Metrics for Evaluation of Student Models

Radek Pelánek
Masaryk University Brno
pelanek@fi.muni.cz

Researchers use many different metrics for evaluation of performance of student models. The aim of this paper is to provide an overview of commonly used metrics, to discuss properties, advantages, and disadvantages of different metrics, to summarize current practice in educational data mining, and to provide guidance for evaluation of student models. In the discussion we mention the relation of metrics to parameter fitting, the impact of student models on student practice (over-practice, under-practice), and point out connections to related work on evaluation of probability forecasters in other domains. We also provide an empirical comparison of metrics. One of the conclusion of the paper is that some commonly used metrics should not be used (MAE) or should be used more critically (AUC).

1. Introduction

A key part of adaptive educational systems are models that estimate knowledge of students. To compare and improve these models we use metrics that measure quality of model predictions; these metrics are also called scoring rules (Gneiting and Raftery, 2007). Metrics are also used (sometimes implicitly) for parameter fitting, since many fitting procedures aim to optimize parameters with respect to some metric (e.g., log-likelihood). There is no single universal metric for model evaluation and thus researchers have to decide which metric to use. The choice of a metric is an important step in the research process. Differences in predictions between competing models are often small and the choice of a metric can influence results more than the choice of a parameter fitting procedure. Moreover, fitted model parameters are often used in subsequent steps in educational data mining and thus the choice of a metric can indirectly influence many other aspects of research.

This problem of performance metric choice is not specific only to student modeling. Particularly for classification problems and probabilistic predictions there are many possible metrics and it is not obvious which one to use in experiments and thus the issue has been thoroughly discussed in literature. Metrics have been studied (including experimental evaluation of relationships between metrics) in the general context of machine learning (Caruana and Niculescu-Mizil, 2004; Ferri et al., 2009), with special attention to behaviour of metrics for imbalanced sets (Jeni et al., 2013). For specific domains researchers have provided discussions of metrics specifically for a particular domain, e.g., Herlocker et al. (2004) discuss metrics for evaluation of recommender systems, Liu et al. (2011) provide overview of metrics used in ecology (species distribution models) with a focus on statistical tests for model comparison. Particularly evaluation of models for weather forecasting (Toth et al., 2003) can provide interesting inspiration for evaluation of student models, e.g., the concept of Brier score decomposition (discussed in Section 5.1.).

In the area of student modeling the discussion of performance metric has been very limited so far. Despite the importance of metrics and absence of consensus on their usage, the topic gets very little attention in most research papers. Many authors do not provide any rationale for their choice of a metric[1]. Sometimes it is not even clear what metric is exactly used (particularly in the case of R^2 metric), so it may be even difficult to use the same metric as previous authors.

The main aim of this paper is to fill this gap and to provide an overview of performance metrics relevant for evaluation of student models and to discuss issues specific to student modeling. One of the goals is also to raise awareness of issues that are not well-known in educational data mining community (e.g., that MAE is not a proper score). We also provide an empirical comparison of metrics and discuss whether small differences in performance metric matter – an issue raised for example by Beck and Xiong (2013). Based on the arguments in the paper we finally provide summary of specific recommendations for future evaluations of student models, e.g., the MAE metric should not be used for evaluation and the AUC metric should be used more critically (based on the intended use of studied model).

2. Types of Student Models

Before we try to assess which metrics are suitable for evaluation of student models, it is useful to discuss main types of student models and their typical uses.

2.1. Skill Modeling

The most often used type of student models (Desmarais and de Baker, 2012) are models of student skills; typical example is Bayesian knowledge tracing (Corbett and Anderson, 1995). These models predict performance of students; their evaluation is done by comparing the predicted performance and the actual observed performance (using suitable metric). In most cases the observed performance is binary, typically correctness of an answer to an exercise. Sometimes the performance measure can have multiple values or be continuous (taking into account hint use or response time).

Skill models are used to guide adaptive behaviour of educational systems, particularly for mastery learning (deciding whether a student has reach mastery of a topic) and for selection of examples of appropriate difficulty (Klinkenberg et al., 2011; Papoušek et al., 2014). These applications use directly model prediction. However, we are often interested not only in model predictions, but also in model parameters (e.g., problem difficulty, degree of learning). Model parameters may provide useful feedback to content authors, system developers, and students, e.g., in the form of open learner models (Bull, 2004). Skill models may be also used for discovery with models (Baker and Yacef, 2009), i.e., models are used in another analysis or in higher level modeling (Beck and Mostow, 2008; Cocea et al., 2009).

2.2. Models of Affect and Motivation

In addition to skill modeling, educational systems recently try to model also other aspects of student behaviour, particularly affective states like boredom, concentration, confusion, or frustration (Baker et al., 2012), and behaviours like gaming the system (Baker et al., 2004). Evalua-

[1] For example a recent EDM paper (González-Brenes et al., 2014) uses sophisticated modeling and provides extensive discussion of experimental results, but uses only the AUC metric without any justification for the choice of the metric (beside the fact that it is “a popular machine learning metric”).

tion of these models is done by comparing the predicted state (e.g., "bored" versus "not bored") with the observed state (by human observers) or by evaluating agreement between two types of detectors, e.g., with and without physiological sensors.

Models of affective states may be used as an additional input for the choice of exercises (e.g., adjusting difficulty level to the actual student affective state), or to provide personalized feedback to students using text messages or animated agents (Arroyo et al., 2014). Models can also be used as a tool for evaluation of educational systems and as a guide in their development (e.g., hint availability).

3. Overview of Metrics

As the discussion of student models shows, in student modeling the most common type of models predict binary events (correctness of answer or student state). To attain clear focus and readability we will in the following discuss only this type of models. Extensions of discussed metrics for multiple classes are described for example by Ferri et al. (2009). In the case of continuous predictions (e.g., response time) the choice of a metric is usually simpler. Common choice is root mean square error, which is in this case equivalent to log-likelihood with the assumption of normally distributed noise (Bishop, 2006). Previous experience in student modeling with continuous predictions suggest that even other choices of metric do not influence results of model comparison (Jarušek and Pelánek, 2012).

In the description of metrics we use the following notation. 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 $o_i \in \{0, 1\}$. A model performance metric is a function $f(\vec{p}, \vec{o})$. 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.

The standard terminology of metrics is, unfortunately, slightly confusing. Some metrics (MAE, RMSE) are "errors" (lower is better), others (accuracy, AUC, likelihood-based metrics) are "rewards" (higher is better). This lack of clear convention unnecessarily complicates understandability of research results. However, definitions of these metrics are so firmly established that we do not try to fight the tradition and provide definitions in their standard (non-systematic) way.

Note that since we are interested in using metrics for comparison, monotone transformations (square root, logarithm, multiplication by constant) are inconsequential and are used mainly for better interpretability or for traditional reasons.

