 11 items are ordered according to their difficulty level from easy to difficult, the idea of the graded response model (Samejima, 1969) can be adopted to extract the boundary response function of each category as:

$P_{i<1}(\beta_i) = 1 - P_1(\beta_i)$
$P_{1<i<2}(\beta_i) = P_1(\beta_i) - P_2(\beta_i)$
…
$P_{10<i<11}(\beta_i) = P_{10}(\beta_i) - P_{11}(\beta_i)$
$P_{11<i}(\beta_i) = P_{11}(\beta_i)$

The final estimation of β_i is obtained by maximizing the log likelihood, while fixing b_js, i.e. the difficulty parameters of item 1 to 11 as estimated by the preliminary IRT analysis.

2.1.5 *Expert Rating.* Another approach to obtain item parameter estimates is allowing subject domain experts to estimate the value of the difficulty parameter (Yao 1991, Linacre 2000, Fernandez 2003, Lu, Li, Liu, Yang, Tan & He 2007). There is some evidence in the measurement literature that test specialists are capable of estimating item difficulties with reasonable accuracy (e.g., Chalifour & Powers 1989), although other studies found contradictory results (Hambleton, Bastari & Xing 1998). As indicated by Impara and Plake (1998), a distinction has to be made between the ability of experts to rank order items accurately with reference to the difficulty level, and the ability of experts to estimate the proportion of persons who will answer the items correctly. Experts seem to be capable conducting the former task, but have difficulties conducting the latter where they have to be able to conceptualize the reference group and predict how well such persons will perform on each item.

Hence, two methods for obtaining expert ratings were included in this study: a paired comparison method and an evaluation on a proportion correct metric. The formula's to obtain the item difficulty parameter estimates based on these two methods are described in the subsections "Paired Comparison" and "Proportion Correct" respectively.

2.1.6 *Elo Rating.* The Elo Rating approach (Brinkhuis & Maris 2010) for estimating the item difficulty level is an educational implementation of the Elo Rating system used for rating chess performances and sports (Elo 1978). In sport, for example, two players compete with each other, resulting in a win, a loss or a draw. These data are known as paired comparison data, for which the Elo rating system is developed. In the educational field, a person is seen as a player and an item is seen as its opponent. The Elo Rating formula expresses the new rating after an event as a

function of the pre-event 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. Brinkhuis and Maris (2010) estimated the expected score on the new observation by means of the Rasch model. The formula implies that when the difference between the expected score and the observed score is high, the change in both the person's knowledge level and the item difficulty level will be high. Because the estimation of the difficulty level becomes more stable when more persons have answered an item, the weight given to new observations decreases when the rating of items is based on many observations. The same is true for the rating of the persons. When the rating of the person's knowledge level is based on a large amount of answered items, the weight given to new observations decreases.

In this study, the Elo Rating system implemented by Brinkhuis and Maris (2010) was used to estimate the item difficulty level. This Elo Rating system enables continuous measurement, since the rating is updated after every event. The formula for updating the item difficulty level, and on the same time the person's knowledge level, is given by:

$$\beta_n = \beta_0 + W(Y - Y_e),$$

where β_n is the new item difficulty rating after the item is answered by a person, β_0 is the pre-event rating, W is the weight given to the new observation, Y is the actual observation (score 1 for incorrect, 0 for correct), and Y_e is the expected observation which is estimated on the basis of the Rasch model. Hence, the formula for updating the item difficulty level after a correct response becomes:

$$\beta_n = \beta_0 + W\left[0 - \frac{\exp(\theta_0 - \beta_0)}{1 + \exp(\theta_0 - \beta_0)}\right],$$

where θ_0 is the estimated person's knowledge level before that person has answered this specific item. In this study, the weight has been set to 0.4. Preliminary analysis have shown that a weight of 0.4 results in good estimates as it is not too large, resulting in too much fluctuation, and it is not too small, resulting in an nearly invariant difficulty estimate.

Next to the comparison of the different alternative estimation methods with IRT-based calibration, we are interested whether the alternative estimation methods that are based on the binary response data of the learners, i.e. 1 for a correct response and 0 for an incorrect response, are sample dependent. If the correlation between these methods and the true difficulty parameter values are lower than the correlation between these methods and the difficulty parameter values obtained on the basis of IRT-calibration with the data gathered in this study, then these alternative methods are somewhat sample dependent. Furthermore, on the basis of the study of Impara and Plake (1998) it is hypothesized that the correlation between the true difficulty parameter values and the ones obtained by means of expert rating will be lower than the correlation between the true difficulty parameter values and the ones obtained by means of paired comparison conducted by the experts.

