Dueling Metrics: Choosing the Appropriate Error Metric for Models of Cognition in the Learning Analytics Field

Phitchaya Phothilimthana
Seung Yeon Lee
Zachary Pardos

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2018-7
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-7.html

April 15, 2018

A

Dueling Metrics: Choosing the Appropriate Error Metric for Models of Cognition in the Learning Analytics Field

Phitchaya Mangpo Phothilimthana, University of California, Berkeley
Seung Yeon Lee, University of California, Berkeley
Zachary A. Pardos, University of California, Berkeley

Similar to how a machine learning model converges by following the gradient produced by the choice of loss function, a scholarly field converges towards adoption of various model modification by following a type of gradient produced by the choice of error metrics used to report results in its papers. In this way, a field and its practitioners become a part of a larger human-centric process of design. In this paper we argue for the importance of choosing the right error metric for a popular cognitive model called Bayesian Knowledge Tracing (BKT), used in the context of intelligent tutoring systems. According to our analyses with synthetic data—including correlation analysis, gradient visualization, and parameter estimation—we find that error metrics of Root Mean Squared Error (RMSE) and log-likelihood provide the best correspondence to the true generating model. Area Under the Curve (AUC) and accuracy are significantly behind, while precision and recall have extremely poor performance. Our result validates the standard practices of using RMSE as a metric to evaluate BKT models and using RMSE or log-likelihood for BKT parameter estimation. Our result adds to the mounting wisdom against using AUC and accuracy, which are the other metrics that have been frequently used to evaluate BKT models as depicted in our seven-year literature review of the field. Additionally, we investigate the validity of parameters estimated using the different error metrics on real data from ASSISTments, Cognitive Tutor, and Khan Academy. The real data analysis reinforces our finding that log-likelihood and RMSE appear to be superior to the rest of the metrics and should be the metric of choice when applying this model.

CCS Concepts:•**Applied computing** → **Computer-assisted instruction;**

General Terms: Algorithms, Human Factors

1. INTRODUCTION

The development of a machine learning model occurs not just within the confine of a project limited to several contributors. In the case of a model specific to a discipline, a much larger community contributes to the development of the model. The choice of the error metric used in the evaluation of models can influence the magnitude of performance difference between one model and another, and change how the field selects which introduced extensions are kept and which are discarded. In this paper, we study a model frequently used in the Intelligent Tutoring Systems (ITS) and Educational Data Mining (EDM) field to make decisions about when a student has reached mastery of knowledge. We present an argument for which error metrics should and should not be used so that both the model and the field can better converge towards an evolution of the model that best correlates with the objective of improving human learning outcomes.

Mastery learning, in brief, is an instructional strategy whereby students are required to master prerequisite material before being allowed to advance in the curriculum [Bloom 1984]. Many computer-based tutoring programs, aspired by the benefits of this in-person instructional strategy, have implemented mastery learning in their systems in one form or another. Since these systems rely on a machine learning algorithm, instead of a human, to make inferences about cognitive mastery, the effectiveness of mastery learning depends largely on the accuracy of the inferences made by this algorithm. The most popular algorithm in the literature and in practice for modeling cognitive mastery in tutoring systems has been the Bayesian Knowledge Tracing (BKT) model, introduced by Corbett and Anderson [1994]. It has been applied to datasets from a variety of tutoring systems [Koedinger et al. 1997; Beck and Sison

2006] to predict student performance as well as to determine whether students have mastered a particular skill. A standard BKT model is characterized by four model parameters: *prior, learn, guess*, and *slip*. In contrast to other types of machine learning models such as support vector or neural network techniques, BKT is often preferred due to its interpretability. For instance, the BKT parameters can provide pedagogical insights (e.g., learning effect of a tutoring system); therefore, the values of these parameters are often of interest to researchers.

How do we identify whether a model's parameters and inferences accurately represents reality? Unfortunately, since student knowledge is not directly observed, we cannot directly measure discrepancy between the estimated parameters and the true parameters. In practice, we use error metrics to evaluate and report the difference between the predicted performance (i.e., correctness of answers) and the actual observed performance. In the general context of machine learning, the principal objective is often to achieve the best generalized predictive performance on new data. When this predictive performance is the objective, error metrics are used to evaluate the goodness of the model; however, they do not directly evaluate the validity of a latent variable (e.g., student knowledge state) that may be modeled. Some error metrics might conclude that two models' predictive performances are the same, but fail to capture a substantive difference in how student knowledge is represented between the models. For example, Beck and Xiong [2013] showed that the models that are significantly different in their knowledge inferences may result in an only slight amount of difference in their predictive performance (e.g. in the fifth decimal place of RMSE). In addition, Pelánek [2015] identified that small differences in predictive performance between models may have significant impact on student practice and interpretable results; for example, slight differences in RMSE or log-likelihood can lead to significant changes on student over-practice and under-practice [Yudelson and Koedinger 2013], improvements in the tutoring system [Liu et al. 2014], and suggested numbers of instructional problems [Rollinson and Brunskill 2015].

Finding the ground truth model with respect to the model parameters become more and more crucial among researchers and practitioners. Recently, increasing number of studies have applied variants of the BKT model to draw pedagogical conclusions based on the model parameters. For example, Beck et al. [2008] used an extension of a BKT model to investigate the impact of receiving help on future knowledge and performance in a reading tutor. Pardos and Heffernan [2009] followed with instrumenting the model to detect differences in learning gain between different orderings of problems in a math tutor and then using a similar model extension to detect the effect of tutorial strategies such as giving hints versus breaking the problem down into steps [Pardos et al. 2011]. Lin and Chi [2016] applied a similar extension to evaluate the instructional value of having the tutoring system preemptively give help versus asking the student to reflect on what step should come next. They also found that the Bayesian Networks based BKT method performed better than alternative logistic based methods in predicting the students' outcomes.

Despite growing interest of finding the ground-truth model, there is still a lack of studies on how we can assess a model in terms of the model parameters themselves. There is a recent study on evaluating the model based on not only its predictive performance but also the plausibility of its parameters and the consistency with which it arrives at those parameters during fitting [Huang et al. 2015]. Another related work on selection of error metrics has been conducted [Pelánek 2015] which overviews the appropriateness of similar metrics to ours for a variety of models in education including models of affect and skill, such as BKT. However, none of them directly assesses the validity of model parameters or their inferences. Unlike the other works, we look at how the values of error metrics correlate with the validity of the learned parameters,

building on prior early-stage work on the topic [Pardos and Yudelson 2013; Dhanani et al. 2014].