In the following we use classification of metrics into three families as proposed by Ferri et al. (2009): probabilistic understanding of errors, qualitative understanding of errors, and assessing ranking of examples.

3.1. Probabilistic Understanding of Errors

The first set of metrics is based on probabilistic understanding of predictions p_i and of errors, i.e., difference between p_i and o_i. In the case of student modeling this type of metrics is natural mainly for predictions of performance (correctness of answers).

In other domains (e.g., decision analysis, weather forecasting) a summary measure for the evaluation of probabilistic forecasts is called a scoring rule and a useful notion of *proper* scoring

Table 1: Most often used metrics based on probabilistic understanding of errors. For MAE and RMSE lower is better, for LL higher is better.

Mean Absolute Error	MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert o_i - p_i\rvert$
Root Mean Square Error	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2}$
Log-likelihood	LL	$\sum_{i=1}^{n} o_i \log(p_i) + (1 - o_i)\log(1 - p_i)$

rules is studied (Gneiting and Raftery, 2007). Scoring rule is just another name for a performance metric[2] – a function S which for a predictive distribution $\vec{p}$ and an actual outcome $\vec{o}$ assigns a reward $S(\vec{p}, \vec{o})$. By $S(\vec{p}, \vec{q})$ we denote the expected value of $S(\vec{p}, \cdot)$ under distribution $\vec{q}$. The scoring rule is said to be proper if $S(\vec{q}, \vec{q}) \geq S(\vec{p}, \vec{q})$ for all $\vec{p}$ and $\vec{q}$ (strictly proper if the equality holds only if $\vec{p} = \vec{q}$). The intuition behind this definition is that if the forecaster's best judgment is the distributional forecast $\vec{q}$, he has no incentive to predict any $\vec{p} \neq \vec{q}$ (Gneiting and Raftery, 2007).

Table 1 shows definition of most commonly used metrics based on probabilistic understanding of errors. Mean absolute error (MAE) considers absolute differences between predictions and answers. This is not a suitable performance metric, because it prefers models which are biased towards the majority result. In the context of scoring rules this function is also called "linear score" and is well-known to be an improper scoring rule (Gneiting and Raftery, 2007), i.e. it may lead to misleading conclusions (specific example is discussed below). Despite this clear disadvantage, MAE is sometimes used for evaluation of student models (Cen et al., 2006; Pardos and Heffernan, 2010; Qiu et al., 2011).

A similar metric, root mean square error (RMSE), is obtained by using squared values instead of absolute values. As opposed to MAE, RMSE is a proper score (Gneiting and Raftery, 2007). Note that from the perspective of model comparison, the important part is only the sum of square errors. The square root in RMSE is traditionally used to get the result in the same units as the original "measurements" and thus to improve interpretability of the resulting number. In the particular context of student modeling and evaluation of probabilities, this is not particularly useful, since the resulting numbers are hard to interpret anyway. In order to get better interpretability, researchers sometimes use R^2 metric: $R^2 = 1 - \sum_{i=1}^{n}(o_i - p_i)^2 / \sum_{i=1}^{n}(o_i - \bar{o})^2$. With respect to comparison of models, R^2 is equivalent to RMSE since the only model dependent part is again the sum of square errors. In context of the standard linear regression (where it is most commonly used) R^2 has nice interpretation as "explained variability". In the case of logistic regression (which is closely connected to student models) this interpretation does not hold and different "pseudo R^2" metrics are used (e.g., Cox and Snell, McFadden, Nagelkerke). Thus the disadvantage of R^2 is that unless researchers are explicit about which version of R^2 they use (often they are not), a reader cannot know for sure which metric is reported.

In educational data mining the use of RMSE metric is very common, particularly for evaluation of skill models (Beck and Xiong, 2013; Gong et al., 2010; Wang and Beck, 2013; Wang and Heffernan, 2013; Yudelson et al., 2013; Papoušek et al., 2014; Nižnan et al., 2015). RMSE was also used as a metric in the KDD Cup 2010 focused on student performance evaluation. In other

[2] As opposed to metrics, scoring rules are consistently used as rewars, i.e., higher is better.

Table 2: Expected values of MAE, RMSE, and LL on a simple example.

	Model A	Model B
MAE	$0.7 \cdot \lvert 1-0.7\rvert + 0.3 \cdot \lvert 0-0.7\rvert = 0.42$	$0.7 \cdot \lvert 1-0.9\rvert + 0.3 \cdot \lvert 0-0.9\rvert = 0.34$
RMSE	$0.7 \cdot (1-0.7)^2 + 0.3 \cdot (0-0.7)^2 = 0.21$	$0.7 \cdot (1-0.9)^2 + 0.3 \cdot (0-0.9)^2 = 0.25$
LL	$n \cdot 0.7 \cdot \log(0.7)+$ $n \cdot 0.3 \cdot \log(1-0.7) = -0.61n$	$n \cdot 0.7 \cdot \log(0.9)+$ $n \cdot 0.3 \cdot \log(1-0.9) = -0.76n$

domains (particularly in weather forecasting) the mean square error (RMSE without the square root) is called a Brier score (Brier, 1950; Toth et al., 2003) or a quadratic scoring rule[3] (Gneiting and Raftery, 2007). The Brier score is sometimes decomposed into additive components (Murphy, 1973), which provide further insight into behaviour of predictive models. We discuss this issue in more detail in Section 5.

Another related metric is based on the notion of likelihood. The likelihood of data (observed answers) given the model (predicted probabilities) is $L = \prod_{i=1}^{n} p_i^{o_i} \cdot (1-p_i)^{(1-o_i)}$. For reasons of numerical stability and indifference of our concerns to monotonic transformations we typically work with log-likelihood (LL), i.e., the logarithm of the likelihood (Table 1). Note that although the behaviour of this metric is similar to RMSE, the absolute values are interpreted in completely different way than RMSE. Log-likelihood values are negative, it is not averaged (i.e., it decreases with the size of data set), and it has interpretation as a reward (higher is better).

The LL metric can also be interpreted from information theoretic perspective as a measure of data compression provided by a model (Roulston and Smith, 2002), i.e., from theoretical perspective this metric has better foundations than RMSE (Jewson, 2003). The LL metric is used in student modeling (Khajah et al., 2014; Pavlik et al., 2009), but much less frequently than RMSE. Its use is often connected with extensions like Akaike information criterion (AIC) or Bayesian information criterion (BIC). These metrics penalize large number of model parameters and thus aim to avoid overfitting. In the context of student modeling it is typically much better to address the issue of overfitting by cross-validation (Stamper et al., 2013). Since AIC and BIC provide faster way to asses models than cross-validation, they may be useful as heuristics in some algorithms (Cen et al., 2006; Stamper et al., 2013).

