An Interpretable Statistical Ability Estimation in Web-based Learning Environment

Yi-Ting Huang[a*], Meng Chang Chen[b], Yeali S. Sun[a]
[a] *Department of Information Management, National Taiwan University, Taiwan.*
[b] *Institute of Information Science, Academia Sinica, Taiwan.*
*d97008@im.ntu.edu.tw

Abstract: With growing interest in estimating true ability in contemporary learning, the demand for personalized learning and Web-based learning environments has become increasingly important. This paper develops a statistical and interpretable method of estimating ability. This method captures the succession of learning over time and provides an explainable interpretation of a statistical measurement, based on Item Response Theory and the quantiles of acquisition distributions. The results from the simulation and empirical study demonstrate that the estimated abilities can successfully recognize the actual abilities of students. The correlation values between the estimated abilities and the post-test score, which incorporate this testing history, are higher than values that only consider test responses at the time of testing. Furthermore, the pre-test and post-test administered to the experimental group show significant student improvement. These results suggest that this method serves as a successful alternative ability estimation and provides a better understanding of student competence.

Keywords: Ability estimation, Adaptive test, Item response theory

1. Introduction

Recently, theories on learning have focused increasing attention on understanding and measurement student ability. Vygotsky [12] states that a learner's ability in the Zone of Proximal Development (ZPD)—the difference between a learner's actual ability and his or her potential development—can progress well with external help. Instructional scaffolding [11], closely related to the concept of ZPD, suggests that an appropriate support during the learning process helps learners achieve their stated goals. Effective instructional support requires identifying a student's prior knowledge, tailoring an aid to meet their initial needs, and then removing this aid when he or she acquires sufficient knowledge.

Nowadays, estimations of ability offer extensive applications within e-learning systems in various domains. For example, Chen et al. [4] considered a learner's ability for recommending personalized learning paths in a Web-based programming learning system, while Chen and Chung [3] analyzed students' understanding by suggesting English vocabulary on mobile devices. Similarly, within Computerized Adaptive Testing (CAT), Barla et al. [1] calculated an examinee's ability to select suitable questions. All of these studies used Item Response Theory (IRT) to estimate a student's ability, and their results demonstrated improved student performance.

Item Response Theory is a modern theory of testing that examines the relationship between an examinee's responses and items related to abilities measured by the items in the test. Three well-known ability estimations proposed by IRT are maximum likelihood estimation, maximum a posteriori and expected a posteriori [6]. Examples of this research include [1], where researchers used expected a posteriori to score each examinee's ability at

each time of a test. However, IRT has some disadvantages. First, every exercise performed by a student is recorded in most of the web-based learning environments listed above; however, the ability estimations of IRT only consider test responses at the time of testing, rather than incorporating testing history. Moreover, the interpretation of the result of estimating an examinee's ability is often defined in terms of the acquisition of a large portion of knowledge of the specific ability itself—through a test. Unfortunately, this definition is qualitative rather than quantitative.

In response to these issues, this paper proposes a statistical method and a novel interpretation of estimating ability with inherent randomness in the acquisition process. We conduct a simulation study to investigate the property of the proposed approach and an empirical study to evaluate practical performance. Our simulation results demonstrate the convergence between an examinee's current grade and his or her actual ability. We also implement this method on a Web-based learning environment. The empirical results find a strong correlation between the estimated ability and the post-test score that incorporates this testing record, and this correlation is higher than correlations between ability and values that only examine test responses at the time of testing. Moreover, the pre-test and post-test administered to the experimental group demonstrate significant student improvement.

The remainder of this paper is organized as follows. In Section 2, we present the proposed ability estimation. Section 3 reports a simulation and Section 4 contains the empirical procedure and results. Finally, Section 5 summarizes our conclusions.

2. Method

We propose the following interpretation of the quantitative definition: an examinee is said to have ability θ if s percent of items in a test $T = (t_1, \ldots, t_m)$ have been correctly answered each by r percent of the population.

