Estimating student ability and problem difficulty using item response theory (IRT) and TrueSkill

Youngjin Lee
University of Kansas, Lawrence, Kansas, USA and University of North Texas, Denton, Texas, USA

Abstract

Purpose – The purpose of this paper is to investigate an efficient means of estimating the ability of students solving problems in the computer-based learning environment.

Design/methodology/approach – Item response theory (IRT) and TrueSkill were applied to simulated and real problem solving data to estimate the ability of students solving homework problems in the massive open online course (MOOC). Based on the estimated ability, data mining models predicting whether students can correctly solve homework and quiz problems in the MOOC were developed. The predictive power of IRT- and TrueSkill-based data mining models was compared in terms of Area Under the receiver operating characteristic Curve.

Findings – The correlation between students' ability estimated from IRT and TrueSkill was strong. In addition, IRT- and TrueSkill-based data mining models showed a comparable predictive power when the data included a large number of students. While IRT failed to estimate students' ability and could not predict their problem solving performance when the data included a small number of students, TrueSkill did not experience such problems.

Originality/value – Estimating students' ability is critical to determine the most appropriate time for providing instructional scaffolding in the computer-based learning environment. The findings of this study suggest that TrueSkill can be an efficient means for estimating the ability of students solving problems in the computer-based learning environment regardless of the number of students.

Keywords Problem solving, User modeling, Prediction model, Educational data mining (EDM), Log file analysis, Learning analytics (LA)

Paper type Research paper

Introduction

Recently, US Department of Education emphasized the importance of developing computer-based learning environments that can provide customized learning contents tailored to the ability of students (Bienkowski *et al.*, 2012). It is anticipated that such adaptive learning environments can maximize the learning outcome of students because students can be engaged in personalized learning activities matching their level of understanding (Tanenbaum *et al.*, 2013). To develop an adaptive learning environment, it is essential to accurately estimate the ability of students as they are engaged in various learning activities. Typically, computer-based learning environments estimate the ability of students, which is changing over time as a result of their learning, by having students solve a series of problems. The estimated ability of students can then be used to provide differentiated learning experiences.

The simplest way to estimate the ability of students solving a series of problems is to count the number of correct answers or to compute the fraction of correct answers submitted by students. Because of its simplicity, this approach is frequently used in many computer-based learning environments such as massive open online courses (MOOCs); students receive instructional supports and guidance when they submit an incorrect answer a certain number of times. However, the heuristics like this are unlikely to maximize the learning outcome of students because they do not take into account the difficulty of problems and the ability of students. When the problem is difficult, it makes sense to allow more opportunities before providing instructional supports. Likewise, we do not want to postpone providing help to academically weaker students because they are likely to get frustrated, fail the learning task and may give up their learning entirely unless they receive instructional supports and guidance in time. Moreover, the effectiveness of such heuristics has not been thoroughly investigated in the computer-based learning environment.

Item response theory (IRT) is an approach that can address the shortcomings of simple count or fraction of correct answers in estimating the ability of students solving problems in the computer-based learning environment (Baylari and Montazer, 2009; Chen *et al.*, 2005). IRT assumes that the ability of students does not change while taking a test and each problem is independent of other problems in the same test. Under these assumptions, IRT can estimate the ability of students and the difficulty of problems that are invariant to students and problems being used in estimation (Ayala, 2009). As IRT takes into account both the ability of students and the difficulty of problems, solving more difficult problems is treated differently

The current issue and full text archive of this journal is available on Emerald Insight at: **www.emeraldinsight.com/2398-6247.htm**

Information Discovery and Delivery

[DOI 10.1108/IDD-08-2018-0030]

Received 16 August 2018
Revised 20 November 2018
22 December 2018
Accepted 28 December 2018

from solving easier problems in estimating the ability of students. Even though the assumptions made in IRT are not directly applicable to learning, it has been successfully used in estimating the ability of students learning in computer-based learning environments (Lee *et al.*, 2008; Bergner *et al.*, 2015; Colvin *et al.*, 2014; Champaign *et al.*, 2014; Milligan and Griffin, 2016).

Another approach that can be used in estimating the ability of students solving a series of problems is the TrueSkill rating system (Herbrich *et al.*, 2006). TrueSkill is developed by Microsoft Research to rank game players in the Xbox Live system. After assigning an initial rating to each player, TrueSkill updates the rating of players in such a way that the amount of change in the player's rating is in proportion to the unexpectedness of match outcomes. When a stronger player wins against a weaker player, for example, the update in their rating is small because the outcome is not unexpected. However, if a weaker player beats a stronger player, the update in their rating becomes large. TrueSkill can be used to rank multi-player teams, and it takes into account the uncertainty of estimated ratings (Herbrich *et al.*, 2006). By interpreting problem solving as a match between student and problem, it is possible to use the TrueSkill rating system in estimating the ability of students solving a series of problems in the computer-based learning environment. One important characteristic of the TrueSkill rating system, as explained in more detail in the method section below, it updates the rating of players iteratively and does not involve calibration of model parameters. Therefore, it can be used in the computer-based learning environment when it does not have accumulated data that can calibrate model parameters. Although the TrueSkill rating system has these advantages, it has not been used much in education research.

This study investigates whether the TrueSkill rating system can estimate the ability of students who are solving a series of problems in the MOOC environment. In particular, this study compares the ability of students and the difficulty of problems estimated by the TrueSkill rating system to those computed from the two-parameter IRT model. The rest of the paper is organized as follows. The literature review section summarizes previous studies directly related to the present study. The method section describes how the two-parameter IRT model and the TrueSkill rating system are developed and tested with simulated and real problem solving data sets. The result section examines whether the two-parameter IRT model and the TrueSkill rating system can recover the ability of students and the difficulty of problems in the simulated data set, and whether the estimated ability of students can predict their problem solving performance in the MOOC, followed by a discussion.