Although error metrics are defined to measure *discrepancy in performance* (discrepancy between the predicted performance and the observed performance), we presume that an error metric can function as a proxy for *discrepancy in parameters* (discrepancy measure between the estimated model's parameter values and the true parameter values). Considering that student performance is predicted by the model parameters, we expect that a set of parameters which is closer to the ground truth will lead to a more accurate prediction in performance. This is the assumption made when evaluating BKT models: the increase in predictive accuracy translates to improvement in the validity of their inference, used to determine when a student has reached cognitive mastery. In this study, we attempt to answer which error metric serves as the best proxy of discrepancy in parameters, consequently, best representing the ground truth.

For the evaluation of the error metrics, we use synthetic data that are generated from the known ground truth (true model parameter values) so that we can establish the relationship between the discrepancy in performance and the discrepancy in parameters. The use of simulated data has proven to be useful in many studies, for example, to study the impact of prediction accuracy on students' learning experience in an adaptive educational system [Niznan et al. 2014], and to evaluate the appropriateness of various predictive models [Beheshti and Desmarais 2015]. In particular, Niznan et al. [2014] used simulated data to confirm that when a model A achieves better RMSE than a model B, A better corresponds to reality than B, in the studied setting. Similarly, we will use simulation to study not only RMSE but also other metrics in BKT setting. The work of Rosenberg-Kima and Pardos [2014] further supports our use of simulated data by showing that in BKT setting, characteristics of real and simulated data are extremely similar.

The outline of this paper is as follows. We, first, describe the BKT model in Section 2. In Section 3, we provide the overview of error metrics and discuss trends in the selection of metrics in the EDM community. Section 4 describes our data simulation procedure. In Section 5 and 6, we calculate values obtained from different error metrics at different locations over the entire parameter space, and examine the relationship between the discrepancy in performance and the discrepancy in the parameters. In Section 7, we run a gradient decent with each error metric, and the Expectation Maximization algorithm with log-likelihood to simulate actual model estimating processes. We then compare the metrics based on the distances between their estimated parameters and the true parameters. In Section 8, we investigate the validity of parameters estimated using the different error metrics on real data from ASSISTments, Cognitive Tutor, and Khan Academy. Additionally, we show that our simulated data exhibit similar characteristics to the real data. Finally, we present our conclusion in Section 9.

2. OVERVIEW OF BAYESIAN KNOWLEDGE TRACING

Knowledge tracing, popularized by Corbett and Anderson [1994], is a well-known method for modeling student knowledge. Many intelligent tutoring systems use this model to predict students' performance and determine if students have mastered a particular skill. Knowledge tracing uses four model parameters: prior, learn, guess, and slip. The prior parameter is the initial probability that students know the skill at the beginning of using the tutor. The learn parameter is the probability that students' knowledge state will transition from unlearned to learned after interacting with each question. The guess parameter is the probability that students get a correct answer when they do not know the associated skill, and the slip parameter is the probability that students make a mistake when they know the associated skill.

The system computes the probability that a student knows a given skill at a given time by updating the probability based on the observed response at each problem step using Bayesian inference. Let L_n denote the probability that a student knows the skill at time n; the prior parameter can be denoted by L_0. The probability that the student knew the skill before responding to the question at time n, L_{n-1}, is updated as follows:

$$L_{n-1}|Correct_n = \frac{L_{n-1} \times (1 - slip)}{L_{n-1} \times (1 - slip) + (1 - L_{n-1}) \times guess}$$

$$L_{n-1}|Incorrect_n = \frac{L_{n-1} \times slip}{L_{n-1} \times slip + (1 - L_{n-1}) \times (1 - guess)}$$

where $Correct_n$ and $Incorrect_n$ indicate correct and incorrect response to the question at time n respectively. Next, the system incorporates the possibility that the student learned the skill while trying to solve the problem:

$$L_n = L_{n-1}|Action_n + (1 - L_{n-1}|Action_n) \times learn$$

where $Action_n$ is either $Correct_n$ or $Incorrect_n$. An advantage of BKT over other types of machine learning models is its interpretability. The estimated parameters not only allow us to infer student knowledge, but also provide pedagogical insights to researchers.

3. METRICS

3.1. Definition of metrics

This section provides overview of commonly used metrics. We discuss metrics categorized into three families suggested by Ferri et al. [2009] and Pelánek [2015]. To define metrics, we denote that the each data point is subscripted by $i \in \{1, ..., n\}$. y_i and $\hat{y}_i$ denote the actual and predicted value of the i^{th} component of the outcome respectively.

Probabilistic Understanding of Error. The first family of metrics is described as *probabilistic* understanding of error which measures the deviation from the true probability. Pelánek [2015] has suggested that this type of metrics is natural for predictions of performance in student modeling. Most popular metrics in this family are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Log-likelihood (LL). MAE is defined by absolute differences between predicted values and observed values:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Similarly, RMSE is defined by the square root of the mean squared errors:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Likelihood of a set of parameters is the probability of the observed outcome given those parameter values. Log-likelihood is a natural log transformation of the likelihood, defined as:

$$\text{LL} = \sum_{i=1}^{n} y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i)$$

Among theses metrics, Pelánek [2015] showed that MAE is not suitable when the outcome variable is binary, like in BKT, because it is not a proper score and may lead

to misleading conclusion. RMSE is a most frequently chosen metric by researchers in general, including the Educational Data Mining community. Some researchers, however, use R^2 as a substitution of RMSE because of the difficulty of RMSE interpretation under certain circumstances. R^2 is defined as $1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$, where $\bar{y}$ is the mean, and can be interpreted as explained variability by the model.

In MAE and RMSE, the smaller value the better because they measure *errors*. On the other hand, in LL, the higher the better. LL and RMSE, in fact, share similarities in that they both take sum of errors into account, and they are considered to be equivalent in linear models; under the assumption of normally distributed errors, minimizing the residual sum-of-squares (least squares) is equivalent to maximizing likelihood [Hastie et al. 2009]. LL is also used for other model evaluation metrics such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Qualitative Understanding of Errors. The second family is defined as *qualitative* understanding of error. This family consists of metrics based on a threshold such as metrics calculated by a confusion matrix, shown in general form below.