We first consider that each item t_i in a test T has been correctly answered by r percent of the population. In general, there is a specific knowledge behind each tested item t_i. The level of the specific knowledge represents that most people have acquired knowledge of t_i. Most people understand some knowledge at an early age, whereas some understand this knowledge later in life. Here, we precisely denote the level the specific knowledge represents as the age at which r percent of the population has acquired knowledge of t_i, where age can refer to school grades or lifetime. When given a knowledge t_i and a population, the probability distribution of knowledge acquisition $p_t(\theta)$ can be calculated. Let the quantile function q_t of the cumulative distribution function correspond to the acquisition distribution p_t. In other words, $q_t(r)$ represents the age at which r percent of the population has acquired knowledge of t. This assumes a normal distribution,

$$q_t(r) = \mu_t + \Phi^{-1}(r)\sigma_t \quad (1)$$

where μ_t and σ_t represent the mean and standard deviation of the distribution p_t, and $\Phi^{-1}(r)$ is a quantile function representing the probability of exactly r to fall inside the interval of the distribution. When an examinee correctly responds to the item t_i, the examinee's ability is regarded as the age or grade level, etc. To investigate the distribution of the grade level of a test T, we collect the grade level values generated from each quantile function $q_t(r)$ as the distribution of knowledge acquisition within a single test f_Q.

In practice, this is time consuming and costly for each item t_i known in advance by the distribution p_t. Fortunately, under Item Response Theory [6], a response of an examinee to an item is modeled by a mathematical item response function, known as the item characteristic curve. The item characteristic curve is a mathematical family model that describes the probability of a correct response between an examinee's ability and the item parameters. These models employ one or more parameters, such as an item difficulty

parameter and an item discrimination parameter, to define a particular cumulative form. When given the item parameters, the grade level at which *r* percent of the population correctly responds to item *t* can be inferred. Take one-parameter logistic model as an example,

$$q_t(r) = \ln(\frac{r}{1-r}) + b \qquad (2)$$

where variable *b* as item difficulty.

Estimating an examinee's ability through a test relies on the test responses of the test. We consider a percentage of correct responses in a test as variable *s* and define the *s*th quantile of the distribution of knowledge acquisition in a test f_Q as the examinee's ability. The distribution of the *s*th quantile of f_Q, where *s* percent of items in a test have been correctly answered by *r* percent of the population, can be performed using a standard formula for normal approximation of order statistics [5]:

$$q_T(r,s) \sim N(F_Q^{-1}(s), \frac{s(1-s)}{m[\, f_Q(F_Q^{-1}(s))]^2}) \qquad (3)$$

where F_Q is the cumulative distribution function and *m* is the number of items in a test. This result is more certain of the estimated grade level assigned to a large sample item size. In cases where an examinee correctly answered all items or no item, a smooth constant *c* is used (*c=0.01* in this study).

When given an examinee's responses in a test, the current examinee's ability θ_t can be described by the distribution (3) in which *r* percent of the population correctly answer *s* percent of items. We also consider an examinee's history record, and employ Exponential Moving Average (EMA) [2] to combine this history with the current ability, transformed by the following formula:

$$ability_t = \alpha \times \theta_t + (1-\alpha) \times ability_{t-1} \quad (4)$$

where θ_t is the current ability in time *t* obtained from the mean of the equation (3), $ability_{t-1}$ is the past estimated ability in the time *t-1* as history records, and $ability_t$ is the final estimated ability in time *t* after the combination of the current ability and the past estimated ability with EMA. Additionally, $\alpha = 2/(n+1)$ is a smoothing constant represented as an exponential weight, and *n* represents the period as the length of the moving window.

3. Simulation

3.1 Settings

To understand the performance of the proposed method, we conducted a simulation. According to a one-parameter logistic model in Item Response Theory [6], the probability of correct response is *0.5* when an item difficulty is equal to an examinee's ability. In the simulation, we referred to this probability for setting the variable *r*. Moreover, the item response model also provides information in the estimation of the variable *s*. We used a one-parameter logistic model to predict the probability of a correct response when given the ability and an item, and then conditionally randomly sampled the variable *s*.