Literature review

Assessing student learning using item response theory

IRT is a psychometric framework designed to model interactions between the ability of subjects who are taking a test and the difficulty of problems in the test. As it was first developed in the 1950s and 1960s (Lord, 1951; Birnbaum, 1969; Rasch, 1960), IRT has been used extensively in large-scale educational assessments, psychological testing, cognitive diagnosis, health measurement and marketing research (Linden, 2018). Therefore, instead of providing a comprehensive literature review on IRT and its applications, which can be found in Linden (2018), this paper elects to summarize previous studies relevant to the present study; namely IRT was used to assess the ability of students solving problems in the computer-based learning environment. Lee *et al.* (2008) investigated the usefulness of instructional supports by examining how much the ability of students estimated by a two-parameter IRT model changed after students had used different types of instructional supports available in the Web-based physics tutoring environment. In Colvin *et al.*'s (2014) study, student learning was quantified as the difference between the IRT abilities estimated in the pre- and post-tests that were administered to the students enrolled in on-ground and MOOC courses. Champaign *et al.* (2014) examined the correlations between time spent on learning materials, such as eText, formative assessment problems, tutorial videos, discussion board and wiki, and the ability of students measured by Graded Response Model. Bergner *et al.* (2015) compared the ability of students that was estimated by dichotomous and polytomous IRT models to the fraction of correct answers for problems in the MOOC under the two different approaches to handling missing data (incorrect vs. missing at random). Milligan and Griffin (2016) used a partial credit IRT model to assess whether students are capable of doing higher order learning in MOOCs.

Bayesian knowledge tracing

Bayesian Knowledge Tracing or BKT (Corbett and Anderson, 1995) is one of the most popular approaches to quantifying ability of students in educational data mining (EDM) and intelligent tutoring system (ITS). In BKT, the ability of students is modeled as a latent variable that can take two states, mastered or not mastered, and learning is characterized as a transition between these states. Because BKT was originally developed for cognitive theory of learning, each BKT model seeks to estimate the probability of mastery of a fine-grained knowledge or skill (e.g. addition, subtraction, multiplication and division), which is often called Knowledge Component (KC), given the history of responses to problems requiring an understanding of the target KC (Corbett and Anderson, 1995).

Since it was first applied to model students' evolving knowledge required to solve problems in an ITS for LISP programing (Corbett and Anderson, 1995), BKT has been used in many studies in EDM. Qiu *et al.* (2011) used BKT to measure the impact of time between problem solving attempts. Sao Pedro *et al.* (2013) applied BKT to model the science inquiry skill as students were conducting virtual experiments in an interactive simulation environment. San Pedro *et al.* (2013) developed a logistic regression model predicting college enrollment from BKT-based skills and affect features inferred from the log files of a computer-based tutoring system. Wang and Heffernan (2013) extended the standard BKT model to take into account partial credits so that students who have to make more attempts or ask for more hints while solving problems get a score closer to zero. Khajah *et al.* (2014) proposed a new user modeling approach, which is a synthesis of a latent-factor model and BKT, that can predict individual differences in student learning. David *et al.* (2016) used a partial credit BKT model to estimate the ability of students and

to select problems whose difficulty falls within the range of the estimated ability of students. Pardos *et al.* (2013) investigated whether BKT can be adapted to a MOOC. Their study identified three challenges to modeling MOOC data using BKT: lack of an explicit KC model, allowance for unpenalized multiple problem attempts, and multiple pathways through the system that allow for learning outside of the current assessment.

The TrueSkill rating system

Even though the TrueSkill rating system was originally developed to rank game players, it has been used to assess the ability of students in non-gaming environments as well. For example, Lee (2012) used the TrueSkill rating system to efficiently decompose a multi-class classification problem into multiple binary classification problems. His experiment showed that the TrueSkill-based approach was able to reduce the complexity of the multi-class classification problem without losing classification performance. Liu *et al.* (2013) found that the TrueSkill rating-based approach outperformed a PageRank-based approach in estimating the difficulty of questions posted in the community question answering services. Baumann (2017) used the TrueSkill rating system to efficiently create a likability ranking for a large number of speakers from crowdsourced pairwise listener ratings, and examined the TrueSkill-based speaker ranking stability under various conditions. In Kawatsu *et al.* (2018)'s work, the TrueSkill rating system was used to predict students' decision in a training simulation developed for teaching social skills. They were able to predict student choices to a high degree of accuracy when students were making decisions at branches in the social skill simulation.

Method

Two-parameter item response theory model

In the two-parameter IRT model, the probability for a student s who has an ability θ_s to get a problem i correct, which is denoted as $P_{IRT}^{s,i}$, is assumed to be:

$$P_{IRT}^{s,i} = \frac{1}{1 + e^{-\alpha_i(\theta_s - d_i)}} \tag{1}$$

In equation (1), each problem is parameterized by a discrimination coefficient, α_i, and a difficulty parameter, d_i The difficulty parameter on the ability axis is the point for which the predicted probability of correct answer is equal to 0.5. The larger the discrimination coefficient is, the better the problem discriminates students having a low ability from students having a high ability (Ayala, 2009). Theoretically, the difficulty and the discrimination parameters of an IRT model remain unchanged regardless of the ability of students, and the ability of students also remains invariant to problems used for estimating the ability of students (Ayala, 2009). In IRT, students with the same fraction of correct answers can have a different ability, depending on the difficulty of problems they were able to solve correctly.

