Measuring Predictive Performance of User Models: The Details Matter

Radek Pelánek
Masaryk University
Brno, Czech Republic
pelanek@fi.muni.cz

ABSTRACT

Evaluation of user modeling techniques is often based on the predictive accuracy of models. The quantification of predictive accuracy is done using performance metrics. We show that the choice of a performance metric is important and that even details of metric computation matter. We analyze in detail two commonly used metrics (AUC, RMSE) in the context of student modeling. We discuss different approaches to their computation (global, averaging across skill, averaging across students) and show that these methods have different properties. An analysis of recent research papers shows that the reported descriptions of metric computation are often insufficient. To make research conclusions valid and reproducible, researchers need to pay more attention to the choice of performance metrics and they need to describe more explicitly details of their computation.

KEYWORDS

evaluation; metrics; predictive accuracy; student modeling; RMSE; AUC

1 INTRODUCTION

A key approach in the evaluation of user modeling techniques is the analysis of their predictive accuracy, i.e., their ability to predict future actions of users. To measure predictive performance of models, we need to summarize the difference between predictions and observations by some performance metric, e.g., RMSE, AUC, MAE, log-likelihood, or accuracy. Although the choice of a specific metric used for analysis often does not get much attention in published research, this choice can significantly influence interpretation of results of model evaluation [32].

Moreover, it is not just the choice of a metric that matters. Even seemingly small details of metric computation can be important. This is well illustrated by a recent work by Khajan et al. [22]. They studied deep knowledge tracing approach proposed by Piech et al. [33], who claimed that their approach based on deep learning leads to large improvement compared to previous results reported for the same data set in literature [30]. In their analysis, Khajan et al. [22] noticed that although both [33] and [30] used the same performance metric (AUC), the metric was computed in each case in a slightly different way (global computation of the metric versus per-skill computation with averaging). The large improvement reported in [33] was probably to large degree caused by this methodological issue.

Motivated by this case, we analyze in detail two commonly used performance metrics (RMSE and AUC), specifically with the focus on the used averaging approach. To highlight the issues involved, we provide illustration on artificial examples. We also provide discussion of relevant research literature – providing pointers to relevant discussions in other research areas and mapping the use of metrics in user modeling. This overview shows that many studies rely on the AUC metric, which has several known disadvantages, and that research studies typically do not provide explicit details about metrics computation. This is an obstacle to reproducibility and research progress.

To keep the paper compact, we consider examples and studies only from the area of student modeling (models applied for personalization in education). Nevertheless, the raised issues are relevant also for other types of user models and personalization areas.

2 DEFINITION OF METRICS

At first, note that the word "metric" is traditionally used in the context of student modeling in a sense "any function that is used to make comparisons", not in the mathematical sense of a distance function. We focus on two metrics – RMSE and AUC – that are used most commonly in student modeling and represent two significantly different approaches to measuring predictive performance, for overview of other metrics and their usage see [32].

2.1 Basic Definitions of Metrics

We start with a basic definition of metrics, where we treat all predictions and observations uniformly. We assume that we have data about n cases, numbered $i \in \{1, \ldots, n\}$, a student model provides predictions $p_i \in [0, 1]$, and the observed value is given by the binary value $o_i \in \{0, 1\}$. Root mean square error (RMSE) is then given as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (o_i - p_i)^2}$$

RMSE is an "error metric", i.e., lower values mean better predictive performance.

The second metric that we consider is the area under the receiver operating curve (AUC). The receiver operating curve (ROC) summarizes performance of a binary classification over all possible thresholds. The curve has "false positive rate" on the x-axis and "true positive rate" on the y-axis, each point of the curve corresponds to a choice of a threshold; for a detailed introduction to

UMAP'17 Adjunct, July 09-12, 2017, Bratislava, Slovakia

978-1-4503-5067-9/17/07…$15.00
DOI: 10.1145/3099023.3099042

ROC curve construction and interpretation see [9]. Area under the ROC curve (AUC) provides a summary performance measure across all possible thresholds. It is equal to the probability that a randomly selected positive observation has higher predicted score than a randomly selected negative observation. AUC is 1 for a perfect model and 0.5 for a random predictions, i.e., it is interpreted as a reward (higher is better). The area under the curve can be approximated using a metric called A'; this metric is equivalent to the well-studied Wilcoxon statistics [10].