In each simulation, ten items were generated according to an examinee's ability at the time. The distribution of difficulty of these items acts as a normal distribution. For example, given an examinee's ability $\theta=3$, the difficulties of a test are {2, 2, 3, 3, 3, 3, 3, 3, 4, 4}. Ability and difficulty in this study range from one to six, corresponding to the school grades. In practice, an examinee's school grade is considered as their initial ability, and the ability is updated by responses in each test. Thus, the simulation starts with any grade ranging from one to six in order to simulate different grade students with various abilities, and then terminates 100 iterations after the convergence point. We found the convergence point and

then counted the Root Mean Square Error (RMSE) during the 100 iterations. The definition of the convergence point is determined by computing the difference between the estimated ability and the ground truth, and the difference value is continuously four times smaller than a threshold (*thd = 0.25* in the simulation). Each simulation was processed 1000 times. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{k}\sum_{i}(\theta_i - \hat{\theta}_i)^2} \qquad (5)$$

where θ is the actual ability as the ground truth, $\hat{\theta}$ is the estimated ability, k is the number of the iterations. Here, k=1000. This metric represents the average distance between the ground truth and the generated results. The smaller RMSE value indicates that the estimated ability is close to the ground truth. In addition, we also discuss the parameter α in equation (4). The parameter is presented in terms of n time periods and represents the weight of the observation at the present time. The variable n was set from one to twelve.

3.2 Results

Table 1 shows the average convergence points in the number of variable n of parameter α in equation (4) over the degree of difference between the estimated ability and ground truth, and the results of RMSE during the 100 iterations after the convergence points. It is clear that the proposed method can successfully estimate abilities in the finite iterations. Specifically, an examinee's ability can be estimated more precisely when he or she continues to have more tests. Furthermore, the error distances between the estimated abilities and the ground truths are low enough to be acceptable after convergence. That is, an examinee's ability can be steadily measured during a long-term observation.

The parameter $\alpha = 2/(n+1)$ in the equation (4) is an exponential weight of the current ability, and n represents the number of time periods, such as times or days, taken into consideration. When $n=1$, it represents that an examinee's ability only considers the current estimated ability without the history record. In Table 1, the values in screentone present that the average convergence points are fewer than the points generated from $n=1$. This result shows that the estimated abilities are quickly found and the error distances decrease when considering the history record. In particular, it is apparent when the initial grade is equal to the ground truth. When n is small (e.g. $n=2, \alpha = 2/3$; $n=3, \alpha = 1/2$), the estimated ability is mainly decided by the current ability. The convergence points are smallest and the RMSE is slightly smaller than one generated from $n=1$. In contrast, when n increases, the estimated ability is principally composed of abilities from the past to now. If an examinee's initial ability is not close to his or her actual ability, it takes more information to accurately estimate. Although it takes time, the RMSE is clearly shrinking.

Table 1. The results of convergence point and RMSE (each row represents the degree of difference between the initial ability and the actual ability, and each column represents the number of time periods considered by the exponential weight of the current ability)

d \ n	1	2	3	4	5	6	7	8	9	10	11	12
0	20.61	13.88	11.72	11.53	10.98	10.90	10.26	10.52	10.16	10.35	10.18	10.04
1	21.96	16.17	15.74	16.31	17.40	19.07	20.43	22.29	23.98	25.45	26.92	28.42
2	22.91	18.08	18.54	19.91	21.90	24.18	26.64	29.06	31.50	33.53	35.62	38.58
3	23.86	19.67	19.91	21.91	24.59	27.62	30.33	32.90	35.74	38.43	41.52	44.13
4	24.30	20.73	21.52	23.51	26.71	29.68	32.96	36.00	40.19	42.83	45.45	48.65
5	24.50	21.41	22.66	25.22	29.10	31.92	35.97	38.22	42.62	46.40	49.18	53.12
RMSE	0.39	0.32	0.28	0.26	0.24	0.23	0.22	0.21	0.20	0.19	0.19	0.18