TrueSkill rating system

The TrueSkill rating system assumes that player's skill follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ where μ is the current estimation of skill and σ is the uncertainty of the estimated skill. In response to the outcome of games, the TrueSkill rating system iteratively updates the skill of players involved in the match by applying Bayes' theorem. Given the current TrueSkill rating scores of two players (prior) and the outcome of the match (data), the TrueSkill rating system updates the skills of players (posterior). In case of two-player games with no-draw, TrueSkill's Bayesian update procedure can be summarized as follows:

$$\begin{aligned}
\hat{\mu}_{winner} &\leftarrow \hat{\mu}_{winner} + \frac{\hat{\sigma}^2_{winner}}{c} \cdot v\left(\frac{\hat{\mu}_{winner} - \hat{\mu}_{loser}}{c}\right) \\
\hat{\sigma}^2_{winner} &\leftarrow \hat{\sigma}^2_{winner}\sqrt{1 - \frac{\hat{\sigma}^2_{winner}}{c^2} \cdot w\left(\frac{\hat{\mu}_{winner} - \hat{\mu}_{loser}}{c}\right)} \\
\hat{\mu}_{loser} &\leftarrow \hat{\mu}_{loser} - \frac{\hat{\sigma}^2_{loser}}{c} \cdot v\left(\frac{\hat{\mu}_{winner} - \hat{\mu}_{loser}}{c}\right) \\
\hat{\sigma}^2_{loser} &\leftarrow \hat{\sigma}^2_{loser}\sqrt{1 - \frac{\hat{\sigma}^2_{loser}}{c^2} \cdot w\left(\frac{\hat{\mu}_{winner} - \hat{\mu}_{loser}}{c}\right)} \\
c^2 &= 2\beta^2 + \hat{\sigma}^2_{winner} + \hat{\sigma}^2_{loser} \\
v(x) &= \frac{\mathcal{N}(x)}{\Phi(x)} \\
w(x) &= v(x) \cdot [v(x) + x]
\end{aligned} \tag{2}$$

where μ is the mean of the estimated TrueSkill rating score, σ is the standard deviation of the estimated TrueSkill rating score, $\mathcal{N}(x)$ is the probability density of a standard normal distribution, $\Phi(x)$ is the cumulative density of a standard normal distribution, and β is the distance that guarantees 76 per cent chance of winning. See Dangauthier *et al.* (2007) for the technical details of the Bayesian update procedure.

The TrueSkill rating system is conceptually similar to Bayesian IRT in the respect that the model parameters are updated in light of newly available data. While the TrueSkill rating system uses the updating rules [equation (2)] to refine TrueSkill rating scores of players, Bayesian IRT employs a simulation-based estimation approach such as Markov Chain Monte Carlo method in estimating the posterior distribution of model parameters (Fox, 2010). In the present study, the initial TrueSkill rating score, μ_0, was set to 25, the initial standard deviation of TrueSkill rating score, σ_0, to 8.33, and the distance that guarantees 76 per cent of winning, β, to 4.17 as recommended in the original TrueSkill paper (Herbrich *et al.*, 2006).

Once the current TrueSkill rating scores of students and problems are estimated, the probability for this particular student to correctly solve the problem at hand can be computed as follows. Suppose that we want to estimate how likely a student (mean and standard deviation of TrueSkill rating score are μ_1 and σ_1) correctly solve a problem whose mean and

standard deviation of TrueSkill rating score are μ_2 and σ_2. Since TrueSkill rating scores are Gaussian distributions, the difference of these two distributions is also a Gaussian distribution with mean of $\mu_1 - \mu_2$ and standard deviation of $\sqrt{{\sigma_1}^2 + {\sigma_2}^2}$. Then, the probability of this student to correctly solve the current problem can be represented as the area under this Gaussian distribution where TrueSkill rating difference is greater than 0. Figure 1(a) illustrates how this approach is applied to a situation where a student whose TrueSkill rating mean and standard deviation are 27 and 0.85 is trying to solve a problem whose TrueSkill rating mean and standard deviation are 26 and 0.80. In this case, the probability for the student to get this problem correct is estimated to be 0.80, which is shown as the highlighted area under the normal difference distribution of two TrueSkill rating scores [Figure 1(b)].

Data sets

To investigate whether the two-parameter IRT model and the TrueSkill rating system can accurately estimate the ability of students and the difficulty of problems, this study used a simulated data set in which 1,000 students solved 200 problems. The ability of students, θ_s, and the difficulty of problems, d_i, were assumed to follow a standard normal distribution, $\mathcal{N}(0, 1)$, and the discrimination parameters, α_i, were assumed to follow a narrower normal distribution, $\mathcal{N}(0, 0.2)$. A Bernoulli random variable with $p = P_{IRT}^{s,i}$ [equation (1)] was used to represent an answer of a student having an ability of θ_s on the problem i whose difficulty is d_i and discrimination is α_i, which produced a 1,000 (students) by 200 (problems) matrix consisting of either 1 (correct answer) or 0 (wrong answer). An open source R package named ltm (Rizopoulos, 2006) was used to develop a two-parameter IRT model estimating the ability of students and the difficulty of problems from the problem solving performance of simulated students.

Unlike IRT, which requires a matrix of entire problem solving performance to estimate the ability of students and the difficulty of problems, the TrueSkill rating system needs a set of student, problem, and correctness triplet so that it can iteratively update the ability of students and the difficulty of problems as students are solving a series of problems [equation (2)]. Thus, the 1,000 (students) by 200 (problem) matrix used in the IRT model was decomposed into 200,000 triplets of student, problem and correctness. These triplets were then fed into an open source TrueSkill python library (Lee, 2017) to estimate TrueSkill rating scores representing the ability of students and the difficulty of problems. The result section compares in detail the ability of students and the difficulty of problems computed from these two algorithms to the true values used in the simulated data set.