		Observed	
		Correct	Incorrect
Predicted	Correct	True Positive (TP)	False Positive (FP)
	Incorrect	False Negative (FN)	True Negative (TN)

Precision, recall, accuracy, F-measure, and Kappa statistics are defined as:

$$\begin{aligned} precision &= \frac{TP}{TP+FP} \\ recall &= \frac{TP}{TP+FN} \\ accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ F-measure &= \frac{2TP}{2TP+FP+FN} \\ Kappa &= \frac{accuracy - R}{1-R} \end{aligned} \tag{1}$$

$$\text{where } R = (TP+FN)(TP+FP) + (TN+FP)(TN+FN)n^2$$

These metrics depend on what we choose as the classification threshold. For example, in BKT with a threshold of 0.5, the performance predicted as 0.49 and 0.51 are classified into different classes (incorrect and correct respectively), while prediction 0.51 and 0.99 are classified into the same class. Pelánek [2015] discussed that such characteristic is not desirable for student modeling.

Assessing Ranking of Examples. The metrics of the last family measure how well the model *ranks* the examples. This family includes Area under the ROC curve (AUC), a commonly used metric. AUC is defined as the area under the Receiver Operating Characteristic (ROC) curve, which plots False Positive Rate (FPR) vs. True Positive Rate (TPR) for all possible threshold values. A threshold value is a decimal between 0 and 1, where a prediction above that threshold is considered a prediction in favor of the positive class. FPR and TRP are defined in Equation (2). AUC can also be described as the probability that the prediction of a randomly chosen correct response will be greater than the prediction of a randomly chosen incorrect response. An AUC value of 0.5 represents predicting no better than random chance, while 1 represents the perfect model (TPR of 1 and FPR of 0).

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + FN} \quad (2)$$

Several studies have raised some issues about using AUC. Cortes and Mohri [2004] analyzed the relationship between AUC and error rates used in an objective function optimization. Their results showed that while an average AUC value increases as classification accuracy increases, the standard deviation of AUC values is high for uneven distribution and higher error rates. As a result, the best AUC value may not lead to the minimum error rate. In addition, Pelánek [2015] highlighted concern swith using AUC in student modeling. AUC considers only relative ranking between predictions, so it cannot capture the absolute values of predictions. For example, if all predictions are divided by two, the AUC value remains constant. In skill models such as BKT, where well-calibrated absolute values of predictions are needed, AUC may not be appropriate.

3.2. Survey of metrics used for BKT

The BKT model of cognitive mastery is of broad relevance to the field of learning analytics and has been most studied in the overlapping area of Educational Data Mining (EDM). To investigate trends in selection of error metrics for evaluating BKT, we surveyed the proceedings of the last seven years of the EDM conference.

Table I displays the types of error metrics and the number of times they have been used in the EDM conference proceedings from 2010 to 2016. Accuracy, RMSE, AUC, F-measure, recall, and precision are the most frequently used metrics in descending order with AUC, F-measure, precision and recall trending in recent years. We then extracted only papers that evaluated BKT models. Figure 1 presents the number of times each error metric has been used in BKT-related papers. Similarly, RMSE, AUC, and accuracy have been frequently used. In particular, RMSE and AUC are the two most popular metrics.

This paper aims to compare the error metrics in terms of their performance in identifying ground truth. For the comparison, we include the following metrics: LL, RMSE, AUC, accuracy, precision, and recall. Pelánek [2015] has suggested that LL and RMSE are appropriate for the evaluation of skill models. Although there exists a dispute on AUC, it appears as one of the most popular metrics in BKT-related papers based on our survey. In addition, although it has been discussed that metrics based on qualitative understanding of errors are not desirable for skill models, accuracy is quite frequently used in BKT-related papers, and other metrics such as recall and precision are also commonly used in the overall studies based on our survey; particularly, recall and accuracy have been suggested by Pardos and Yudelson [2013] as the metrics that best correspond to accuracy in skill mastery estimation.

Year	RMSE	LL	AUC	Accuracy	Recall	Precision	F-measure	R^2	AIC	BIC	MAE	Kappa	Others
2016	12	3	12	10	12	16	19	10	6	5	1	12	6
2015	22	9	18	20	8	8	14	8	9	5	2	12	10
2014	9	2	9	13	9	8	4	7	4	5	3	4	0
2013	12	1	11	13	4	4	5	4	2	5	5	9	0
2012	5	1	3	10	4	0	0	0	2	3	2	1	5
2011	8	1	3	10	4	4	1	3	4	5	1	0	1
2010	2	2	3	7	2	1	1	5	0	2	2	3	0
Total	70	19	59	83	43	41	44	37	27	30	16	41	22

Table I: Number of times different error metrics appeared in the EDM conference proceedings from 2010 to 2016. R^2 is an abbreviation for R^2 / pseudo R^2.

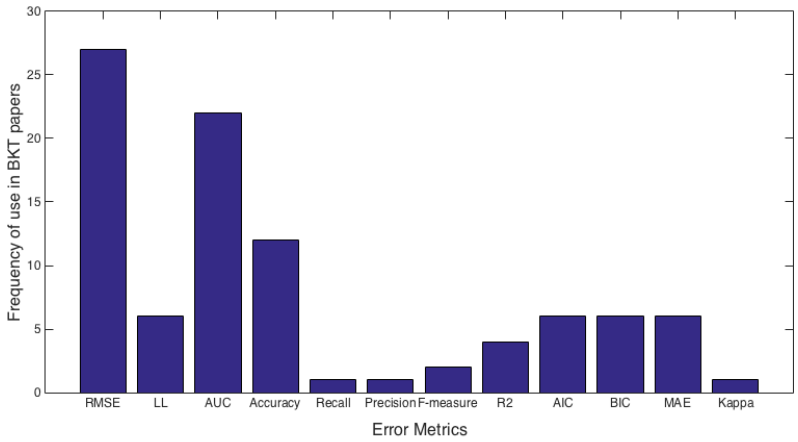

Fig. 1: Error metrics used in the EDM 2010–2016 papers that concentrate on BKT

4. DATASETS

The accuracy of a model's inference depends on its parameters. In real-world datasets, this inference of knowledge can not directly be validated, nor can the accuracy of its parameters. In order to determine which error metric is the best indicator of valid parameters, and thus inferences, we use simulated data, where the parameters generating the data are known. We will refer to these generating parameters as the ground truth parameters. We generated simulated student responses based on predefined ground truth parameters in a similar fashion to the data generated in [Pardos and Heffernan 2010]. We used the standard BKT model with four parameters: prior, learn, guess, and slip. The BKT model was constructed using functions from MATLAB's Bayes Net Toolbox [Murphy 2001]. Each dataset contains N students answering Q questions. Each data point indicates whether the student's answer was correct or incorrect.