The fundamental difference between RMSE and AUC is that RMSE considers absolute values of predictions, whereas AUC takes into account only their relative ordering.

The practical computation of RMSE is straightforward, whereas the computation of the AUC metric is not completely clear and can be done using different approaches. For example, scikit, a popular machine learning library for Python, uses a trapezoid rule, whereas code by Ryan Baker [1] (used in previous student modeling evaluations) uses brute force to examine all tuples of correct/incorrect answers. Fawcett [8] discusses different possibilities of AUC computation in detail. Implementations may differ in their results particularly in cases with small data and repeated values of predictions p_i.

2.2 Averaging

The basic definitions of metrics considers only "one dimensional data". But in student modeling we typically have at least two basic dimensions of data: students and skills[1]. Data (both observations and predictions) can thus be seen as a matrix (typically with missing values). The basic definitions of metrics are based on computations over a flattened matrix. Alternatively, we can compute metrics per row (or column) of the matrix and then compute an average value of the metric.

Thus there are three main approaches to computing any metric:

- *Global computation.* In the metric computation we do not differentiate between students and skills and treat all data points as equal.
- *Averaging across skill.* We compute the metric for each skill and then take an average (in the case of low number of skills we may also report the value for each skill).
- *Averaging across student.* We compute the metric for each student and then take an average.

None of these approaches is "the correct one", since the suitability of each approach depends on a particular application. At the same time, the choice of the approach is important – we will show that these approaches can lead to quite different results.

To get a basic intuition why the results may differ, consider a case of highly uneven distribution of answers, i.e., some students (skills) have much larger number of answers than others – such situation are in fact very typical in real educational systems. With the global computation of a metric all data points have the same weight and thus the results are influenced mainly by students (skills) with many answers. On the other hand, the per student (or per skill) computation gives equal weight to all students (skills) without regard to the number of answers, i.e., answers for students (skill) with many answers have less weight.

[1] In other user modeling application the dimensions would be "users" and "items".

3 LITERATURE REVIEW

As the previous section shows, we can choose from metrics with quite different properties and we can compute them in several different ways. Now we provide an overview of research literature to show what methodical advice is available and what is the current practice in research papers.

3.1 Metrics in Student Modeling

A detailed overview of metrics used for evaluation of predictive accuracy of student models is provided in [32]. Several other works describe general methodological issues connected with performance metrics. Dhanani et al. [7] compare metrics in the case of learning model parameters; they conclude that RMSE it better than AUC for this purpose. Pardos and Yudelson [31] study the ability of models to identify "moment of learning" and analyze the relation between this ability and predictive accuracy metrics; the AUC metric again shows poor results. González-Brenes and Huang [16] briefly mention the differences between global computation and computation per skill and possible relation to the Simpson's paradox.

Both RMSE and AUC are widely used for evaluation of student models. In most cases, however, the exact approach to computation is typically not explicitly specified in papers. In most cases, probably, the used approach is either global (particularly for the RMSE metric) or averaging per skill.

The RMSE metric has been used for example in [5, 12, 26, 27, 36–38]. RMSE was also used as a metric in the KDD Cup 2010, which focused on student performance evaluation. Examples of papers that use both the AUC metric and some other metric are [5, 12, 13, 20, 21]. There are also many papers that use only the AUC metric for evaluation, for example [3, 4, 15, 17, 28, 33].

Some papers that use the AUC metric explicitly describe per skill computation or averaging. Pardos and Heffernan [30] report AUC for individual skill. González-Brenes et al. [14] discuss both global computation and averaging over skills. Khajah et al. [22] use both global computation and averaging over skills and discuss the impact of the choice on comparison with previous work. Several works compute AUC per student and report averages and results of statistical comparisons [2, 29, 34].

3.2 Metrics in Other Areas

Performance metrics are used also in many other research areas. Result and observations from these areas may provide useful insight for evaluation of user modeling techniques.

The RMSE metric is closely connected to sum of square errors and mean square of errors. From the perspective of model comparison all these metrics are equivalent since averaging and square root are monotone operations. The exact equivalence however does hold only for the global computation. In the case of per skill or per student averaging the result may slightly differ. In some domains (particularly in weather forecasting) the mean square error (RMSE without the square root) is called a Brier score [6, 35] or a quadratic scoring rule [11]. The Brier score is sometimes decomposed into additive components [25], which provide further insight into behaviour of predictive models.