After confirming that IRT and TrueSkill algorithms were able to accurately recover the ability of students and the difficulty of problems in the simulated problem solving data set, these two algorithms were tested against a real problem solving data that captures how 452 students enrolled in a MOOC named "edX 8. MReVx Mechanics Review (hereafter, 8.MReVx)" solved 43 homework and quiz problems in the summer 2014 semester. For further exploration of the course structure of 8.MReVx, which was offered by MIT, see the archived course at https://courses.edx.org/courses/MITx/8.MReVx/2T2014/info.

Results

Estimating ability of students and difficulty of problems from simulated data set

As shown in Table I, the two-parameter IRT model was able to accurately recover true parameter values used in the simulated data set. Similarly, TrueSkill rating scores of students and problems were found to have near perfect correlations with the true parameter values that generated the data set. These results suggest that the two-parameter IRT and TrueSkill models can be

Figure 1 Estimating a problem solving probability from two TrueSkill rating score distributions

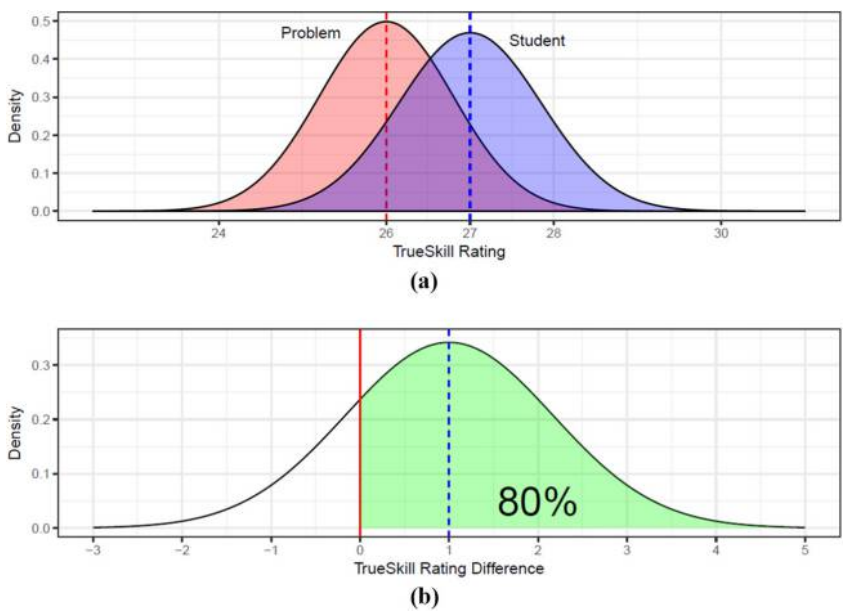


Table I Correlations among true parameters, IRT parameters and TrueSkill rating scores

	True ability	IRT ability	TrueSkill rating of students
True ability	1.00	0.99	0.98
IRT ability	0.99	1.00	0.99
TrueSkill rating of students	0.98	0.99	1.00
	True difficulty	**IRT difficulty**	**TrueSkill rating of problems**
True difficulty	1.00	0.99	0.97
IRT difficulty	0.99	1.00	0.97
TrueSkill rating of problems	0.97	0.97	1.00

used in estimating the ability of students and the difficulty of problems from the binary outcome (correct vs. wrong) of problem solving data.

One useful feature of the TrueSkill rating system is that it does not need the entire problem solving data to estimate current ability of students and difficulty of problems; ability of students and difficulty of problems are updated iteratively as students are solving more problems [see equation (2)]. Figure 2 shows how TrueSkill rating scores are converging to the final values as simulated students with different abilities (at the 25th, 50th and 75th percentile) are solving 200 problems. The error bars in the figure indicate the uncertainty of the estimated TrueSkill rating scores of the simulated students. The student at the 50th percentile happened to get the first six problems correct, resulting initially in a higher TrueSkill rating score than the final value. However, after trying to solve 36 problems, the final TrueSkill rating value, 25.43, falls within the error bars of the currently estimated TrueSkill rating score. In case of the student at the 75th percentile, just after the 12th problem, the error bars of the estimated TrueSkill rating score include the final TrueSkill rating value obtained at the 200th problems, 27.86. For the student at the 25th percentile, the TrueSkill algorithm needed 59 problems before it converges to the final TrueSkill rating score, 22.75.

Estimating ability of students and difficulty of problems in the MOOC

Unlike a test where everyone solves the same set of problems at the same time, students enrolled in MOOCs choose their own learning activities at their own pace. As a result, the data capturing how students solve problems in MOOCs is growing over an extended period of time. For instance, the first 1 per cent of 8.MReVx data set (in the increasing order of when students solved homework and quiz problems) contains problem solving performance of only six students. Therefore, it would be important to investigate whether the two-parameter IRT and the TrueSkill rating models can correctly estimate the ability of students and the difficulty of problems from a problem solving data set *growing over time*. To answer this question, this study compared the ability of students and the difficulty of problems estimated from 1, 2.5, 5, 7.5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 per cent of the problem solving performance data (in the increasing order of when students solved homework and quiz problems) obtained from 8.MReVx data set.