Parameter	≤ 0.5	> 0.5
prior	16	10
learn	19	7
guess	17	9
slip	15	11

Table II: Distribution of datasets' parameter values. Column 2 and 3 contain numbers of datasets whose specific parameter value is less than and greater than 0.5 respectively.

We generated 26 datasets with diverse parameter values.[1] Fifteen datasets contain data for 3,000 students, and 11 datasets contain data for 30,000 students. Nineteen of the datasets had responses for five questions per student, and seven of the datasets had responses for ten questions per student. Each dataset, and respective set of generating parameters, can be considered as a separate skill or Knowledge Component (KC) that our simulated students interacted with. Table II shows the distribution of prior, learn, guess, and slip parameter values in our datasets. The actual parameters, numbers of

[1]Datasets, code for their generation, and result files are available at https://github.com/CAHLR/publications/tree/master/JEDM_metrics

students, and numbers of questions used in generating the datasets can be found in Appendix A. Most of our datasets have low guess values ($guess \leq 0.5$). However, nine datasets with high guess values were generated as well to account for some problem sets with high guess values such as the exercises in the Reading Tutor [Beck and Chang 2007].

5. CORRELATIONS TO GROUND TRUTH

In this section, we evaluate how good each of the selected error metrics are at indicating the closeness (or distance) of a model's parameters to the ground truth values, i.e., the discrepancy between the estimated parameter values and the true parameter values. An instructional designer may use a BKT model to evaluate the effectiveness of an instruction by inspecting the $learn$ parameter value given a particular stimulus or intervention; this is, a human perception matters. We assume the instructional designer naturally interprets that the distance between different values of the parameter is at face and linear. Therefore, we use Euclidean distance to measure the closeness of parameter values.

5.1. Methodology

For each dataset, we calculated (i) the error metric value measuring predictive performance and (ii) the distance to the ground truth of each parameter point in the four dimensional parameter space (dimensions were prior/learn/guess/slip) with a interval of 0.1. Each point P is defined as follows.

$$P = (P_1, P_2, P_3, P_4) = (prior, learn, guess, slip)$$

On each point P, we calculated students' predicted responses (probability that students will answer questions correctly) conditioned on previous responses. We then used these predicted responses along with the actual simulated responses to calculate LL, RMSE, AUC, precision, recall, and accuracy for all points.

To determine which error metric was best at indicating ground truth, we looked at the correlations between values calculated by the error metrics and the Euclidean distances from the corresponding points to the ground truth. The distance from a point P to the ground truth R is define as follows.

$$d(P, R) = \sqrt{\sum_{i=1}^{4} (P_i - R_i)^2}$$

For each error metric, we plotted the error metric values against distances. Note that we used -RMSE instead of RMSE to standardize our convention across different error metrics; as a result, with this change, higher error metric values indicate smaller error—closer to the ground truth—for all error metrics. In addition to visualizing the results, we calculated correlation coefficients between the six error metric values and distances. For a particular set of parameters, we call a better indicator of the ground truth if values calculated by an error metric have stronger positive correlations with the distances to the ground truth.

After evaluating correlations over the entire parameter space, called *entire space*, we conducted a focused analysis on the area close to the ground truth. Assuming the

ground truth is R, we define *nearby space* to be the area that covers all points p:

$$p \in \{(p_1, p_2, p_3, p_4) \quad | \quad R_1 - 0.1 \leq p_1 \leq R_1 + 0.1, \\ R_2 - 0.1 \leq p_2 \leq R_2 + 0.1, \\ R_3 - 0.1 \leq p_3 \leq R_3 + 0.1, \\ R_4 - 0.1 \leq p_4 \leq R_4 + 0.1\}$$

For each error metric, a correlation coefficient is calculated considering points in the nearby space with a 0.02 interval.

We hypothesized that higher correlations with the distance in the entire space do not always imply higher correlations with the distance in the nearby space. We also considered the situation in which one metric is the best at directing to the right region but another metric is better at greater precision. This is important as we are looking for an error metric that can guide the parameter estimation not only to an area close to the ground truth but also to a specific location that is as close to the ground truth as possible.

5.2. Results

Figure 2 displays the scatter plots of values calculated by the error metric vs. distances from the ground truth of dataset 2. The dataset was generated with prior = 0.2, learn = 0.444, guess = 0.321, and slip = 0.123. As shown in the figure, LL and -RMSE appear to have stronger correlations with the distance from the ground truth. This pattern, in fact, is common in all datasets. This indicates LL and RMSE as good measures of distances from the ground truth. In the case of AUC, precision, recall and accuracy, we are not able to discern any observable relationship between distances from the ground truth and error metric values. In certain cases, they exhibit a similar pattern to that of RMSE and LL, but the appearance of this pattern is inconsistent.

The correlation coefficients we calculated between the error metric values and the distances from the ground truth further support the findings from our visual analysis. Table IIIa summarizes the correlation coefficient values over the entire space. The table includes mean, maximum and minimum values of the correlations from the 26 datasets. An entry in 'best' and 'worst' row shows the number of the datasets for which a particular error metric has the highest and lowest correlation value, respectively, among all metrics. The individual correlation coefficients of all datasets can be found in Appendix B.

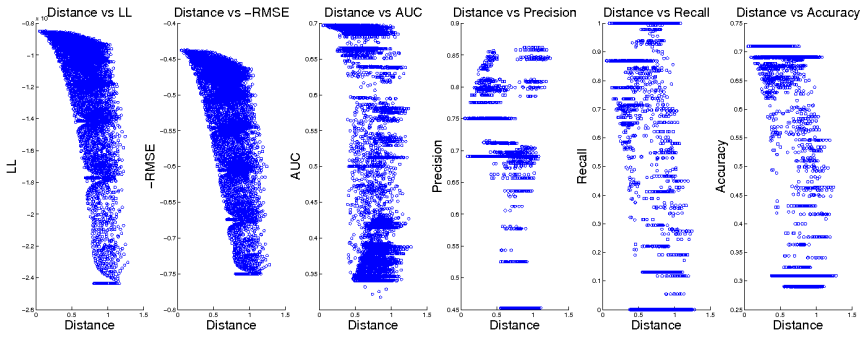

Fig. 2: Values calculated by different error metrics vs. distances to the ground truth. Each dot in the plots corresponds to an individual point in the four dimensional parameter space.