The main part of MAE, RMSE, and LL is in all cases "sum of error contributions for individual predictions". The metrics differ in the function which specifies the error contribution, Figure 1 shows graphically this difference. Note mainly the comparison between RMSE and LL – they are quite similar, the main difference is in the interval [0.95, 1], i.e., in cases where the predictor is confident and wrong. These cases are penalized very prohibitively by LL, whereas RMSE is relatively benevolent. In fact the LL metric is unbounded, so single wrong prediction (if it is too confident) can ruin the performance of a model. This property is usually undesirable and an artificial bound is used (bound 0.01 is used in a below presented experiment). Note that in the context of student modeling this corresponds to forcing a possibility of a slip and guess behaviour into a model. After this modification the error contributions for RMSE and LL are rather similar. Nevertheless, the LL approach "penalize mainly predictions which are confident and wrong" is reasonable and it can be argued that it is preferable to RMSE (Jewson, 2003).

[3] Quadratic scoring rule differs from mean square error by some technical transformations, which change it from an error to a reward.

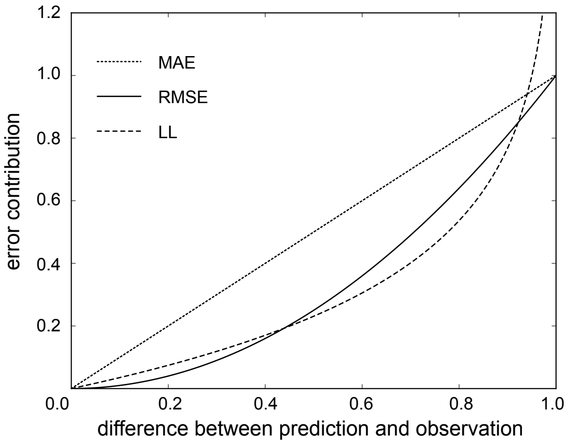

Figure 1: Comparison of error contributions for individual predictions for MAE, RMSE, LL metrics. For LL metric the function $\log(1-x)$ is scaled by $\frac{1}{3}$ to make it better comparable to the function x^2 for RMSE.

Although MAE seems quite similar to RMSE and LL, there is a fundamental difference between these metrics as MAE is an improper score and RMSE and LL are proper scores. As this fact is not commonly known in educational data mining community, it is worthwhile to illustrate the improper behaviour in an explicit situation. As a simple illustration consider a simulated student who answers correctly with constant probability 0.7. Let us compare two simple student models: Model A always predicts probability of correct answer 0.7, Model B always predicts probability of correct answer 0.9. Table 2 shows expected values of MAE, RMSE and LL for these two models. Model B achieves better MAE, but in this case Model A is clearly a better model, since it is equivalent to the ground truth. Thus the use of MAE leads to a wrong conclusion. Both RMSE and LL choose the correct model, in fact it is easy to verify that Model A achieves optimal RMSE and LL.

3.2. Qualitative Understanding of Errors

Another set of metrics is based on qualitative understanding of errors, i.e., either the prediction is correct or incorrect (0-1 loss). In student modeling this approach is suitable mainly for predictions of student state (e.g., boredom). In other domains this type of metrics is used mainly for evaluation of classification tasks in pattern recognition and information retrieval.

When we use qualitative understanding of errors, predictions have to classified into just two classes. If the predictions are in the interval $p_i \in [0,1]$, the classification can be done easily by choosing a threshold and doing the classification by comparison to this threshold. Once predictions are binarized, they can be classified as as true/false positives/negatives by a confusion matrix (Table 3). Qualitative metrics are then defined using the values from this matrix, the most common ones are shown in Table 4. The F1-score is a harmonic mean of precision and recall. The kappa statistic was originally proposed to measure interrater agreement as an improvement over agreement by chance.

As opposed to metrics based on probabilistic understanding of errors, values of these metrics depend on an external parameter – the choice of the classification threshold. They are also sensitive to prediction only with respect to this threshold. For example if we use a threshold 0.5, predictions 0.49 and 0.51 are considered different, whereas predictions 0.51 and 0.99 are

Table 3: Confusion matrix.

		Observed (true class)	
		Positive	Negative
Predicted	Positive	true positive (TP)	false positive (FP)
	Negative	false negative (FN)	true negative (TN)

Table 4: Metrics associated with a confusion matrix (for all of these metrics higher is better).

Accuracy	$(TP + TN)/n$
Precision	$TP/(TP + FP)$
Recall (sensitivity)	$TP/(TP + FN)$
F-measure (F1 score)	$2TP/(2TP + FP + FN)$
Kappa statistic	$(Accuracy - R)/(1 - R)$
	R = $((TP + FN)(TP + FP) + (TN + FP)(TN + FN))/n^2$

treated as same. Such behaviour can be for many student modeling applications undesirable. It is possible to use variants of these metrics also in a threshold independent way, e.g., to use maximum accuracy (Liu et al., 2011), but this approach is not common in student modeling.

These types of metrics are mainly used for evaluation of models of affective states (concentration, confusion, boredom, frustration, joy, distress), researches use mainly the kappa metric (San Pedro et al., 2013; Baker et al., 2012; D'Mello et al., 2008), or accuracy (Conati and Maclaren, 2009). Accuracy has also been used for evaluation of skill models (Pardos and Yudelson, 2013; Käser et al., 2014), but for this type of models it is much less appropriate and common.

3.3. Assessing Ranking of Examples

The third possible approach to evaluation of predictions takes into account ranking of predictions, i.e., the values of p_i are considered relatively to each other. There is only one commonly used approach of this type – receiver operating characteristics (ROC) curve and the related area under the ROC curve (AUC) metric.

The ROC curve summarizes the qualitative error of the prediction model over all possible thresholds. The curve has "false positive rate" $FP/(FP + TN)$ on the x-axis and "true positive rate" $TP/(TP + FN)$ on the y-axis, each point of the curve corresponds to a choice of a threshold; for a detailed introduction to ROC curve construction and interpretation see Fawcett (2006). 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).

Figure 2 shows examples of ROC curves for two student models evaluated over data from geography learning application (Papoušek et al., 2014). For illustration the figure explicitly highlights the false and true positive rates for three selected threshold values. Note that although the two curves have similar overall shape, locations of points for individual thresholds significantly differ for the two models. The AUC value is 0.792 for Model 1 and 0.784 for Model 2.