Figure 3 shows how the correlations between IRT parameters and TrueSkill rating score estimated from the various fractions of problem solving data set. The size of circles in Figure 3 reflects the number of students contained in the analyzed data sets. Note that when 1, 2.5 and 5 per cent of the data sets were analyzed, the correlation could not be computed because the two-parameter IRT model had failed to estimate the ability of students and the difficulty of problems. The correlation between IRT ability and TrueSkill rating score of students drops from 0.98 to 0.95, and remains essentially the same when more than 10 per cent of data set, which includes 50 students and 43 problems, was analyzed. The correlations

Figure 2 Estimated TrueSkill rating score vs. number of problems students solved in the simulated data set

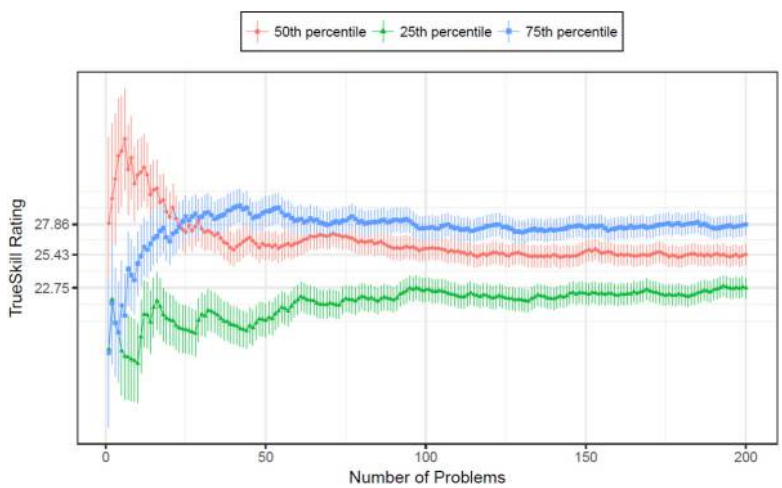
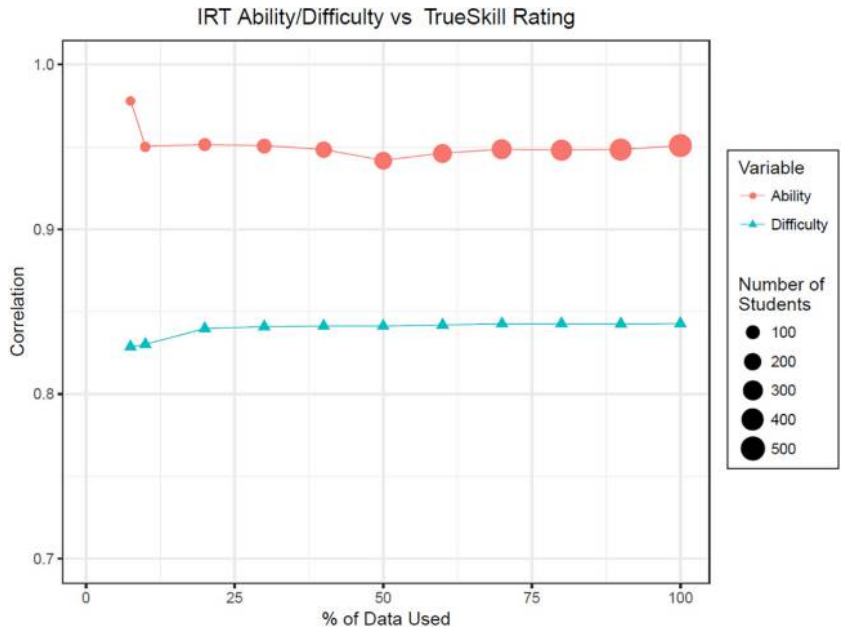


Figure 3 Correlations between IRT parameters and TrueSkill rating scores on the partial data sets based on the problem submission time



between the difficulty of problems from the two-parameter IRT model and the TrueSkill rating score are not as strong. It converges to 0.84 when more than 20 per cent of the problem solving data set, which includes the problem solving performance of 102 students on 43 problems, was analyzed.

Predicting problem solving performance of students in the massive open online course

To examine the usefulness of the TrueSkill rating system, this study developed a TrueSkill-based data mining model predicting whether students enrolled in 8.MReVx would be able to solve homework and quiz problems correctly, and compared it to the two-parameter IRT model in terms of their predictive power. To estimate the predictive power of the data mining models without bias, this study employed a training/test set approach. First, 1, 2.5, 5, 7.5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 per cent of the problem solving performance data sets (in the increasing order of when students solved homework and quiz problems) were divided into corresponding training and test sets. When creating a test set, which consists of 20 per cent of each data set, stratified random sampling was used to ensure that the ratio of correct to incorrect answers in both training and test sets are similar. Each training set was fitted to two-parameter IRT and TrueSkill models to estimate the difficulty and discrimination parameters of problems and the ability of students. These parameters were then used when the probability of submitting a correct answer was estimated in the test sets.