	LL	-RMSE	AUC	precision	recall	accuracy
mean	0.4705	0.4926	0.4090	N/A	0.0623	0.3712
max	0.7133	0.7343	0.5930	N/A	0.5472	0.5931
min	0.2418	0.2601	0.0879	N/A	-0.5875	0.1031
best	0	15	11	0	0	0
worst	0	0	3	0	23	0

(a) Entire parameter space with a 0.1 interval

	LL	-RMSE	AUC	precision	recall	accuracy
mean	0.5479	0.5811	0.2702	-0.0035	0.0450	0.2787
max	0.7337	0.8758	0.5152	0.3194	0.3073	0.4749
min	0.4247	0.4965	-0.0080	-0.2961	-0.3089	0.0144
best	6	20	0	0	0	0
worst	0	0	5	9	12	0

(b) Nearby space with a 0.02 interval

Table III: Summary of correlation coefficients on all points over entire space (top) and the nearby space (bottom). An entry in 'best' and 'worst' row shows the number of the datasets for which a particular error metric has the highest and lowest correlation value respectively.

According to the entire space table, RMSE has the best performance. RMSE has the highest mean, min, and max correlation values. Additionally, it has the highest correlation value in 15 out of 26 datasets and does not has the smallest correlation value in any dataset. Although LL does not have the highest correlation value in any dataset, its performance is the second best according to mean, max, and min. AUC, on the other hand, has the highest correlation value in 11 datasets, but its min and max are much lower than those of RMSE and LL, and it performs worst in three datasets. This statistic reveals that the performance of AUC is inconsistent across datasets. Thus, despite being the best in 11 datasets, AUC may not be a good metric. The correlations of precision values cannot be calculated because some precision values are undefined. According to its definition in Equation (1), precision considers only the student responses that are predicted as correct; as a result, when the denominator is zero, the precision value can be undefined.[2] This demonstrates the limitation of precision. Overall, recall also performs poorly. It has the smallest mean, max, and min correlation values, has the smallest correlation values in 23 datasets, and has negative correlation values in eight datasets. Lastly, accuracy shows comparable performance with AUC in terms of mean, max, and min. It clearly performs worse than LL and RMSE but better than recall.

Table IIIb shows the correlation coefficient values over the nearby space. The individual correlation coefficients of all datasets on the nearby space can be found in Appendix B. Superiority of RMSE is even more noticeable in the nearby space analysis; RMSE has the largest values in mean, max, and min, and is the best in 20 datasets. LL is the second best metric as it is in the entire space analysis. AUC, in the nearby space, does not have the highest correlation value in any dataset and, furthermore, produces negative correlations for one dataset. This result further suggests that AUC is not a reliable metric. The mean correlation values of precision and recall are almost equal to zero. With this result, we can conclude that precision and recall are poor metrics for identifying the ground truth. Accuracy performs slightly better than AUC, as it does

[2]With some parameter values, the BKT model predicts that all students always answer incorrectly; for each response of each student, a probability that such response is correct is always less than 0.5. According to Equation (1), TP and FP are both zero. As a result, precision = 0/0 = undefined.

Comparison	Δ of correlations	t	p-value
RMSE > LL	0.0221	7.1975	< 0.001
RMSE > AUC	0.0835	2.3487	0.0135
RMSE > Accuracy	0.1214	12.023	< 0.001
RMSE > Recall	0.4302	6.0144	< 0.001
LL > AUC	0.0614	1.7233	0.0486
LL > Accuracy	0.0993	8.3329	< 0.001
LL > Recall	0.4082	5.6814	< 0.001
AUC > Accuracy	0.0379	0.9774	0.1689
AUC > Recall	0.3467	5.9946	< 0.001
Accuracy > Recall	0.3089	4.1642	< 0.001

(a) Entire space

Comparison	Δ of correlations	t	p-value
RMSE > LL	0.0332	3.972	< 0.001
RMSE > AUC	0.3109	11.347	< 0.001
LL > AUC	0.2778	10.71	< 0.001

(b) Nearby space

Table IV: T-test statistics for comparing $A > B$. A is defined as correlation between values calculated from an error metric A and distances to the ground truth.

not have a negative correlation value in any dataset, but it is still much worse than LL and RMSE.

In order to determine if the differences in correlations to the ground truth between different metrics are statistically significant, we evaluated the one-tailed paired t-test on correlation values of all 26 datasets comparing all pairs of metrics excluding precision because of its missing correlation coefficients. Recall and accuracy were also excluded from the nearby space t-test because of their missing correlation coefficients on some datasets. The result is shown in Table IV. In summary, the differences in all pairs of the metrics, except for AUC-accuracy pair, are statistically significant at 5% significance level. Although the mean RMSE correlation is only slightly greater than that of LL, the difference is significant.

Hence, the results in both the entire space and nearby space analyses suggest that RMSE and LL are good indicators of the distance to the ground truth with RMSE being best at any proximity to the ground truth.

6. GRADIENT VISUALIZATION

In this section, we examine further why RMSE and LL appear to be better indicators of ground truth than the other metrics by visualizing the gradient of the values calculated from the error metrics across the domain of the model parameters.

6.1. Methodology

We visualized error metric values of all points over the two dimensional guess/slip space with an interval of 0.02. We fixed prior and learn parameter values to the ground truth values. Using the guess and slip parameters as the axes, we visualized the values calculated from the error metrics by colors ranging from dark red to dark blue corresponding to the values ranging from low to high. We call this plot a *heat map*.