The area under the curve can be approximated using a metric called A' (Fogarty et al., 2005):

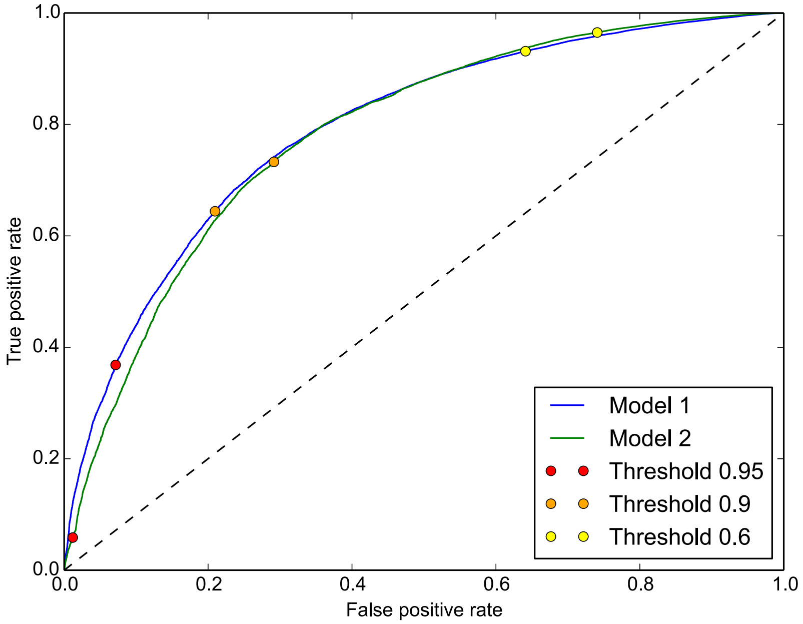

Figure 2: Illustration of ROC curves.

$A' = 1/(|C||I|) \cdot \sum_{i \in C} \sum_{j \in I} v(p_i, p_j)$, where $C = \{i | o_i = 1\}$ (positive observations), $I = \{i | o_i = 0\}$ (negative observations), $v(x, y) = 1$ if $x > y$, $v(x, y) = 0.5$ if $x = y$, and $v(x, y) = 0$ if $x < y$. This A' metric is equivalent to the well-studied Wilcoxon statistics (Fogarty et al., 2005), which provides ways to study statistical significance of results (but requires attention to assumptions of the tests, e.g., independence).

The ROC curve and the AUC metric are successfully used in many different research areas, but their use is sometimes criticized (Lobo et al., 2008; Hand, 2009), e.g., because the metric summarizes performance over all possible thresholds, even over those for which the classifier would never be practically used. From the perspective of student modeling it is important to take into account that the AUC metric considers predictions only in relative way – if all predictions are divided by 2, the AUC metric stays the same. For this reason the AUC metric should not be used (as the only metric) in cases where we need absolute values of predictions to be well calibrated.

To address disadvantages of the AUC metric, several extension have been proposed, e.g., scored AUC (Ferri et al., 2009), which incorporates absolute values of probabilities in the definition, and partial AUC (Dodd and Pepe, 2003), which considers only part of the ROC curve that is relevant for a particular application. These extensions are currently not commonly used in student modeling, although they may be useful (e.g., in student modeling we are often interested mainly in model decisions with high thresholds).

In student modeling the AUC (resp. A') metric is often used for evaluating models of affective states (San Pedro et al., 2013; Baker et al., 2012), and related student behaviour like gaming the system (Baker et al., 2004; Baker et al., 2008). These are natural classification problem and the use of AUC in these cases seems appropriate. The AUC metric, however, is also widely used for evaluation of skill models (Baker et al., 2008a; Beck and Chang, 2007; Baker et al., 2008b; Baker et al., 2010; González-Brenes and Mostow, 2013; Pardos et al., 2013; Pardos et al., 2012; Pardos and Heffernan, 2011; Sao Pedro et al., 2013; González-Brenes et al., 2014).

The suitability of AUC in these cases is disputable, particularly in cases where AUC is used as the only metric, since applications of skill models often need well calibrated absolute values of predictions. Moreover, in some cases AUC is used as the only metric for final evaluation, but the parameter fitting procedure uses (implicitly) different metric (RMSE or LL). Particularly in cases of brute force fitting this approach is strange. If AUC is really a proper metric for a specific application, it should be used not just for final evaluation, but also for parameter fitting (this may not be easy to do with gradient descent or expectation maximization, but it is easy to realize in case of brute force search).

4. Metrics and Usefulness of Models

Predictive ability of student models is not an end in itself, but mainly a mean to improving behaviour of educational systems and for getting insight into the learning process. Now we look at the relation of performance metrics to these goals.

4.1. Impact on Student Practice

Model predictions are often used to guide the behaviour of an adaptive educational systems. The goal of these systems is to provide efficient learning. Models are used mainly to select practice opportunities and the goal of efficient learning thus means choosing suitable practice items. This is often operationalized as minimizing over-practice (practice of items that the student already mastered, i.e., "wasted time" of students) and under-practice (missing practice that is necessary for mastery of a topic and for further progress).

Over-practice and under-practice are clearly more important than predictive ability of models , but they cannot be measured directly. Since we do not have any direct information about student mastery, we can only make indirect inferences about over-practice and under-practice. Moreover, the amount of over-practice and under-practice depends not only on the student model, but also on other aspects of the educational system, e.g., settings of thresholds for mastery learning.

Nevertheless, it is important to study over-practice and under-practice and to analyze their relation to model performance metrics to see if the easily measurable performance metrics provide a good proxy measure for these more important goals. Previous research (Yudelson and Koedinger, 2013) demonstrated on real student data that small differences in predictive performance can lead to significant impact on student under-practice and over-practice (7-20%). Other authors (Fancsali et al., 2013; Fancsali et al., 2013; Lee and Brunskill, 2012; Pardos and Yudelson, 2013; Pelánek, 2015) performed experiments with simulated data (in most cases using Bayesian knowledge tracing model) to investigate the impact of different models or different mastery thresholds on the amount of student practice. Pardos and Yudelson (2013) used simulated data to study the relation between different performance metrics and identification of the "moment of learning" (step in which a student learned a skill). They obtained good results for the RMSE metric and poor results for the AUC metric.

4.2. Impact on Parameter Fitting

Student models are valuable not only for their predictions, but also for their parameters. Even through "black box" techniques (like ensembles) can potentially improve predictive performance (Pardos et al., 2012), most researchers and system developers prefer interpretable models like Bayesian knowledge tracing. Model parameters can provide us with valuable insight into the learning process and can be used for interventions (Liu et al., 2014) or even in higher-level modeling (Hershkovitz et al., 2013). For example, Cen et al. (2007) discuss a specific case where analysis of model parameters leads to changes in the tutoring system reducing the over-practice of students.

Even if parameter fitting is the real goal of a particular application of student modeling, metrics of predictive performance are still important. Parameter fitting is typically done by optimizing performance of a model with respect to a chosen metric – although this step is often not explicitly mentioned, e.g., when researchers use the EM algorithm, which implicitly optimizes the LL metric. However, we should not rely too much on performance metric for analysis of model parameters. The fact that a model with particular parameter values achieves best results with respect to a choosen metric does not mean that these parameters are "correct" or even stable. Stability of parameters should get more attention and be studied for example with bootstraping experiments (estimating parameters from different samples of data).