The ROC curve and AUC metric are successfully used in many different research areas, but their use is criticized for several reasons [19, 23], e.g., because the metric summarizes performance over all possible thresholds, even over those for which the classifier would never be practically used. Marzban [24] discusses AUC in the meteorology context and shows that "AUC discriminates well between good and bad models, but not between good models".

Fawcett [8] provides a detailed discussion of the ROC curve and the AUC metric, discussing also averaging issues (with a focus on the construction of the ROC curve). Hamill and Juras [18], using the context of meteorology, discuss the issue of metric interpretation in the case when the frequency of observed events is not invariant in all samples (which is closely relevant to varying success rates for different skills in student models).

4 ILLUSTRATION OF METRIC PROPERTIES

We use artificial simplified examples to highlight issues with metric computation. To do so we use a very simple model of learning – simple error curve model, where the probability of an error by a student decreases exponentially with the number of attempts. This model can be easily used for both generating data and as a predictive model. All reported experiments use 10,000 simulated students.

4.1 Absolute and Relative Values of Metrics

Before we consider issues connected to averaging, we provide a clarification related to interpretation of absolute and relative values of metrics (by relative value we mean the difference in values of two models). Sometimes these values are used to make judgments about the quality of a model or about the significance of a model improvement. Such use of metrics is, however, rather misleading. From the perspective of model evaluation it makes sense to consider only the ordering of metric values; the magnitude and differences of metrics values are dependent mainly on the data available, not on the quality of models.

For the RMSE metric, the value of the metric is closely related to the average error rate. When average error rate is near 50%, the RMSE value will be near 0.5, unless we have very good predictor, which is in the case of predicting student noisy behaviour unlikely. If the average error rate is low, the RMSE value will go towards zero even for a simple constant predictor. For the AUC metric, the value will be high (near 1) even for a simple model when there is high heterogeneity in data (e.g., differences among skills, students, or pronounced learning leading to large difference between the beginning and the end of each student's sequence). With homogeneous data the AUC value will be typically low (near 0.5) even for a complex model.

As a specific example, consider the error curves in Figure 1. We consider only cases where we fit the data by the same model that generated them. If we consider only the curve A, the metric values are poor: RMSE 0.492, AUC 0.593. If we consider only the curve B, RMSE is much better: 0.160 (because the error rate is low). If we consider model with both curves A and B, AUC is much better: 0.834 (because of the heterogeneity in data). Note that in all cases we are evaluating optimal predictions, i.e., the differences in metric

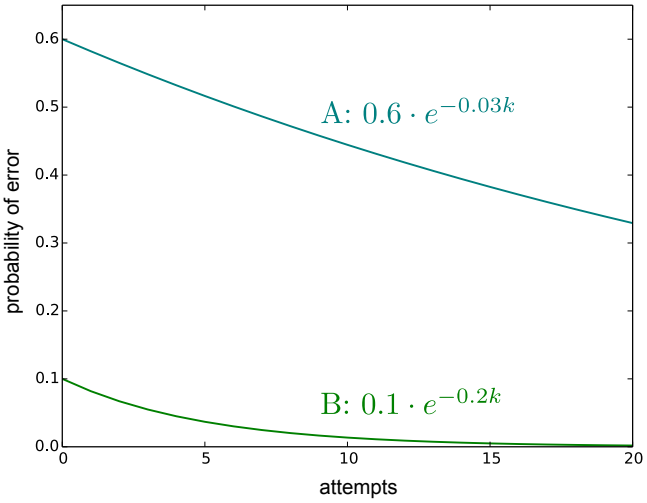


Figure 1: Error curves used for illustration of metric values.

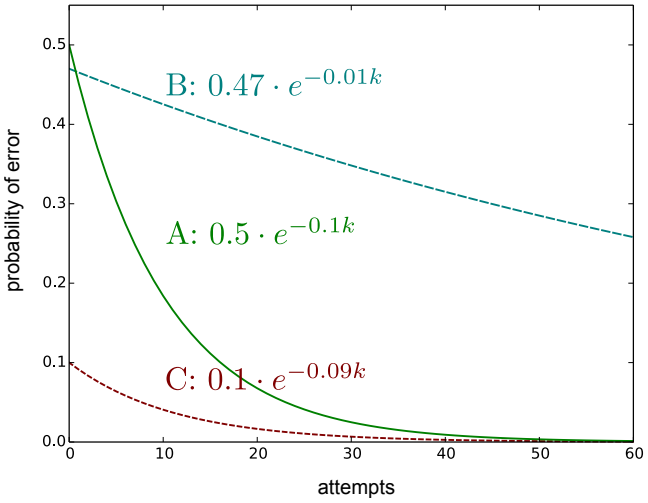
values are not caused by inherent changes in predictive ability of models, but by the characteristics of data.