6.2. Results

According to the visual characteristics of the heat maps, we could roughly categorize the datasets into four groups shown in Table V. Figure 3 shows the heat maps of all error metrics using a representative dataset from each of the four groups. The white dot in each graph indicates the location of the ground truth (generating parameter val-

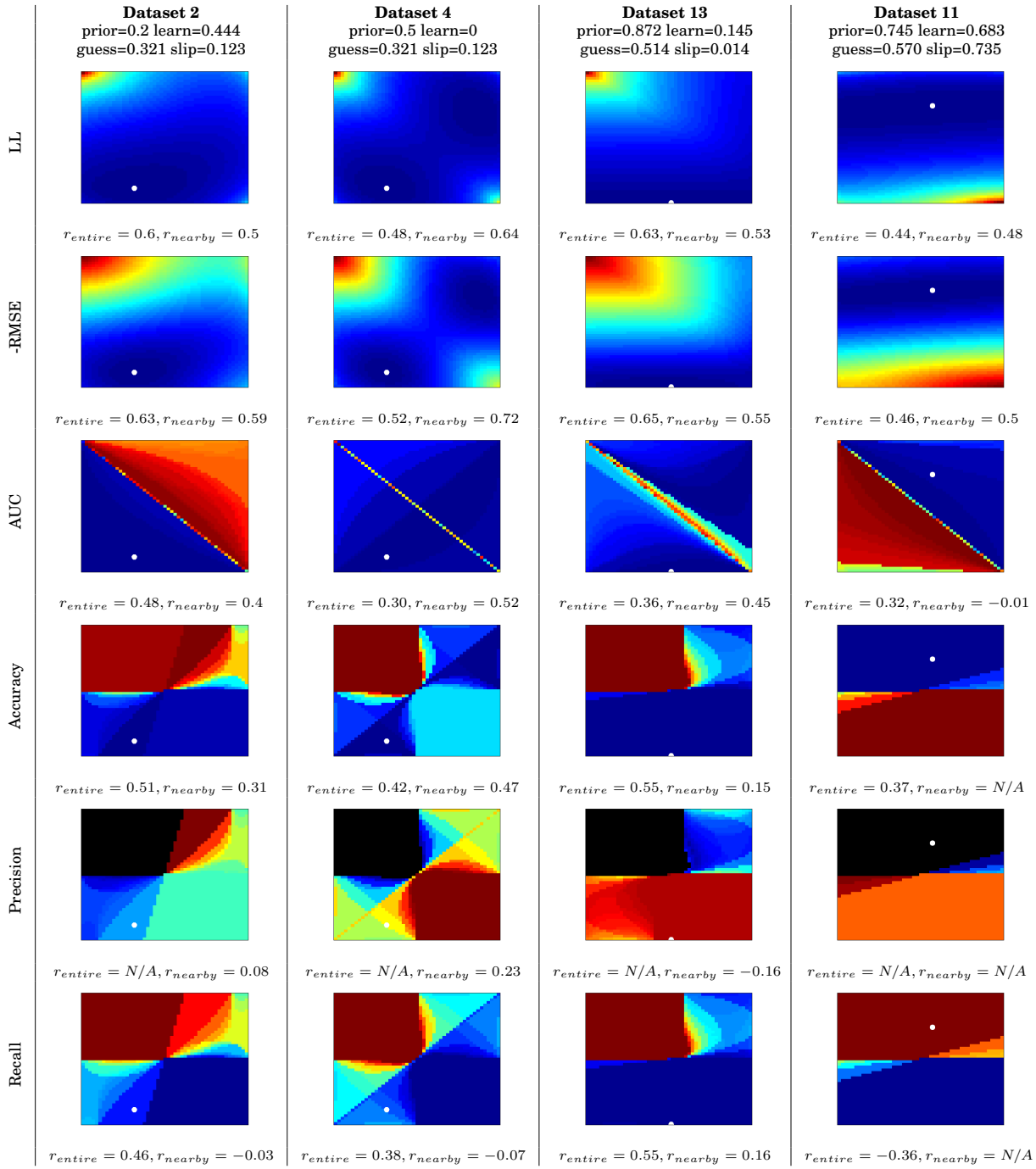

Fig. 3: Guess-slip heat maps of some simulated datasets when fixing prior and learn to true values. X-axis is guess, an Y-axis is slip. The colors range from red to blue with blue representing high prediction accuracy or low error depending on the metric (blue is good, red is poor). The white dot represents the ground truth. r_{entire} is a correlation coefficient in the entire space, while r_{nearby} is a correlation coefficient in the nearby space for that dataset from Tables 3 and 4.

ues). Black represents undefined values resulting from dividing by zero. These regions only exist in precision heat maps.

In all LL, -RMSE, AUC, and accuracy heat maps, the actual ground truth parameters are located in the expected regions, the dark blue regions, the areas with high error metric values. However, in precision and recall heat maps, the dark blue regions often do not contain the ground truth. In precision and recall heat maps of Group A and B, the actual parameters are located in regions with colors ranging from light blue to green. Note that these two groups contain more than half of the datasets. Group B differs from Group A only that Group B has two dark blue regions instead of one. In precision heat maps of Group C, the actual parameters are located in the regions with colors ranging from orange to red, opposite from the expected colors. In Group D, the actual parameters are located in N/A (black) regions of precision heat maps, and in dark red regions of recall heat maps. Thus, these heat maps clearly suggest that precision and recall are, again, poor error metrics for identifying the ground truth in BKT.

Although the dark blue regions of AUC, and accuracy heat maps contain the ground truth parameters, the AUC and accuracy heat maps do not have smooth gradients from low to high values compared to the LL and -RMSE heat maps. In particular, accuracy heat maps show more severe pattern, segmenting the parameter space into several sections without smooth gradients. The same pattern is also observed in precision and recall heat maps. This pattern of the heat maps can be connected to the threshold issue addressed in section 3: accuracy, recall, and precisions do not distinguish predictions classified into the same class. The lack of smooth gradients in the heat maps of AUC, accuracy, precision, and recall agrees with their low correlations with the distances to the ground truth, discovered in Section 5. This result further suggests that these metrics might not be good error metrics for indicating the closeness to the ground truth.

These heat maps also explain why RMSE has higher correlation with the distances to the ground truth than does LL, particularly in the area around the ground truth. In each dataset, if we follow the gradient from the dark red region to the dark blue region, we can see that RMSE has a smoother, more gradual transition than LL and that the region surrounding the ground truth has more level of gradation for RMSE than for LL. However, we still cannot conclude that RMSE is more accurate metric than LL at identifying the ground truth from these results, but rather that it might be a better metric for conducting a guided search of the parameter space.

7. ESTIMATING MODEL PARAMETERS

Thus far, the correlational analysis and visualization of the gradient space have shown that LL and RMSE are good indicators of the closeness of a model's parameters to the ground truth. This section compares the performance of the error metrics in identifying the ground truth when used for guiding the parameter estimation process. We exclude precision and recall from our analysis in this section as results from the previous sections demonstrate those metrics have very low correspondence to ground truth.