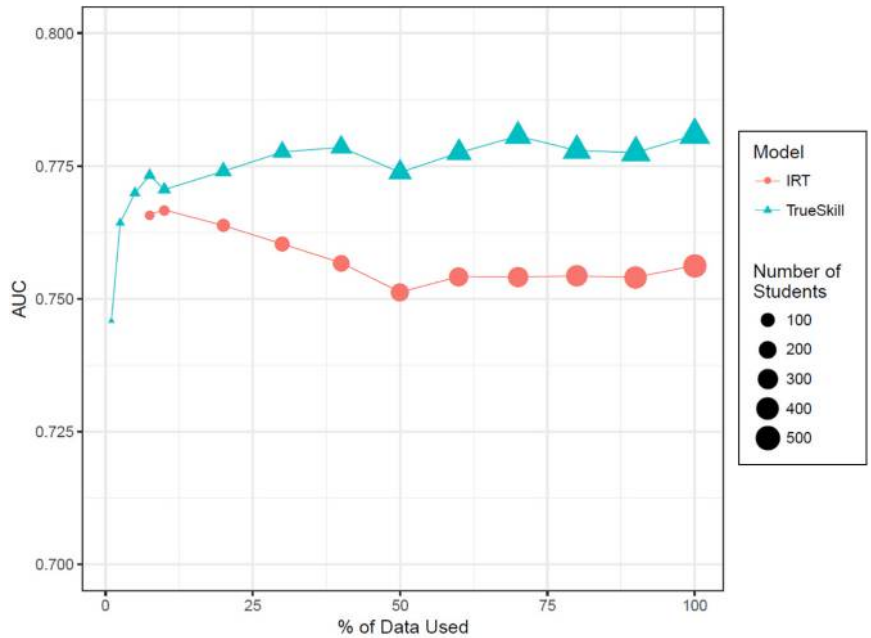
Figure 4 shows how Area Under the receiver operating characteristic Curve (AUC) values computed from the two-parameter IRT and TrueSkill models were changing when correct problem solving performance was estimated from a different amount of 8.MReVx data set. AUC is frequently used as a means for quantifying the quality of a binary classification model (being able to vs. fail to solve a problem in the context of the present study) because it does not depend on a single cutoff probability (Fawcett, 2006). AUC can vary from 0.5 (predicting no better than random guessing) to 1.0 (making perfect predictions), and it is known to be equal to the probability that a predictive model will rank a randomly chosen positive instance higher than a randomly chosen a negative instance. Roughly speaking, AUC computed on the test set can be interpreted as a probability for a binary classification algorithm to make correct predictions on the new, unseen data in the future (Fawcett, 2006).

Overall, both models showed a comparable predictive power. When the full data set was used, the AUC value of the TrueSkill rating system ($AUC_{TrueSkill,\ 100\ per\ cent}$ = 0.781) was larger by 3.3 per cent than that of the IRT model ($AUC_{IRT,\ 100\ per\ cent}$ = 0.756). However, the IRT model did not converge when the size of data set was small (1, 2.5 and 5 per cent of the full data set), resulting in no ACU values as shown in Figure 4. Considering the fact that the smallest data set contains the problem solving records of just six students, it is quite remarkable that the TrueSkill rating system was able to achieve the AUC value ($AUC_{TrueSkill,\ 1\ per\ cent}$ = 0.746) that is just 4.7 per cent smaller than the AUC value obtained from the full data set ($AUC_{TrueSkill,\ 100\ per\ cent}$ = 0.781) which contains problem solving records of 453 students. In addition, the AUC values from the IRT model decreased slightly until 50 per cent of data was used, and leveled off approximately at 0.754. On the other hand, the AUC values from the TrueSkill rating system appear to increase slightly as the size of a data set increases.

Discussion

To facilitate student learning, it is critical to provide instructional supports and guidance. However, it does not mean that we need to provide instructional supports immediately when students make mistakes or fail to complete the learning task at hand because it will prevent students from exerting enough cognitive

Figure 4 AUC values computed from the two-parameter IRT and TrueSkill models using different amounts of problem solving data set



Note: The size of symbols reflects the number of students included in the analyzed data sets

efforts (Kapur, 2008; Schmidt and Bjork, 1992). On the other hand, academically weaker students will flounder, get frustrated, and may give up their learning unless they receive instructional supports in time. Therefore, the real issue is how we can determine the most appropriate time to provide instructional supports when students having a different ability are trying to resolve learning tasks with varying difficulties, which requires an accurate estimation of the ability of students in relation to the difficulty of learning tasks. This issue would become more important in the computer-based learning environment where students learn mostly on their own and there is no teacher who can provide appropriate instructional supports and guidance to struggling students. Unfortunately, however, the ways in which instructional supports are provided in the computer-based learning environment do not take into account the ability of students. Regardless of their ability, students get instructional supports after submitting the same number of incorrect answers, which is unlikely to maximize the learning outcome of students. The findings from this study suggest that the two-parameter IRT model and the TrueSkill rating system may be able to address this issue because these approaches were able to estimate the ability of students solving a series of problems in relation to the difficulty of those problems, and to predict how likely students will be able to solve problems correctly.

One important characteristic of a computer-based learning environment is that students are not required to complete the same set of learning activities at the same time. Although this flexibility allows students to have more personalized learning experiences, it makes it more difficult to accurately estimate the ability of students and the difficulty of learning tasks. This study found that the two-parameter IRT model was not able to estimate the ability of students and the difficulty of problems when the analyzed data set included a small number of students, which is consistent with previous studies investigating accuracy and stability of IRT model parameters vs. sample size (Chen *et al.*, 2013; Jiang *et al.*, 2016; Torre and Hong, 2010). On the other hand, the TrueSkill rating system was able to estimate the ability of students and the difficulty of problems relatively accurately when a small number of students solved the problems. Another strength of the TrueSkill rating system is that it requires less computation than IRT. In IRT, new problems can be added to the problem set or homework only after the item parameters are calibrated on a different set of students. On the other hand, since the TrueSkill rating system involves no calibration process, new problems can be added to the problem set or homework any time without any additional computation. The ability estimation made on the newly added problems may be less accurate initially, but as more students are solving these problems, the estimation accuracy will improve quickly. These features would make the TrueSkill rating system more appropriate for inferring the ability of students in the computer-based learning environment in which the contents (e.g. adding new problems to problem sets or homework) are changing more frequently, compared to standardized tests such as SAT or GRE for which IRT is originally developed.