A choice of metric with respect to which model parameters are optimized can have large impact on parameter values. Dhanani et al. (2014) performed experiments with simulated data based on the Bayesian knowledge tracing model. They studied the relation between metrics and a distance of the fitted parameters to ground truth parameters (parameters that were used to generate the data). The results show that RMSE and LL metrics are able to retrieve parameters well (RMSE being slightly better), whereas the AUC metric leads to poor results.

5. Beyond Single Number

Metrics summarize performance of a student model by a single number, which gives a simplified view of model behaviour. This simplification may be necessary for parameter fitting, but for model comparison and for better understanding of model behaviour it may be useful to quantify model performance in more detail. One commonly used possibility is to use several performance metrics and check whether they agree. In this way we obtain several 'views' of model performance, but they all still capture only the summary performance and give limited insight into details of model performance.

Another possibility is to decompose a particular metric into several components or decompose predictions into disjoint groups and analyze model performance for individual groups. We discuss some of these possibilities in more detail.

5.1. Brier Score Decomposition

In Section 3.1. we mentioned that mean square error (RMSE without the square root) is sometimes called the Brier score (Brier, 1950; Toth et al., 2003). In this context the metric is often decomposed into additive components (Murphy, 1973; Toth et al., 2003; Cohen and Goldszmidt, 2004), which provide further insight into the behaviour of the predictor.

Assume that predictions p_i take only c different values (if there is a large number of different values we can bin them into c classes, with c typically between 10 and 30). Let us denote these values q_k, the number of predictions in the same category n_k, and the observed frequency of these predictions $f_k = \sum_{i,p_i=q_k} o_i/n_k$. Finally, let f be the base rate (proportion of positive observations). Then the Brier score can be decomposed as (Toth et al., 2003):

$$\begin{aligned} BS &= \tfrac{1}{N}\textstyle\sum_k n_k(q_k - f_k)^2 - \tfrac{1}{N}\sum_k n_k(f_k - f)^2 + f(1-f) \\ BS &= REL - RES + UNC \end{aligned}$$

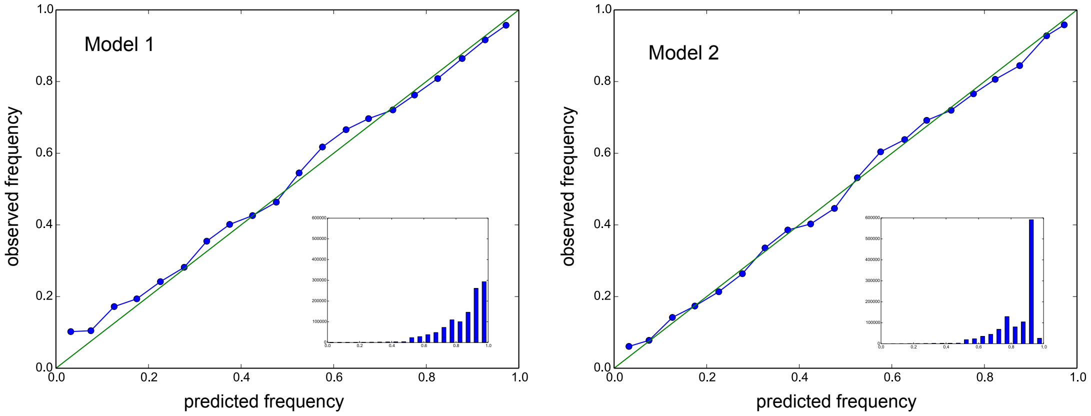

Figure 3: Reliability diagrams and sharpness graphs for two student models.

The first term (REL) is *reliability*, which measures the difference between predicted probabilities and observed probabilities. It is not difficult to achieve perfect reliability if the model just predicts the base rate of events. Useful model should also predict wide range of probabilities. This is captured by the second term (RES) called *resolution*. Resolution measures how much do predictions differ from the base rate. A good model needs both reliability and resolution. The third term (UNC) is *uncertainty*, which quantifies the inherent uncertainty of events. This part is independent of a model.

5.2. Reliability Diagrams

Closely related to Brier score decomposition are reliability diagrams (Bröcker and Smith, 2007; Niculescu-Mizil and Caruana, 2005; Toth et al., 2003). Reliability diagram shows predicted frequency versus observed frequency for individual classes (bins). The number of cases falling into individual bins is often highly uneven, so it is customary to show the reliability diagram together with a 'sharpness graph' (a histogram of number of cases in each bin), see Figure 3. The reliability graph is directly related to the reliability term in the Brier score decomposition, the sharpness graph is related to the resolution term. Bröcker and Smith (2007) provides thorough discussion of possibilities for the choice of bins and their centers, and also of methods for assessing reliability of reliability diagrams (e.g., with consistency bars).

Figure 3 gives a specific example of reliability diagrams for two student models (the same models which were used for construction of Figure 2). The difference in RMSE is only very small (Model 1: 0.344, Model 2: 0.345), but the behaviour of the models is actually quite different. Model 2 has quite poor resolution, with most predictions being around 0.93; Model 1 has better resolutions. Note that this difference can also be detected in the ROC curve when we highlight specific thresholds (Figure 2). Both models have good reliability; Model 2 being only slightly better. For Model 1 we can notice a trend in the reliability diagram – the model tends to slightly overestimate students when probability of correct answer is low, and underestimate students when probability of correct answer is high. Such observation can lead to an improvement of the student model or the learning system. Similar kind of analysis can be performed in terms of residual analysis, Käser et al. (2014) provides a specific example (with analogical observations about underestimation and overestimation).

6. Usage of Metrics

In this section we discuss some recurring issues about practical application of performance metrics.

6.1. Interpretability

Interpretability is sometimes an important reason why researchers use a particular metric. However, even for seemingly interpretable metrics like AUC, the interpretation can be difficult and misleading. Particularly when data are not homogeneous (widely different difficulties of items or skills of students), metrics can indicate good performance even through good results are achieved mainly due to the variability in data, which could be easy captured by a simple model. Hamill and Juras (2006) discuss this effect in detail in the context of weather forecasting. To make the point they present analysis of a hypothetical scenario with predicting weather on two islands with different climatology. Their scenario is analogical to student modeling for two classes of students with different background knowledge.

Even if we ignore this issue, interpretability is a wrong reason for use of a particular metric for model comparison or parameter fitting. In model comparison and parameter fitting we care only about relative performance of models and we should use metrics most suitable for the particular context. Specifically, for evaluation of skill models we should prefer RMSE or LL to AUC. The AUC metric may be reported additionally to provide a more complete (interpretable) picture of the evaluation results.