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. Three hundred eighteen students decided to participate. Sixteen teachers French from the above mentioned educational programs were contacted as experts. Thirteen experts decided to participate.

2.2.1 Material and Procedure. The study took approximately half an hour. The learning material consisted of items on French verb conjugation, supposedly measuring one single skill. The instructions, consisting of information on the login procedure for the learning environment and on the proceedings of the experimental study were sent to the participants by email. Once logged into the learning environment, the procedure for students was different from the procedure for experts.

Students were given an informed consent. Next, they completed the pretest used as an example. This pretest consisted of one item with three subquestions. First, the student had to fill in the correct French verb conjugation. Second, the student was asked: "How difficult did you find the previous item?" and the student has to answer on a 5 point Likert scale, ranging from -2 ("Very easy") to 2 ("Very difficult"). Finally, the student was asked to judge where the presented item should be located in the given series of 11 items ordered by difficulty level from easy to difficult. After the pretest sample, students completed the actual test, which consisted of 25 items each with three subquestions.

Experts completed the pretest used as an example. This pretest consisted of one item with three subquestions. First, the expert had to fill in the correct French verb conjugation. Second, the expert was asked: “What is, according to you, the percentage of students that will answer this item correctly after completing secondary education?”. Finally, the expert was asked to judge where the presented item should be located in the presented series of 11 items ordered by difficulty level from easy to difficult. After the pretest sample, experts completed the actual test, which consisted of 25 items each with three subquestions.

2.3 Results

The inter-rater agreement for the classification of the item difficulty was calculated by means of the intraclass correlation coefficient (ICC; Shrout & Fleiss 1979). Shrout and Fleiss (1979) report the magnitude for interpreting ICC values where ICC<0.40 = ”poor”, 0.40≤ICC≤0.59 = “fair”, 0.60≤ICC<0.74 = “good”, and ICC≥0.74 = “excellent”. The inter-rater agreement for the classification of the item difficulty by students was fair (ICC[3,1]=0.42 for learner feedback; ICC[3,1]=0.43 for paired comparison). The inter-rater agreement for the classification of the item difficulty by experts was good (ICC[3,1]=0.68 for expert rating and for paired comparison). The inter-rater agreement for the classification of the item difficulty for paired comparison by experts and learners combined was fair (ICC[3,1]=0.44). The inter-rater agreement, when considering the mean of the paired comparison feedback given by learners and the mean of the paired comparison feedback given by experts, was excellent (ICC[3,1]=0.88).

The criterion used to evaluate the efficacy of the item difficulty estimation methods was the Pearson correlation between the estimated item parameter and its corresponding true parameter. The true difficulty parameter value for each item was estimated in advance by Selor, using examinee data for conducting the IRT analysis. Additionally the Pearson correlation was measured between the estimated item parameter and its corresponding IRT difficulty parameter value based on calibration with the data gathered in this study. The Pearson correlation between the estimated item difficulty parameter and the true item difficulty parameter is a measure for the strength of their linear relationship.

Detailed correlation results for the item difficulty estimates are shown in table I.

Table I. Pearson correlation matrix of the item difficulty estimates for the different estimation methods.

Item Difficulty Estimation Method	**Item Difficulty Estimation Method**							
	True β	**IRT-Study**	**Proportion Correct**	**Learner Feedback**	**Expert Rating**	**Paired Comparison (Learner)**	**Paired Comparison (Expert)**	**Elo Rating**
True β	1.00							
IRT-Study	.90	1.00						
Proportion Correct	.90	1.00	1.00					
Learner Feedback	0.88	0.88	0.88	1.00				
Expert Rating	0.80	0.80	0.80	0.95	1.00			
Paired Comparison (learner)	0.62	0.50	0.50	0.58	0.55	1.00		
Paired Comparison (Expert)	0.56	0.44	0.44	0.53	0.51	0.98	1.00	
Elo Rating	0.85	0.92	0.92	0.81	0.73	0.45	0.39	1.00

The results of the Pearson correlation between the estimated item difficulty parameter and the true item difficulty parameter indicates that proportion correct has the strongest relation ($r(23)=0.90$, $p<0.01$), followed by learner feedback ($r(23)=0.88$, $p<0.01$), Elo rating ($r(23)=0.85$, $p<0.01$), expert rating ($r(23)=0.80$, $p<0.01$), paired comparison based on learners' feedback ($r(23)=0.62$, $p<0.01$) and paired comparison based on expert data ($r(23)=0.56$, $p<0.01$). The Pearson correlation between the estimated item difficulty parameter and the difficulty parameter estimated by means of IRT with the data of the 318 students in this study shows similar results. The correlation with proportion correct is the highest ($r(23)=1.00$, $p<0.01$), followed by Elo rating ($r(23)=0.922$, $p<0.01$), learner feedback ($r(23)=0.88$, $p<0.01$), expert rating ($r(23)=0.80$, $p<0.01$), paired comparison based on learners' feedback ($r(23)=0.50$, $p<0.05$) and finally paired comparison based on expert data ($r(23)=0.44$, $p<0.05$).