Consider a dramatic example to explain the properties of the proposed method. Assume that a first grade student, whose real ability is the sixth grade, learns and has a test in a web-based learning system once a day. Figure 1 illustrates the changes in the estimated ability computed from the proposed method in different weights. The black horizontal line at the sixth grade represents the student's actual ability as the ground truth. The other curves depict the estimated abilities under the different weights: a red dotted line, *n=1*; a green solid line, *n=3*; a purple solid line, *n=6*; and a blue solid line, *n=12*. The mark labels on each line are the convergence points (the value is continuously four times smaller than *thd = 0.25*). It is clear that the estimated abilities are converging as *n* decreases in size. Although these estimated abilities are estimated using few iterations when *n=1*, the red-dotted line drastically fluctuates after the convergence point. In other words, if the ability estimation only takes the current responses into consideration, instead of past performance, the variance of every estimated ability may be large. In this situation, question selection in a test using inaccurate ability estimation could result in confusion by the examinee. In contrast, the estimated error gradually decreases when *n>1*, even though the estimated abilities when *n=1* take more time to estimate. In this situation, the students' abilities were gradually updated and the difficulties of items incrementally increased. This is thus a trade-off problem between time and precision.

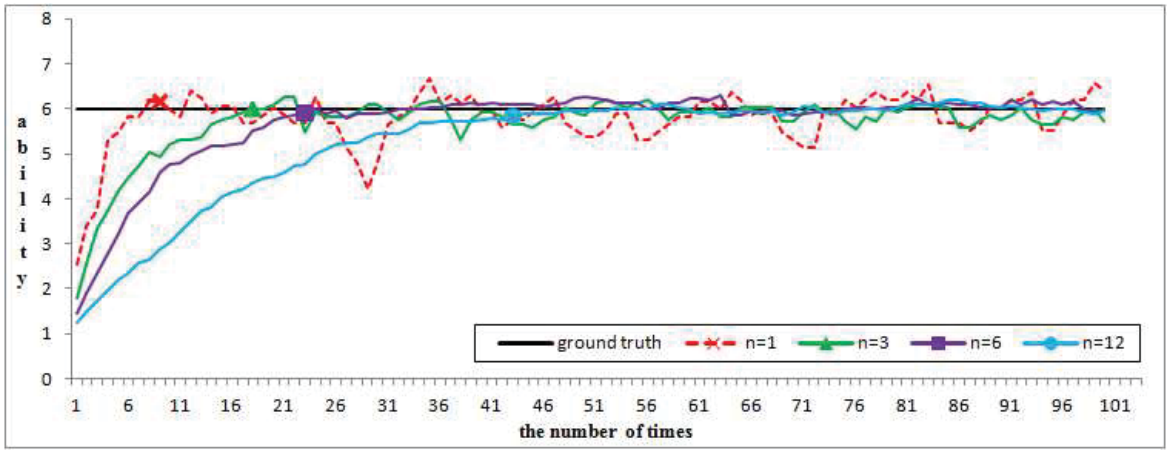

Figure 1. The changes in the estimated ability computed from the proposed method for the different weights (*n=1, n=3, n=6, n=12*)

4. Empirical Study

4.1 Materials

The measurement approach proposed in this study is implemented on a Web-based learning system developed by the AutoQuiz Project of the IWiLL learning platform [8]. It provides English language learners online English reading materials collected from up-to-date online news websites and multiple-choice tests and automatically generates related quiz material [9][10]. Each test was composed of ten vocabulary questions, five grammar questions, and three reading comprehension questions. A total of 2,425 items were automatically generated based on 72 reading materials. The grade level of the vocabulary and grammar questions are defined according to the semester of high school in which the correct answer is taught, while the difficulty of the reading comprehension questions are measured by a reading difficulty estimation [7]. In other words, the grade level in this experiment is defined from one to six, corresponding to the six semesters of senior high school.

4.2 Participants and Procedure

The participants in this study were high school students in Taiwan, divided into two groups: a control group where ability is estimated only based on current responses, and an experimental group that incorporates the history record into the current ability estimation. 30 students participated within the control group, while 47 students participated in the experimental group.