6.2. Does the choice of metric really matter?

Is the choice of metric really practically important? Do different metrics actually lead to different conclusions? We report an experiment inspired by comparison of metrics for evaluation of recommender systems (Herlocker et al., 2004).

To compare metrics we used the following experiment with simulated data. To generate data we used standard student models – Bayesian knowledge tracing (Corbett and Anderson, 1995) and additive factor model based on logistic function (Käser et al., 2014). Using different parameter settings we generated 20 datasets. The datasets differ in the number of simulated students and length of student traces. The total number of attempts is, however, same in all datasets (10,000). For predictions we used Bayesian knowledge tracing (Corbett and Anderson, 1995), Performance Factor Analysis (Pavlik et al., 2009), and adaptation of the Elo rating system (Pelánek, 2014). Using different parameter values we specified 15 concrete models. For each dataset and model (300 combinations) we evaluated the performance using the most commonly used metrics for evaluation of skill models: MAE, RMSE, LL, AUC.

Figure 4 shows scatter plots and correlation coefficients for all combinations of studied metrics. The presented comparison includes all combinations of datasets and models, including some cases of significant mismatch between data and a model. Such cases would not occur in practical situations. To test the relations among metrics on “good models”, we counted how often do metrics agree on the best model (from the given fixed set of 15 models). Table 5 shows these results.

The three metrics based on probabilistic understanding of errors (MAE, RMSE, and LL) have large overall correlations – this is not surprising, since they all have similar functional form (as discussed in Section 3.1.). Particularly RMSE and LL agree very strongly with each other,

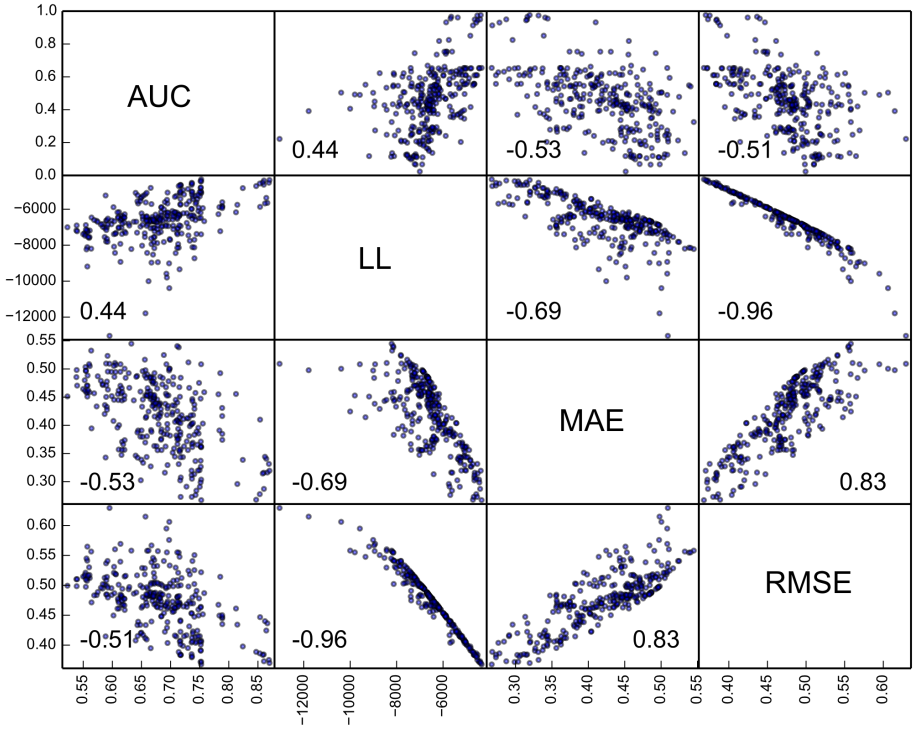

Figure 4: Correlations among metrics.

both in overall correlation and in the choice of the best model. The MAE metrics, on the other hand, has only good overall correlation with RMSE and LL, it does not agree on the choice of the best model, i.e., these result show that the difference between MAE and RMSE/LL occurs also in practice and not only in artificially constructed examples like the one in Table 2. The AUC metric behaves differently from RMSE and LL, both with respect to the overall correlation and with respect to the choice of the best model. These results show that the choice of metric is really important and can influence conclusions of experiments. An exception is the choice between RMSE and LL, which is pronounced neither in the reported experiment nor in the literature.

Table 5: Agreement in the choice of the best model (out of 20 cases).

	AUC	LL	MAE	RMSE
AUC	–	5	2	7
LL	5	–	2	17
MAE	2	2	–	3
RMSE	7	17	3	–

6.3. Do Small Differences in Metrics Matter?

Evaluations of student models often report only small differences in the final values of performance metrics (e.g., third decimal place in RMSE). A frequent topic in educational data mining research[4] (raised explicitly by Beck and Xiong (2013)) is how much do these small differences in predictive ability matter.

To be important, differences between models should either have impact on student practice or lead to interpretable and actionable results. Previous research have shown that even small differences in RMSE can have significant impact on student over-practice and under-practice (Yudelson and Koedinger, 2013) and can be interpretable and lead to improvements in the tutoring system (Liu et al., 2014).

As a practical tool for estimating the importance of model differences it is useful to analyze correlations of model predictions. Small differences in performance metric can be caused by quite different situations. The model predictions may be highly correlated, with one model systematically achieving slightly better predictions. Such result may be interesting, but in most cases it will have only minimal practical impact. On the other hand, it can also happen that models differ only slightly in a summary metric, but their predictions are not well correlated, i.e., individual predictions are actually quite different (at least in some cases). In such a case the model difference may be of significant practical importance.

As a specific example consider a hypothetical scenario, where we have two similar models based on Q-matrices (Barnes, 2005). One of them is ‘correct’ and has 10 skill (knowledge components), the other one is ‘incorrect’ and merges two of the skills together. The difference between performance metrics of these two models will be necessarily small, since in most cases their predictions will be identical. The difference is, however, of practical significance, because if a tutoring system uses the incorrect model, students may miss practice of one of the skills. A realistic scenario of this type is reported by Liu et al. (2014).

This issue may be further highlighted by a feedback loop between models and data collection (Nižnan et al., 2015). The data that are used for model evaluation are often collected by an intelligent tutoring system which uses mastery learning. The mastery is judged by a student model, i.e., the used model influences which data are collected and used for evaluation. So it can happen that the used model does not collect data that would show its deficiencies. As a specific example consider the above mentioned scenario. If the tutoring system uses the incorrect model with merged skills, then it may happen that students answer very few questions on one of the merged skills, because their mastery is declared based on their correct answers to the second merged skill. This distortion of collected data further reduces differences in performance metrics.