We typically use metrics for model comparison and the focus is thus not on their absolute values, but on relative performance of models (difference between metric values for different models). These relative values also cannot be easily interpreted. Specifically with the AUC metric we can have vastly different models and obtain the same or nearly the same values of the metric. For example, if we divide all predictions by two, the value of AUC metric remains the same, since the metric considers only relative ordering of predictions. In the error curve model with single skill, even an arbitrary model for which error predictions decrease with k achieves the same AUC value as the optimal model. A solid difference in AUC typically means a model improvement, but a lack of difference in AUC clearly does not mean "absence of improvement" (see also discussion in [24]). This means that by relying solely on the AUC metric, researchers can miss important results!

4.2 Averaging Across Students

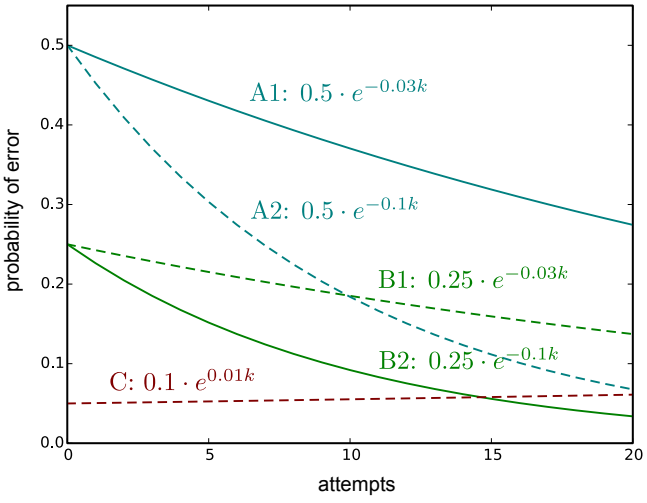
Now we turn to discussing issues related to different averaging methods, starting with averaging across students. The differences in global computation of metrics and averaging across students are important in cases where the number of available data from individual students is unevenly distributed. This corresponds to a rather typical case in any user data – typically we have many users with few responses and few users with many responses. The global computation of metrics gives the same weight to all responses, whereas averaging across users gives the same weight to all users (and thus lower weight to responses by users with many responses).

For specific illustration let us consider the case illustrated in Figure 2. The data are generated according to the curve A with an uneven distribution of number of answers among students: 70% of students have only 5 attempts, 30% of students have 60 attempts. Data are fitted by two models, the first one (curve B) fits only the beginning of a sequence, the second one (curve C) only the end of the sequence. If we compare the models with respect to the RMSE



model	RMSE global	RMSE per student
B	0.40	0.46
C	0.35	0.48

Figure 2: Illustration of the per student computation. Data are generated according to curve A, with 70% of students having 5 attempts and 30% of students having 60 attempts. The table compares predictions by curves B and C.



model	AUC global	AUC per skill
A1, B2 (correct)	0.73	0.63
A2, B1 (speed mismatch)	0.60	0.63
A1, C (negative learning)	0.68	0.45

Figure 3: Illustration of the impact of AUC computation. Data generated by a model with skills A1, B2 and fitted by three models.

metric computed globally, model C is better. If we compare the models with respect to the RMSE metric averaged across students, model B is better.

The use of the AUC metric with averaging across students brings one additional problem. The AUC metric is not well defined when all responses are the same (e.g., all answers are correct). When we consider the computation of the metric per student, such cases are likely to happen (particularly for students with small number of answers). It is not clear how to treat these cases. The basic approach is to ignore these undefined cases (as done for example in [34]), which is however, not completely fair – we want predictors to behave well even for these students and thus we want to take these predictions into account.

4.3 Averaging Across Skills

The basic difference between global computation and averaging across skills is the same as with the averaging across students – each method distributes differently weights to the available data points. The issue can be again quite pronounced for practical systems, as often the distribution of responses among skills is very highly uneven – in the case of skills even more than in the case of students, popular basic skills often have orders of magnitude more responses than specific or advanced skills.