This study found that the correlations between the difficulty parameter of IRT and the TrueSkill rating score of problems was lower than the correlations between the ability parameter of IRT and the TrueSkill rating score of students who solve the problems. This result may be related to the number of IRT parameters describing the ability of students and the difficulty of problems. The two-parameter IRT model uses two parameters, difficulty and discrimination, to characterize a

problem whereas the TrueSkill rating system uses just one parameter, TrueSkill rating score. Since different combinations of difficulty and discrimination parameters of a problem can fit the same problem solving performance data equally well, it is possible that the correlation between the TrueSkill rating scores of problems and the difficulty parameters of the two-parameter IRT model can be low. This explanation seems to support a higher correlation between the TrueSkill rating score of students and the IRT ability parameter because both IRT and TrueSkill rating system uses only one parameter to model the ability of students.

Limitations of study and future works

Although the TrueSkill rating system has proven to be an effective algorithm for matchmaking in an online gaming system, and it can be used in estimating the ability of students and the difficulty of problems in an online learning environment, it can only deal with binary outcomes (win vs. lose; to be able vs. fail to solve a problem). In other words, the TrueSkill rating system cannot handle partial scores, which is not uncommon in many educational settings. Similarly, to use the TrueSkill rating system, complex learning outcomes have to be simplified. For example, students who are able to get a problem correct at their second attempt or after using hints available in the computer-based learning environment may have to be treated as "not able to solve the problem" because the TrueSkill algorithm can only take binary outcomes.

As this study analyzed the log files of an already completed MOOC, it was not possible to examine whether students were able to perform better when they received instructional supports and guidance based on the IRT– or TrueSkill-based prediction models. Therefore, it would be important to develop an adaptive computer-based learning environment capable of providing instructional supports and guidance based on the prediction model, and examine whether it can actually improve the learning outcome of students.

Unlike IRT, which has been studied extensively in psychometrics, very little research has been conducted to examine the usefulness of the TrueSkill rating system in education. Therefore, there are many important questions that have to be answered about the TrueSkill rating system as a means for assessing the ability of students in the computer-based learning environments. For example, does it work well regardless of the difficulty of problems and the ability of students just like IRT? What are the minimum number of students and problems that will yield stable estimates of the ability of students and the difficulty of problems when we have various compositions of students and problems in terms of their ability and difficulty? Similarly, in predicting problem solving performance of students based on the estimated ability of students and the difficulty of problems, prediction would become more accurate when the data set includes more students and problems. However, how many students and problems do we need to make reasonably accurate predictions consistently? Further research is warranted to answer these practical, but important, questions about the usefulness of the TrueSkill rating system.

Finally, we may be able to use the estimated ability of students and the difficulty of problems in building a data mining model predicting the academic success of students in the computer-based learning environment. For example, we may be able to predict the performance of students in the final exam from their ability estimated from the problem solving performance observed in the first a few weeks of the semester. Similarly, we may be able to build an early warning system that can estimate a dropout probability of students in the MOOC or e-learning environment by using the ability of students estimated from the problem solving performance observed early in the semester as a predictor. As a future work, it would be interesting to investigate whether such predictive data mining models can be developed.

References

Ayala, R.J. (2009), *The Theory and Practice of Item Response Theory*, The Guilford Press, New York, NY.

Baumann, T. (2017), "Large-Scale speaker ranking from crowdsourced pairwise listener ratings", *Interspeech 2017, Stockholm*, pp. 2262-2266.

Baylari, A. and Montazer, G.A. (2009), "Design a personalized e-learning system based on item response theory and artificial neural network approach", *Expert Systems with Applications*, Vol. 36 No. 4, pp. 8013-8021.

Bergner, Y., Colvin, K. and Pritchard, D.E. (2015), "Estimation of ability from homework items when there are missing and/or multiple attempts", *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, New York, NY*, pp. 118-125.

Bienkowski, M. Feng, M. and Means, B. (2012), "Enhancing teaching and learning through educational data mining and learning analytics: an issue brief", available at: https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf (accessed 16 November 2018).

Birnbaum, A. (1969), "Statistical theory for logistic mental test models with a prior distribution of ability", *Journal of Mathematical Psychology*, Vol. 6 No. 2, pp. 258-276.

Champaign, J., Colvin, K.F., Liu, A., Fredericks, C., Seaton, D. and Pritchard, D.E. (2014), "Correlating skill and improvement in 2 MOOCs with a student's time on tasks", *Proceedings of the First ACM Conference on Learning @ Scale, ACM Press, New York, NY*, pp. 11-20.

Chen, C.-M., Lee, H.-M. and Chen, Y.-H. (2005), "Personalized e-learning system using item response theory", *Computers & Education*, Vol. 44 No. 3, pp. 237-255.

Chen, W.-H., Lenderking, W., Jin, Y., Wyrwich, K.W., Gelhorn, H. and Revicki, D.A. (2013), "Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data", *Quality of Life Research*, Vol. 23 No. 2, pp. 485-493.

Colvin, K.F., Champaign, J., Liu, A., Zhou, Q., Fredericks, C. and Pritchard, D.E. (2014), "Learning in an introductory physics MOOC: all cohorts learn equally, including an on-campus class", *The International Review of Research in Open and Distributed Learning*, Vol. 15 No. 4, pp. 263-283.

Corbett, A. and Anderson, J. (1995), "Knowledge tracing: modeling the acquisition of procedural knowledge", *User Modelling and User-Adapted Interaction*, Vol. 4 No. 4, pp. 253-278.