Note that the presence of this feedback loop is an important difference compared to other forecasting domains. In weather forecasting models do not directly influence the system and cannot distort collected data, in student modeling they can. This aspect of evaluation of student models deserves further attention.

[4] Note that this question is not specific to educational data mining, e.g., Herlocker et al. (2004) discuss the same issue in the context of collaborative filtering recommender systems.

7. Conclusions

Finally, let us summarize the main issues relevant for researchers who need to evaluate student models. In most cases there is not an unequivocal answer to the question "Which metric should I use for my experiments?" The choice of suitable metric depends on the intended use of a student model. Are the absolute values of estimated probabilities important? Is the relative ordering of predictions more relevant? Do we care only about binary classification? Many papers which propose or evaluate student models lack any discussion of the specific purpose of described models. Such discussion or specific arguments for a particular choice of metric should become standard part of papers that include evaluation of student models.

The arguments presented in this work suggest some guidelines for choice of metrics. A clear conclusion concerns the MAE metric. For predictors of binary outcomes, which are typical in student modeling, the MAE metric should not be used and reported at all, since it is not a proper score and can lead to misleading conclusions.

Metrics based on the confusion matrix and the AUC metric may be appropriate for evaluation of models of affective states, but they are less suitable for evaluation of skill models. Particularly the use of AUC metric deserves attention. It is widely used for evaluation of skill models as it is seemingly more interpretable than RMSE or LL. But with respect to skill models it has several important disadvantages: it considers only relative ordering of predictions, previous research indicates that it has poor relation to over-practice and under-practice (Pardos and Yudelson, 2013) and that it is unsuitable for optimizing parameter values (Dhanani et al., 2014). For particular application of skill models the use of the AUC metric may be justifiable, but the justification needs to be explicitly discussed (which is not a current practice).

A reasonable choice of metric for evaluation of skill models seems to be the LL metric or the RMSE metric. The RMSE metric is currently much more used than the LL metric, and it is a plausible choice due to its connection to Brier score decomposition and reliability graphs, which provide further insight into properties of predictions. The LL and RMSE metrics have similar form, the main difference occurs for predictions that are confident and wrong. This difference may be in some application quite important and the log-likelihood deserves more attention than it currently gets.

The issue of performance metrics is widely discussed in the context of general machine learning and applications in other domain (Caruana and Niculescu-Mizil, 2004; Ferri et al., 2009; Jeni et al., 2013; Gneiting and Raftery, 2007; Herlocker et al., 2004; Liu et al., 2011; Toth et al., 2003). The educational data mining community may find useful inspiration in this rich literature.

Acknowledgements

The author thanks anonymous reviewers for very helpful comments which led to significant improvement of the paper.

References

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., and Tai, M. 2014. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education 24,* 4, 387–426.

BAKER, R. S., CORBETT, A. T., AND ALEVEN, V. 2008a. Improving contextual models of guessing and slipping with a truncated training set. In *Educational Data Mining*. 67–76.

BAKER, R. S., CORBETT, A. T., AND ALEVEN, V. 2008b. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems*. Springer, 406–415.

BAKER, R. S., CORBETT, A. T., GOWDA, S. M., WAGNER, A. Z., MACLAREN, B. A., KAUFFMAN, L. R., MITCHELL, A. P., AND GIGUERE, S. 2010. Contextual slip and prediction of student performance after use of an intelligent tutor. In *User Modeling, Adaptation, and Personalization*. Springer, 52–63.

BAKER, R. S., CORBETT, A. T., AND KOEDINGER, K. R. 2004. Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems*. Springer Berlin Heidelberg, 531–540.

BAKER, R. S., CORBETT, A. T., ROLL, I., AND KOEDINGER, K. R. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction 18,* 3, 287–314.

BAKER, R. S., GOWDA, S. M., WIXON, M., KALKA, J., WAGNER, A. Z., SALVI, A., ALEVEN, V., KUSBIT, G. W., OCUMPAUGH, J., AND ROSSI, L. 2012. Sensor-free affect detection in cognitive tutor algebra. In *Educational Data Mining*. 126–133.

BAKER, R. S. AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining 1,* 1, 3–17.

BARNES, T. 2005. The q-matrix method: Mining student response data for knowledge. In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*.

BECK, J. E. AND CHANG, K.-M. 2007. Identifiability: A fundamental problem of student modeling. In *User Modeling 2007*. Springer, 137–146.

BECK, J. E. AND MOSTOW, J. 2008. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Intelligent Tutoring Systems*. Springer, 353–362.

BECK, J. E. AND XIONG, X. 2013. Limits to accuracy: How well can we do at student modeling. In *Educational Data Mining*. 4–11.

BISHOP, C. 2006. *Pattern recognition and machine learning*. Springer.

BRIER, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review 78,* 1, 1–3.

BRÖCKER, J. AND SMITH, L. A. 2007. Increasing the reliability of reliability diagrams. *Weather and forecasting 22,* 3, 651–661.

BULL, S. 2004. Supporting learning with open learner models. In *Information and Communication Technologies in Education*.

CARUANA, R. AND NICULESCU-MIZIL, A. 2004. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 69–78.

CEN, H., KOEDINGER, K., AND JUNKER, B. 2006. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems*. Springer, 164–175.

CEN, H., KOEDINGER, K. R., AND JUNKER, B. 2007. Is over practice necessary?-improving learning efficiency with the cognitive tutor through educational data mining. *Frontiers in Artificial Intelligence and Applications 158*, 511.

COCEA, M., HERSHKOVITZ, A., AND BAKER, R. S. 2009. The impact of off-task and gaming behaviors on learning: immediate or aggregate? In *Artificial Intelligence in Education*. 507–514.

COHEN, I. AND GOLDSZMIDT, M. 2004. Properties and benefits of calibrated classifiers. In *Knowledge Discovery in Databases: PKDD 2004*. Springer, 125–136.

CONATI, C. AND MACLAREN, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction 19,* 3, 267–303.

CORBETT, A. AND ANDERSON, J. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction 4,* 4, 253–278.

DESMARAIS, M. C. AND DE BAKER, R. S. J. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Model. User-Adapt. Interact. 22,* 1-2, 9–38.

DHANANI, A., LEE, S. Y., PHOTHILIMTHANA, P., AND PARDOS, Z. 2014. A comparison of error metrics for learning model parameters in bayesian knowledge tracing. Tech. rep., Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley.

D'MELLO, S. K., CRAIG, S. D., WITHERSPOON, A., MCDANIEL, B., AND GRAESSER, A. 2008. Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction 18,* 1-2, 45–80.

DODD, L. E. AND PEPE, M. S. 2003. Partial AUC estimation and regression. *Biometrics 59,* 3, 614–623.