The difference between the correlation coefficient of proportion correct with the true difficulty parameter value and the correlation coefficient of proportion correct with the difficulty parameter value estimated by means of IRT with the data of this study is significant ($t(22)=-19.18$, $p<0.05$). The difference between the correlation of the Elo rating system with the true difficulty parameter value and the correlation of the Elo rating system with the difficulty parameter value estimated by means of IRT with the data of this study is also significant ($t(22)=-2.09$, $p<0.05$). The correlation coefficient of proportion correct with the IRT calibration based on the study data differs significantly from the correlation coefficient of the Elo rating system with the IRT calibration based on the study data ($t(22)=20.7485$, $p<0.05$). The significance disappears when proportion correct and the Elo rating system are compared with the true difficulty parameter value ($t(22)=1.46$, $p=0.16$).

There is no significant difference between the correlation of the true item difficulty parameter values with the ones obtained by means of expert rating, and the correlation of the true item difficulty parameter values with the ones obtained by means of paired comparison based on expert ratings ($t(22)=1.89$, $p=0.07$). The difference between the correlation coefficient of learner feedback with the true difficulty parameter value and the correlation coefficient of expert rating with the true difficulty parameter value is significant ($t(22)=2.71$, $p<0.05$). However, the difference between the correlation coefficient of paired comparison based on learner feedback with the true difficulty parameter value and the correlation coefficient of paired comparison based on expert rating with the true difficulty parameter value is not significant ($t(22)=1.85$, $p=0.08$).

3. DISCUSSION

As the tracking and logging data of many learning environments fail to contain the required amount and structure of data needed for IRT estimation, this article searches for appropriate alternative methods to estimate the difficulty level of items. Based on the response data and the judgment data of a sample of learners and experts, the difficulty level of twenty five items was estimated by means of six estimation methods: (1) IRT calibration based on the study data, (2) proportion correct, (3) learner feedback, (4) expert rating, (5) paired comparison (based on learners' judgment and based on experts' judgment), and (6) the Elo rating system.

The findings indicate that proportion correct has the strongest relation with the true difficulty parameter values, followed by learner feedback, the Elo rating system, expert rating and paired comparison. Furthermore, proportion correct also has the strongest relation with the difficulty estimates obtained with IRT calibration on the study data, followed by the Elo rating system, learner feedback, expert rating and paired comparison. Considering the alternative estimation methods that are based on the binary response of the learners (correct vs. incorrect response to an item), it is shown that IRT calibration, proportion correct and the Elo rating system do not differ. The high correlation found between IRT calibration (both true difficulty parameter and IRT calibration on the study data) and proportion correct is not surprising as the total score is a sufficient statistic for the Rasch model. Furthermore, it is clear that proportion correct and the Elo rating system are sample dependent as they correlate higher with the IRT calibration on the study data than with the true difficulty parameter values.

Results contradict the postulation of Impara and Plake (1998) that experts perform better in estimating the difficulty by rank ordering the items than by estimating the proportion of persons who will answer the items correctly. Furthermore, findings indicate that learners perform better on judging the difficulty of items than experts. However, this difference disappears when learners and experts need to rank order the items according to their difficulty level. It needs to be considered that the estimation by means of learner feedback is based on a larger sample than the estimation by means of expert rating, which could explain the difference between learner feedback accuracy and expert rating accuracy. The finding that the correlation of paired comparison with the true difficulty parameter is moderate could be due to the small sample size, resulting in some outlier estimations. The paired comparison data are analyzed by means of the graded response model, which is a more complex IRT model than the Rasch model, and hence may need a larger sample size to obtain reliable item difficulty estimates.

Even though this study indicates that the difficulty of items can be estimated on the basis of alternative estimation methods, it should be considered that the size of the item set that was used to compare the alternative estimation methods was rather small. We recognize that a total number of twenty five items is limited, but considering raters fatigue, we were compelled to keep the item set rather small. Furthermore, we made sure that the twenty five items covered a broad range of difficulty.