Dangauthier, P., Herbrich, R., Minka, T. and Graepel, T. (2007), "TrueSkill through time: revisiting the history of chess", *NIPS '07 Proceedings of the 20th International Conference on Neural Information Processing Systems, Curran Associates, Vancouver, British Columbia*, pp. 337-344.

David, Y.B., Segal, A. and Gal, Y.K. (2016), "Sequencing educational content in classrooms using Bayesian knowledge tracing", *Proceedings of the Sixth International Learning Analytics & Knowledge Conference, Edinburgh*, pp. 354-363.

Fawcett, T. (2006), "An introduction to ROC analysis", *Pattern Recognition Letters*, Vol. 27 No. 8, pp. 861-874.

Fox, J.-P. (2010), *Bayesian Item Response Modeling*, Springer, New York, NY.

Herbrich, R., Minka, T. and Graepel, T. (2006), "TrueSkill^{TM}: a Bayesian skill rating system", *Advances in Neural Information Processing System*, Vol. 19, pp. 569-576.

Jiang, S. Wang, C. and Weiss, D.J. (2016), "Sample size requirements for estimation of item parameters in the multidimensional graded response model", *Frontiers in Psychology*, Vol. 7, available at: www.ncbi.nlm.nih.gov/pubmed/26903916 (accessed 16 November 2018).

Kapur, M. (2008), "Productive failure", *Cognition and Instruction*, Vol. 26 No. 3, pp. 379-424.

Kawatsu, C., Hubal, R. and Marinier, R.P. (2018), "Predicting students' decisions in a training simulation: a novel application of TrueSkill", *IEEE Transactions on Games*, Vol. 10 No. 1, pp. 97-100.

Khajah, M., Wing, R., Lindey, R. and Mozer, M.C. (2014), "Integrating latent-factor and knowledge-tracing models to predict individual differences in learning", *Proceedings of International Conference on Educational Data Mining (EDM) 2014, London*, pp. 99-106.

Lee, H. (2017), "TrueSkill", available at: https://trueskill.org (accessed 28 December 2018).

Lee, J.-S. (2012), "TrueSkill-based pairwise coupling for multi-class classification", *ICANN'12 Proceedings of the 22nd international conference on Artificial Neural Networks and Machine Learning, Springer, Berlin, Heidelberg*, pp. 213-220.

Lee, Y., Palazzo, D.J., Warnakulasooriya, R. and Pritchard, D.E. (2008), "Measuring student learning with item response theory", *Physical Review Special Topics – Physics Education Research*, Vol. 4, p. 010102.

Linden, W. J. V D L. (2018), *Handbook of Item Response Theory, Three Volume Set*, CRC Press, Boca Raton.

Liu, J., Wang, Q., Lin, C.-Y. and Hon, H.-W. (2013), "Question difficulty estimation in community answering services", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, DC*, pp. 85-90.

Lord, F.M. (1951), "A theory of test scores and their relation to the trait measured", *ETS Research Report Series*, Vol. 1951 No. 1, pp. 1-79.

Milligan, S. and Griffin, P. (2016), "Understanding learning and learning design in MOOCs: a measurement-based interpretation", *Journal of Learning Analytics*, Vol. 3 No. 2, pp. 88-115.

Pardos, Z., Bergner, Y. and Seaton, D.T. (2013), "Adapting Bayesian knowledge tracing to a massive open online course in edX", *Proceedings of International Conference on Educational Data Mining (EDM) 2013, Memphis*, pp. 137-144.

Qiu, Y., Qi, Y., Lu, H., Pardos, Z.A. and Heffernan, N.T. (2011), ""Does time matter? Modeling the effect of time with Bayesian knowledge tracing", *Proceedings of International Conference on Educational Data Mining (EDM) 2011, Eindhoven*, pp. 139-148.

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, University of Chicago Press, Chicago.

Rizopoulos, D. (2006), "LTM: an R package for latent variable modeling and item response analysis", *Journal of Statistical Software*, Vol. 17 No. 5, available at: www.jstatsoft.org/article/view/v017i05 (accessed 16 November 2018).

San Pedro, M.O., Baker, R., Bowers, A. and Heffernan, N.T. (2013), "Predicting college enrollment from student interaction with an intelligent tutoring system in middle school", *Proceedings of International Conference on Educational Data Mining (EDM) 2013, Memphis*, pp. 177-184.

Sao Pedro, M.S., Baker, R. and Gobert, J.D. (2013), "Incorporating scaffolding and tutor context into Bayesian knowledge tracing to predict inquiry skill acquisition", *Proceedings of International Conference on Educational Data Mining (EDM) 2013, Memphis*, pp. 185-192.

Schmidt, R.A. and Bjork, R.A. (1992), "New conceptualizations of practice: common principles in three paradigms suggest new concepts for training", *Psychological Science*, Vol. 3 No. 4, pp. 207-217.

Tanenbaum, C. Floch, L.K. and Boyle, A. (2013), "Are personalized learning environments the next wave of K-12 education reform?", available at: www.air.org/sites/default/files/AIR_Personalized_Learning_Issue_Paper_2013.pdf (accessed 16 November 2018).

Torre, J. and Hong, Y. (2010), "Parameter estimation with small sample size a higher-order IRT model approach", *Applied Psychological Measurement*, Vol. 34 No. 4, pp. 267-285.

Wang, Y. and Heffernan, N. (2013), "Extending knowledge tracing to allow partial credit: using continuous versus binary nodes", *Proceedings of International Conference on Artificial Intelligence in Education, Memphis*, pp. 181-188.

Corresponding author

Youngjin Lee can be contacted at: yjlee@ku.edu