Group	Datasets	Characteristics
A	1 2 3 6 7 9 12 17 19 20 21 24 25	Ground truth lies in high accuracy regions for all metrics.
B	4 5	Heat maps are symmetric along $guess + slip = 1$ line.
C	8 13 15 18 22	Ground truth lies in a low precision region.
D	11 10 14 16 23 26	Ground truth lies in a low recall and undefined precision region.

Table V: Groups of datasets categorized according to the visual similarities of their heat maps

7.1. Methodology

Part I. We implemented BKT model parameter estimation using Matlab's $fminsearch$ optimization function to find prior, learn, guess, and slip parameters such that each of -LL, RMSE, -AUC, and -accuracy is minimized. $Fminsearch$ employs the Nelder-Mead simplex direct search algorithm [Nelder and Mead 1965]. We selected ten random starting points to be used for the parameter estimation procedure. For each error metric, we ran the estimation process ten times starting from the ten predetermined random points, and selected the best parameters converged to among the ten runs. We set the termination tolerance on x (parameters) to 10^{-6}. We then recorded the distance between the best converged parameters, according to the respective error metric, and the ground truth values for each error metric on each dataset.

Part II. Although our overall focus is not on determining which fitting algorithm is best at finding accurate model parameters, we are still interested in the effect of fitting algorithms on the accuracy of the solution, and whether the choice of error metric is the more significant factor. Therefore, we also included the standard modeling fitting approach [Chang et al. 2006; Pardos and Heffernan 2010] using EM with LL to compare to the Nelder-Mead search with LL. EM from the xBKT library[3] and the Nelder-Mead search ($fminsearch$) are compared using the same methodology as in Part I.

7.2. Results

Part I. Table VIa summarizes the distances between the ground truth and the parameters estimated using the different error metrics. LL estimates the closet parameters to the ground truth in half of the datasets, while RMSE does in the other half. In contrast, AUC and accuracy never estimate the closet parameters to the ground truth. The average distances to the ground truth when using LL and RMSE are both around 0.12, while the average distances to the ground truth when using AUC and accuracy are around 0.53. This result further supports our speculation that LL and RMSE are better at indicating the closeness to the ground truth. Although Table IIIa and IIIb show that -RMSE values have slightly higher correlation with the distances to the ground truth than do LL values, the result here shows that they are equally accurate at identifying the ground truth. The individual results of all datasets can be found in Appendix C.

Part II. Table VIb shows the distances between the ground truth and the parameters estimated using LL as an error metric with $fminsearch$ and EM. $Fminsearch$ and EM estimate the closest parameters to the ground truth in 19 and 14 datasets, re-

[3]https://github.com/CAHLR/xBKT

	Distances to ground truth			
dataset	LL	-RMSE	AUC	Accuracy
mean	0.1211	0.1192	0.5269	0.5370
min	0.0020	0.0018	0.1470	0.1382
max	0.7925	0.7932	0.9611	1.0002
best	13	13	0	0

(a) Comparing error metrics using $fminsearch$

	Distances to ground truth	
dataset	fminsearch	EM
mean	0.1211	0.1080
min	0.0020	0.0020
max	0.7925	0.7925
best	19	14

(b) Comparing search algorithms using LL as an error metric

Table VI: Distances between the estimated parameters and the ground truth on 26 datasets. The 'best' row summarizes the numbers of datasets the corresponding metrics predict the closet parameters to the ground truth.

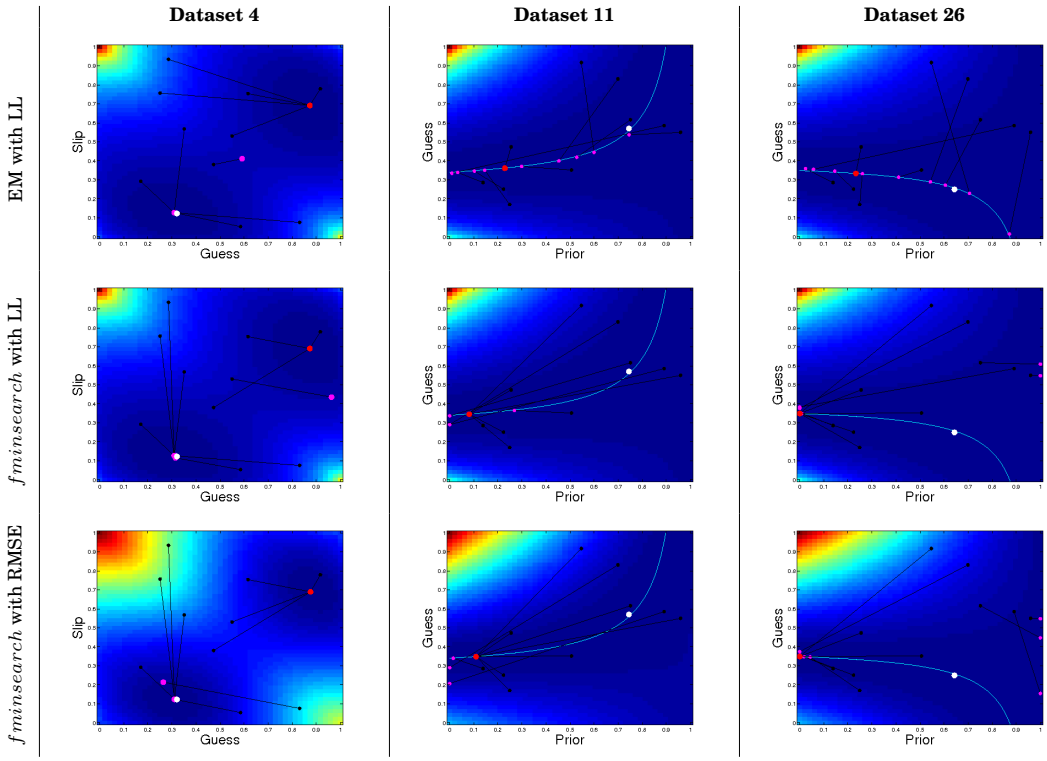

Fig. 4: LL and -RMSE heatmaps. Colors ranging from red to blue represent low to high values. Black dots represent the initial parameter values of the 10 runs. Each line connects an initial point to the corresponding solution, marked in pink, for each run. The red dot represents the final solution, the best of all runs. The white dot represents the ground truth. The thin light blue line is the $guess$ and $prior$ values derived from van de Sande's functional form.

spectively. $Fminsearch$ and EM tied in 7 dataset. However, the average distance to the ground truth when using $fminsearch$ is slightly larger than when using EM. Therefore, we conclude that whether using EM or $fminsearch$ does not significantly affect the accuracy of the solution.