Future research will focus on the sample size requirement for reliable difficulty estimates. The different alternative estimation methods will be compared for different sample sizes. If results would indicate that alternative estimation methods provide reasonable accurate difficulty level estimates, these estimation methods could be used to provide adaptive curriculum sequencing. Those alternative estimation methods could also be used to make IRT estimation more efficient by using the estimates as prior in a Bayesian estimation method. A limitation of this study, which should be tackled in future research, is the fact that even though some of the alternative item difficulty estimation methods seem to be a viable alternative for IRT-based calibration in this study, no generalization can yet be made to other domains and to items requiring more than one skill.

REFERENCES

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, Arhnem.

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

CHALIFOUR, C.L., AND POWERS, D.E. 1989. The Relationship of Content Characteristics of GRE Analytical Reasoning Items to Their Difficulties and Discriminations. *Journal of Educational Measurement*, 26, 120-132.

CHEN, C.M., LEE, H.M., AND CHEN, Y.H. 2005. Personalized e-Learning System Using Item Response Theory, *Computers & Education*, 44, 237-255.

CHEN, C.M., LIU, C.Y., AND CHANG, M.H. 2006. Personalized Curriculum Sequencing Utilizing Modified Item Response Theory for Web-Based Instruction. *Expert Systems with Applications*, 30, 378-396.

CHEN, C.M., AND DUH, L.J. 2008. Personalized Web-Based Tutoring System Based on Fuzzy Item Response Theory. *Expert Systems with Applications*, 34, 2298-2315.

EGGEN, T.J.H.M. 1993. Itemresponstheorie en onvolledige gegevens. In Eggen, T.J.H.M., Sanders, P.F. (Eds) *Psychometrie in de Praktijk*, Cito, Arhnem.

ELO, A.E. 1978. *The rating of chess players, past and present*, B.T. Batsford Ltd., London.

FERNADEZ, G. 2003. Cognitive Scaffolding for a Web-Based Adaptive Learning Environment. *Advances in Web-Based Learning - Icwl 2003, Proceedings*, 2783, 12-20.

HAMBLETON, R.K., BASTARI, AND XING, D. 1998. Estimating Item Statistics. Report No.298, Laboratory of Psychometric and Evaluative Research, School of Education, University of Massachusetts.

IMPARA, J.C., AND PLAKE, B.S. 1998. Teachers' Ability to Estimate item Difficulty: a Test of the Assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement*, 35(1), 69-81.

JOHNS, J., MAHADEVAN, S., AND WOOLF, B. 2006. Estimating Student Proficiency Using an Item Response Theory Model. *Intelligent Tutoring Systems, Lecture Notes in Computer Science*, 4053, 473-480.

KIM, S. 2006. A Comparative Study of IRT Fixed Parameter Calibration Methods. *Journal of Educational Measurement*, 43, 355-381.

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.

LIKERT, R. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22, 1-55.

LINACRE, J.M. 1994. Sample Size and Item Calibrations Stability. *Rasch Measurement Transaction*, 7(4), 328.

LINACRE, J.M. 2000. Computer-Adaptive Testing: A Methodology whose Time has Come. In Chae S., Kang U., Jeon E., Linacre J.M. (Eds) *Development of Computerized Middle School Achievement Test*, Komesa Press, Seoul.

LU, F., LI, X., LIU, Q.T., YANG, Z.K., TAN, G.X.,AND HE, T.T. 2007. Research on Personalized e-Learning System Using Fuzzy Set Based Clustering Algorithm. *Computational Science - ICCS 2007, Part 3, Proceedings*, 4489, 587-590.

OZAKI, K., AND TOYODA, H. 2006. Paired Comparison IRT Model by 3-Value Judgment: Estimation of Item Parameters Prior to the Administration of the Test. *Behaviormetrika*, 33, 131-147.

OZAKI, K., AND TOYODA, H. 2009. Item Difficulty Parameter Estimation Using the Idea of the Graded Response Model and Computerized Adaptive Testing. *Japanese Psychological Research*, 51, 1-12.

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.

SAMEJIMA, F. 1969. Estimation of Latent Trait Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph*, 17.

SHROUT, P.E., AND FLEISS, J.L. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86, 420-428.

THURSTONE, L.L. 1994. A Law of Comparative Judgment. *Psychological Review*, 101(2), 266-270.

TSUTAKAWA, R.K., AND JOHNSON, J.C. 1990. The Effect of Uncertainty of Item Parameter Estimation on Ability Estimates. *Psychometrika*, 55, 371-390.

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

WAINER, H. 2000. *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.

YAO, T. 1991. CAT with a Poorly Calibrated Item Bank. *Rasch Measurement Transactions*, 5, 141.