FANCSALI, S. E., NIXON, T., AND RITTER, S. 2013. Optimal and worst-case performance of mastery learning assessment with bayesian knowledge tracing. In *Proceedings of the 6th International Conference on Educational Data Mining*.

FANCSALI, S. E., NIXON, T., VUONG, A., AND RITTER, S. 2013. Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED Workshops*. Citeseer.

FAWCETT, T. 2006. An introduction to roc analysis. *Pattern recognition letters 27,* 8, 861–874.

FERRI, C., HERNÁNDEZ-ORALLO, J., AND MODROIU, R. 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters 30,* 1, 27–38.

FOGARTY, J., BAKER, R. S., AND HUDSON, S. E. 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.

GNEITING, T. AND RAFTERY, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102,* 477, 359–378.

GONG, Y., BECK, J. E., AND HEFFERNAN, N. T. 2010. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Intelligent Tutoring Systems*. Springer, 35–44.

GONZÁLEZ-BRENES, J., HUANG, Y., AND BRUSILOVSKY, P. 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.

GONZÁLEZ-BRENES, J. P. AND MOSTOW, J. 2013. What and when do students learn? fully data-driven joint estimation of cognitive and student models. In *Proceedings of the 6th International Conference on Educational Data Mining*. 236–240.

HAMILL, T. M. AND JURAS, J. 2006. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society 132,* 621C, 2905–2923.

HAND, D. J. 2009. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning 77,* 1, 103–123.

HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., AND RIEDL, J. T. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS) 22,* 1, 5–53.

HERSHKOVITZ, A., DE BAKER, R. S. J., GOBERT, J., WIXON, M., AND SAO PEDRO, M. 2013. Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist 57,* 10, 1480–1499.

JARUŠEK, P. AND PELÁNEK, R. 2012. Analysis of a simple model of problem solving times. In *Proc. of Intelligent Tutoring Systems*. LNCS, vol. 7315. Springer, 379–388.

JENI, L. A., COHN, J. F., AND DE LA TORRE, F. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 245–251.

JEWSON, S. 2003. Use of the likelihood for measuring the skill of probabilistic forecasts. arXiv preprint physics/0308046.

KÄSER, T., KOEDINGER, K. R., AND GROSS, M. 2014. Different parameters-same prediction: An analysis of learning curves. In *Proc. of Educational Data Mining*. 52–59.

KHAJAH, M., WING, R. M., LINDSEY, R. V., AND MOZER, M. C. 2014. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *Proc. of Educational Data Mining*.

KLINKENBERG, S., STRAATEMEIER, M., AND VAN DER MAAS, H. 2011. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education 57,* 2, 1813–1824.

LEE, J. I. AND BRUNSKILL, E. 2012. The impact on individualizing student models on necessary practice opportunities. In *Educational Data Mining*. 118–125.

LIU, C., WHITE, M., AND NEWELL, G. 2011. Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography 34,* 2, 232–243.

LIU, R., KOEDINGER, K. R., AND MCLAUGHLIN, E. A. 2014. Interpreting model discovery and testing generalization to a new dataset. In *Educational Data Mining*. 107–113.

LOBO, J. M., JIMÉNEZ-VALVERDE, A., AND REAL, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography 17,* 2, 145–151.

MURPHY, A. H. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology 12,* 4, 595–600.

NICULESCU-MIZIL, A. AND CARUANA, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 625–632.

NIŽNAN, J., PELÁNEK, R., AND PAPOUŠEK, J. 2015. Exploring the role of small differences in predictive accuracy using simulated data. In *AIED Workshop on Simulated Learners*.

NIŽNAN, J., PELÁNEK, R., AND ŘIHÁK, J. 2015. Student models for prior knowledge estimation. In *Educational Data Mining*.

PAPOUŠEK, J., PELÁNEK, R., AND STANISLAV, V. 2014. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining*. 6–13.

PARDOS, Z. A., BERGNER, Y., SEATON, D. T., AND PRITCHARD, D. E. 2013. Adapting bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining*. 137–144.

PARDOS, Z. A., GOWDA, S. M., BAKER, R. S., AND HEFFERNAN, N. T. 2012. The sum is greater than the parts: ensembling models of student knowledge in educational software. *ACM SIGKDD explorations newsletter 13,* 2, 37–44.

PARDOS, Z. A. AND HEFFERNAN, N. T. 2010. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*. Springer, 255–266.

PARDOS, Z. A. AND HEFFERNAN, N. T. 2011. Kt-idem: Introducing item difficulty to the knowledge tracing model. In *User Modeling, Adaption and Personalization*. Springer, 243–254.

PARDOS, Z. A. AND YUDELSON, M. V. 2013. Towards moment of learning accuracy. In *AIED 2013 Workshops Proceedings Volume 4*. 3.

PAVLIK, P. I., CEN, H., AND KOEDINGER, K. R. 2009. Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of Artificial Intelligence in Education (AIED)*. Frontiers in Artificial Intelligence and Applications, vol. 200. IOS Press, 531–538.

PELÁNEK, R. 2014. Time decay functions and elo system in student modeling. In *Educational Data Mining*. 21–27.

PELÁNEK, R. 2015. Modeling student learning: Binary or continuous skill? In *Educational Data Mining*.

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. In *Educational Data Mining*. 139–148.

ROULSTON, M. S. AND SMITH, L. A. 2002. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review 130,* 6.

SAN PEDRO, M. O. Z., BAKER, R. S., GOWDA, S. M., AND HEFFERNAN, N. T. 2013. Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system. In *Artificial Intelligence in Education*. Springer, 41–50.

SAO PEDRO, M. A., BAKER, R. S., AND GOBERT, J. D. 2013. Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *Educational Data Mining*. 185–192.

STAMPER, J. C., KOEDINGER, K. R., AND MCLAUGHLIN, E. A. 2013. A comparison of model selection metrics in datashop. In *Educational Data Mining*. 284–287.

TOTH, Z., TALAGRAND, O., CANDILLE, G., AND ZHU, Y. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, Chapter Probability and ensemble forecasts, 137–163.

WANG, Y. AND BECK, J. 2013. Class vs. student in a bayesian network student model. In *Artificial Intelligence in Education*. Springer, 151–160.

WANG, Y. AND HEFFERNAN, N. 2013. Extending knowledge tracing to allow partial credit: using continuous versus binary nodes. In *Artificial Intelligence in Education*. Springer, 181–188.

YUDELSON, M. V. AND KOEDINGER, K. R. 2013. Estimating the benefits of student model improvements on a substantive scale. In *EDM 2013 Workshops Proceedings*.

YUDELSON, M. V., KOEDINGER, K. R., AND GORDON, G. J. 2013. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education*. Springer, 171–180.