The experiment was held from January 30th to March 4th, 2012. During the experiment, the subjects were asked to participate in twelve activities, consisting of reading an article and then taking a test. In each activity, the subjects in both groups received an up-to-date article and a series of quizzes automatically generated based on their abilities. In addition, there was a pre-test and post-test for evaluating their abilities as the ground truth. The variable r was set as 0.5 based on IRT, and the variable s defined as the percentage of correctly answered items. Furthermore, the parameter $n=12$ in the exponential weight of the experimental group was equal to the period of activity, because all test records were taken into consideration.

4.3 Results

To validate the accuracy of the proposed ability estimation, the subjects' abilities in the two groups were estimated with twelve continuous activities. Table 2 reports the Pearson's correlation coefficient between the estimated abilities (the estimated grade is rounded by the estimated score) and the post-test scores among the three quiz types. All of the measures are significantly positively correlated. The results in the experimental group ranged from *0.44* to *0.69*, while ones in the control group ranged from *0.47* to *0.54*. Most of the correlation values in the experimental group are higher than the values in the control group; this suggests that estimating ability with the history record leads to a clearer relationship between the estimated ability and the ground truth.

Table 2. The correlation result between the estimated ability and the post-test in the control group and the experimental group

	vocabulary		**grammar**		**reading comprehension**	
	score	grade	score	grade	score	grade
Control group	0.47*	0.49**	0.54**	0.51**	0.54**	0.47*
Experimental group	0.51***	0.44**	0.55***	0.55***	0.69***	0.65***

*$p<0.05$, **$p<0.01$, ***$p<0.001$

Comparing the post-test score in each estimated ability (grade) is another way to assess the accuracy of the proposed ability estimation. If the estimated abilities are accurate, the subject performance of each ability will differ from that of other abilities. Table 3 presents the mean post-test score of the subjects of different estimated abilities between the control group and the experimental group. Intuitively, a subject estimated a higher ability should has higher post-test score than one estimated a lower ability. One-way Analysis of Variance revealed that there were differences in the estimated vocabulary ability ($F=5.75$, $p=0.001$), the estimated grammar ability ($F=4.71$, $p=0.003$) and the estimated reading comprehension ability ($F=5.98$, $p<0.001$) in the experimental group, while there were no statistical differences between the estimated vocabulary and grammar ability in the control group. Noticeably, although the estimated reading comprehension ability in the control group has a significant difference, the mean scores among every ability fluctuated. The bolded values in Table 3 are unreasonable, because the averaged scores of the higher estimated abilities (grade 2, grade 4 and grade 5) in the control group were lower than ones

of the lower estimated abilities (grade 1 and grade 3). Though there was an unreasonable value for grade 6 of the estimated vocabulary ability in the experimental group, this is likely because only two students were assigned to grade 6. This sample size is likely unrepresentative. Moreover, in the experimental group, a Bonferroni post hoc test indicated that the performance of the estimated ability 1 and 2 were significantly different from the estimated ability 5 and 6. This indicates that the proposed ability estimation can effectively distinguish higher ability examinees from lower ones.

Table 3. The mean post-test score of the subjects in different estimated ability groups between both groups and the result of ANOVA

Estimated ability	**Control group**			**Experimental group**		
	vocabulary	grammar	reading	vocabulary	grammar	reading
1	-	37.50	46.80	-	-	37.67
2	48.33	47.00	**40.00**	23.00	34.33	46.63
3	**38.00**	51.40	52.57	52.86	52.80	53.50
4	54.40	41.40	**41.00**	62.33	54.94	64.50
5	61.22	62.83	**32.67**	69.71	66.81	66.90
6	65.83	65.56	70.18	**57.67**	72.00	78.00
F score	2.67	2.54	**6.12*** **	5.75***	4.71**	5.98***