The AUC metric again brings some specific issues. When AUC is computed per skill, the metric does not require any calibration of the model, since the metric only considers relative ordering of predictions. For example, in our simple error curve model, the only important aspect of predictions is that they are decreasing, it does not matter what is the exact shape of curve (all decreasing curves lead to the same value of the AUC metric). The global computation of the AUC metric takes into account the relative calibration among skills, e.g., if predictions for one of the skills are too low relative to other skills, it will decrease the metric value. However, in this case the overall AUC is easily dominated by "differences among skills" with only limited effect of the "ability to predict within skill".

As a specific example, consider the case demonstrated in Figure 3. We generate the data with a model with skills A1 and B2 (full lines). For model comparison we consider three models and compare the AUC values when computed globally and averaged across skills. A model with "speed of learning mismatch" (using skills A2, B1) achieves the same performance as the correct model when AUC is computed per skill, whereas for globally computed AUC it achieves poor performance (because the curves A2 and B1 cross, whereas the correct curves A1 and B2 do not). A model which uses one correct skill (A1) and one very poor skill (C, which models "negative" learning) has the results other way around – it achieves very poor AUC when computed across skills (due to the inappropriate model of negative learning), whereas it achieves quite good AUC when computed globally (due to large differences between skills which are captured correctly in the model).

5 DISCUSSION

As the presented examples clearly demonstrate, there can be large differences between different methods ("global", "per skill", "per student") of computation of metrics of predictive accuracy. A natural question is: "Which method is the correct one?" Unfortunately,

there is no simple answer to this question – the choice of an appropriate method depends on the specific use case. In some applications we may care mainly about "long-term users" and we do not worry about users who just try a system for a short while, e.g., for systems used schoolwide in a formal educational settings. In other cases the "initial impression" is important and we want the model to work well even for users with few responses, e.g., for commercial systems targeting individual students where the initial impression influences the decision whether to buy a licence. Each of these cases requires different approach to evaluation of predictive accuracy.

Thus the solution is not to choose a single universal metric and to apply it in all user modeling research. The choice of metric, however, clearly deserves more attention in research. Researchers should provide rationale for the choice of metric and also enough technical details about the computation of the metric to make their research reproducible. Our analysis of literature suggest that, at least in student modeling, the current state-of-the-art is inadequate in this respect – in many cases it is not possible to determine whether the reported metric was computed globally or averaged over skills or students.

Our examples also show that the AUC metric can be potentially misleading in several ways. Some of these features have been already noted in research outside of user modeling. In user modeling, however, the AUC metric remains to be heavily used and in many studies it is the only metric that is reported. In the light of discussed deficiencies, these kinds of results should be reevaluated using other metrics and taking into account different methods of metric averaging.

The points raised in this paper are relevant not only to educational applications of student modeling or to the two specific metrics discussed. The issues described illustrate that great care has to be taken in evaluation of user models and that it is necessary to pay attention to all details of evaluation.

REFERENCES

[1] Ryan Baker. 2013. A'/AUC Code. http://www.columbia.edu/~rsb2162/edmtools.html. (2013).

[2] Ryan SJ Baker, Albert T Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Proc. of Intelligent Tutoring Systems*. Springer, 406–415.

[3] Joseph Beck. 2007. Difficulties in inferring student knowledge from observations (and why you should care). In *Proc. of Educational Data Mining*. 21–30.

[4] Joseph E Beck and Kai-min Chang. 2007. Identifiability: A fundamental problem of student modeling. In *User Modeling 2007*. Springer, 137–146.

[5] Joseph E Beck and Xiaolu Xiong. 2013. Limits to accuracy: How well can we do at student modeling. In *Educational Data Mining*. 4–11.

[6] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[7] Asif Dhanani, Seung Yeon Lee, Phitchaya Phothilimthana, and Zachary Pardos. 2014. *A comparison of error metrics for learning model parameters in bayesian knowledge tracing*. Technical Report. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley.

[8] Tom Fawcett. 2004. ROC graphs: Notes and practical considerations for researchers. *Machine learning* 31, 1 (2004), 1–38.

[9] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[10] James Fogarty, Ryan S Baker, and Scott E Hudson. 2005. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *Proc. of Graphics Interface 2005*. 129–136.

[11] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

[12] Yue Gong, Joseph E Beck, and Neil T Heffernan. 2010. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Intelligent Tutoring Systems*. Springer, 35–44.

[13] Yue Gong, Joseph E Beck, and Neil T Heffernan. 2011. How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis. *International Journal of Artificial Intelligence in Education* 21, 1-2 (2011), 27–46.

[14] JP González-Brenes, Yun Huang, and Peter Brusilovsky. 2014. General features in knowledge tracing: Applications to multiple subskills, temporal item response theory, and expert knowledge. In *Proc. of Educational Data Mining*. 84–91.

[15] José P González-Brenes. 2015. Modeling Skill Acquisition Over Time with Sequence and Topic Modeling.. In *AISTATS*.

[16] José P González-Brenes and Yun Huang. 2015. Your model is predictive - but is it useful? theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. In *Proc. of Educational Data Mining*.

[17] José P González-Brenes and Jack Mostow. 2013. What and when do students learn? Fully data-driven joint estimation of cognitive and student models. In *Proc. of Educational Data Mining*. 236–240.

[18] Thomas M Hamill and Josip Juras. 2006. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society* 132, 621C (2006), 2905–2923.

[19] David J Hand. 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning* 77, 1 (2009), 103–123.

[20] T. Käser, S. Klingler, A. G. Schwing, and M. Gross. 2014. Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks. In *Proc. of Intelligent Tutoring Systems*. 188–198.

[21] Tanja Käser, Kenneth R Koedinger, and Markus Gross. 2014. Different parameters - same prediction: An analysis of learning curves. In *Proc. of Educational Data Mining*. 52–59.

[22] Mohammad Khajah, Robert V Lindsey, and Michael C Mozer. 2016. How deep is knowledge tracing?. In *Proc. of Educational Data Mining*.

[23] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography* 17, 2 (2008), 145–151.

[24] Caren Marzban. 2004. The ROC curve and the area under it as performance measures. *Weather and Forecasting* 19, 6 (2004), 1106–1114.

[25] Allan H Murphy. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* 12, 4 (1973), 595–600.

[26] Juraj Nižnan, Radek Pelánek, and Jiří Řihák. 2015. Student Models for Prior Knowledge Estimation. In *Educational Data Mining*.

[27] J. Papoušek, R. Pelánek, and V. Stanislav. 2014. Adaptive Practice of Facts in Domains with Varied Prior Knowledge. In *Educational Data Mining*. 6–13.

[28] Zachary A Pardos, Yoav Bergner, Daniel T Seaton, and David E Pritchard. 2013. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. In *Proc. of Educational Data Mining*. 137–144.

[29] Zachary A Pardos, Sujith M Gowda, Ryan SJ Baker, and Neil T Heffernan. 2012. The sum is greater than the parts: ensembling models of student knowledge in educational software. *ACM SIGKDD explorations newsletter* 13, 2 (2012), 37–44.

[30] Zachary A Pardos and Neil T Heffernan. 2011. KT-IDEM: Introducing item difficulty to the knowledge tracing model. *User Modeling, Adaption and Personalization* (2011), 243–254.

[31] Zachary A Pardos and Michael V Yudelson. 2013. Towards Moment of Learning Accuracy. In *AIED 2013 Workshops Proceedings Volume 4*. 3.

[32] Radek Pelánek. 2015. Metrics for Evaluation of Student Models. *Journal of Educational Data Mining* 7, 2 (2015).

[33] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*. 505–513.

[34] Michael Sao Pedro, Ryan Shaun Baker, and Janice D Gobert. 2013. Incorporating Scaffolding and Tutor Context into Bayesian Knowledge Tracing to Predict Inquiry Skill Acquisition.. In *Proc. of Educational Data Mining*. 185–192.

[35] Zoltan Toth, Olivier Talagrand, Guillem Candille, and Yuejian Zhu. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, Chapter Probability and ensemble forecasts, 137–163.

[36] Yutao Wang and Joseph Beck. 2013. Class vs. Student in a Bayesian Network Student Model. In *Artificial Intelligence in Education*. Springer, 151–160.

[37] Yutao Wang and Neil Heffernan. 2013. Extending knowledge tracing to allow partial credit: using continuous versus binary nodes. In *Artificial Intelligence in Education*. Springer, 181–188.

[38] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized Bayesian Knowledge Tracing Models. In *Artificial Intelligence in Education*. Springer, 171–180.