***p<0.01, ***p<0.001*

To further understand the impact of employing the proposed ability estimation on learners, we investigated the performance between the control group and the experimental group. In keeping with the previous results, the estimated subjects' abilities in the experimental group were more accurate than those in the control group. We assume that appropriate instructional scaffolding could help students advance their learning, when effectively identifying their abilities. Table 4 presents the descriptive statistic and results of a T-test between the pretest and post-test. The results of the independent T-test (*p=0.92* in the pre-test and *p=0.51* in the post-test) showed no significant effect on the post-test between the experimental group and the control group. It is noticeable that the average score of the experimental group in the pretest was lower than the control group, but that of the experimental group in the post-test made great progress and surpassed the control group. Additionally, the paired sample T-test showed a significant effect of the pre-test and the post-test in the experimental group (*p<0.001*), while the performance of the control group had no statistically significant effect (*p>0.05*). This indicates that the subjects in the experimental group with an appropriate support can exceed the past themselves when successfully recognizing their learning status.

Table 4. The results of the pretest and post-test between the control group and the experimental group

	Pretest		**Post-test**		**Paired sample t-test**
	mean	std.	mean	std.	
Control group	53.23	19.35	56.70	17.99	1.57
Experimental group	52.83	16.67	59.28	16.01	3.71***
independent t-test	0.20		0.66		

****p<0.001*

5. Conclusion

This work develops a statistical and interpretable method of estimated ability that captures the succession of learning over time in a Web-based test environment. Moreover, it provides an explainable interpretation of the statistical measurement based on Item Response Theory and the quantiles of acquisition distributions. The result from the simulation demonstrated that the estimated abilities obtained from the proposed method could successfully approximate the actual abilities of students, and estimated abilities can be steadily measured during long-term observation. This proposed approach was also implemented on a Web-based learning environment. The empirical results show that the correlation values incorporating this testing history were higher than the values that only consider test responses at the time of testing. Additionally, the pretest and post-test administered to the experimental group demonstrated significant student improvement. This paper presents preliminary results of a pilot experiment; future research will be further expanded to include long-term evaluation of the effectiveness of the proposed approach under changes in student learning.

Acknowledgment

This work was partially supported by National Science Council, Taiwan, with Grant No. 100-2511-S-008-005-MY3 and NSC97-2221-E-001-014-MY3.

References

[1] Barla, M., Bielikova, M., Ezzeddinne, A. B., Kramar, T., Simko, M. and Vozar, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computer & Education*, 55(2), 846-857.

[2] Brown, R. G. (2004). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Dover Publications.

[3] Chen, C.M. and Chung, C.J. (2008). Personalized Mobile English Vocabulary Learning System Based on Item Response Theory and Learning Memory Cycle. *Computers & Education*, 51 (2), 624–645.

[4] Chen, C.M. Lee, H.M. and Chen, Y.H. (2005). Personalized E-learning System Using Item Response Theory. *Computers & Education*, 44 (3), 237–255.

[5] David, H. A. and Nagaraja, H. N. (2003), *Order Statistics*. Marblehead, MA: Wiley.

[6] Embertson, S., & Resise, S. (2000). *Item response theory for psychologists*. New Jersey, USA: Lawrence Erlbaum.

[7] Huang, Y. T., Chang, H. S., Sun Y. and Chen M. C. (2011). A Robust Estimation Scheme of Reading Difficulty for Second Language Learners. *In Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies*, 58-62.

[8] Kuo, C. H., Wible, D., Chen, M. C., Sung, L. C., Tsao, N. L. and Chio, C. L. (2002). Design and Implementation of an Intelligent Web-based Interactive Language Learning System. *Journal of Educational Computing Research*, 27(3), 785 – 788.

[9] Lin, Y. C., Sung, L. C. and Chen, M. C. (2007). An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding. *In Proceedings of the 15th International Conference on Computers in Education*, 137-142.

[10] Lin, Y. T., Chen, M. C. Sun, Y. S. (2009). Automatic Text-Coherence Question Generation Based on Coreference Resolution. *In Proceedings of the 17th International Conference on Computers in Education*.

[11] Wood, D., Bruner, J.S., and Ross, G. (1976). The Role of Tutoring and Problem Solving. *Journal of Child Psychology and Psychiatry*, 17, 89-100.

[12] Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Process*. Cambridge, MA.: Harvard University Press.