7.3. Issues of Identifiability and Label Switching

Identifiability and label switching are the two main empirical problems of BKT, which may occur in our synthetic datasets. In this section we look deeper into the outlier datasets to explore if their poor convergence was caused by issues of identifiability or label switching that can be explained by the previous works. The outlier datasets include datasets 4, 11, and 26, in which the parameter estimation procedure did not converge to points within 0.2 distance to the ground truth using LL or RMSE. While the identifiability issue and label switching should theoretically present themselves regardless of the fitting procedure, we investigated them using the following subset of fitting procedures to observe their susceptibility in depth: (1) EM with LL, (2) $fminsearch$ with LL, and (3) $fminsearch$ with RMSE.

Label Switching Problem. In dataset 4, all three estimation procedures estimate similar parameter values. The differences between the estimated values and the true values

were less than 0.05 for prior and learn, but were more than 0.5 for guess and slip. Thus, we investigated this problem further by visualizing the heat maps of LL and -RMSE values, shown in Figure 4, when varying guess and slip (the problematic parameters), while fixing prior and learn (the non-problematic parameters) to the true values. We use the same colors ranging from dark blue to dark red as in Figure 3. Black dots represent the initial parameter values of the 10 runs. Each line connects an initial point to the corresponding solution, marked in pink. The red dot represents the best converged parameters. The white dot represents the ground truth.

It can be observed that there are two regions of good fit in dataset 4. Some starting positions resulted in convergence very close to the true parameters, while others converged to a region far from ground truth. With all three fitting procedures, the error metrics indicated that the points converged to in the far away region were slightly better than the region around the ground truth.

This divergence issue is an artifact of a known issue called label switching [Redner and Walker 1984; Celeux 1998]. Applied to BKT, the word *label* refers to the value of the knowledge node and its corresponding relationship to the observable question responses. The assumed relationship in the context of learning is that a value of true for the knowledge node, the learned state, should be associated with true values for the observable question nodes (correct answers). Label switching is when the underlying distribution of the data, or generating parameters in our case, allows for a switching of the label to result in an identical fit to the data. This occurs when the learning rate is zero. In this case, knowledge could be represented instead by a negative value, in which case guess and slip parameters would likewise swap interpretations. By avoiding parameter values where guess + slip >= 1, as done in Pardos and Heffernan [2010], we can avoid a switched label, and the parameter estimation will converge towards the ground truth. Our result of dataset 4 is an example of when fitting is vulnerable to label switching as the generating learn rate in this dataset was zero.

Prior-Guess Identifiability Problem. Unlike dataset 4, dataset 11 and 26 do not have two regions of high LL or -RMSE values, yet the parameter estimation processes still did not converge to locations close to the ground truth. All three fitting procedures again found similar parameter values. The differences between the estimated slip and the true slip was less than 0.002 in this dataset, unlike dataset 4. In contrast, the differences between the estimated prior and the true prior was greater than 0.4. For learn and guess parameters, the differences were on average 0.08 and 0.16 respectively. Therefore, instead of generating the heat maps by varying guess and slip, we varied the two most problematic parameters in this dataset, prior and guess. Columns 2 and 3 of Figure 4 shows the heat maps of dataset 11 and 26. In these two datasets, 10 runs of EM with LL converged to different locations, of which a few were close to the true parameters. For $fminsearch$ with LL and RMSE, most runs converged to the same small area that was far away from the ground truth.

We believe that this divergence issue is an artifact of the identifiability problem. Beck and Chang [2007; van de Sande [2013] show that the identifiability problem exists in the BKT function form (in which prediction is made without considering observable student's responses). The authors show that when learn and slip parameters are assigned to specific values, multiple combinations of prior and guess fit the student performance data equally well. This is a troubling potential when the goal of model fitting is to draw pedagogical conclusions from the learned parameters. Furthermore, van de Sande [2013] analytically proves that there are not merely a few sets of parameters that explain the data equally well but instead a continuous line of best fit in the prior vs. guess space. This analytical proof applies to the functional form of BKT, which van de Sande refers to as the Hidden Markov Model (HMM) form. The same

phenomenon cannot be proved analytically for the algorithm form (in which prediction is made based on observable student's responses). Although previous works cannot show that the identifiability issue may exist in the BKT algorithm form, our result here—which was derived from the BKT algorithm form—suggests that to some degree the same effect can exist in the BKT algorithm form as well.

Before we can conclude that the identifiablity problem indeed exist in the BKT algorithm form, we first investigated how well van de Sande's explanation fits our results. According to van de Sande, the probability that a student gets opportunity n correct is:

$$Correct_n = 1 - slip - Ae^{-\beta n}$$

where

$$A = (1 - slip - guess)(1 - prior)$$
$$\beta = -log(1 - learn)$$

There are different combinations of $guess$ and $prior$ that give the same value for A; these different combinations will result in models that give the same predictions. If $n, Correct_n, learn,$ and $slip$ are given, then $guess$ and $prior$ values can be derived. We hypothesize that if van de Sande's explanation for the identifiablity problem is applicable to the BKT algorithm form, the estimated $guess$ and $prior$ from the fitting algorithms should be close to the $guess$ and $prior$ values derived from the HMM functional form.

To test this hypothesis, in each of datasets 11 and 26, we derived $guess$ and $prior$ values from the functional form of $Correct_0$, the probability that students answer the first question correctly ($n = 0$). We calculated $Correct_0$ from the simulated student performance data, and $Correct_0$ of datasets 11 and 26 were 0.340 and 0.349 respectively. Setting $learn$ and $slip$ parameters to the ground truth values, we derived $guess$ and $prior$ values from the HMM functional form. The derived $guess$ and $prior$ values are shown as thin light blue lines on the heat maps of datasets 11 and 26 in Figure 4. The true parameters and almost all estimated parameters perfectly lie on top of the derived values. Although according to our result, points on van de Sande's best fit line do not have identical LL and RMSE values as they would be if using the functional form, many points along the line do have identical LL and RMSE values. With this evidence, we conclude that van de Sande's explanation for the identifiablity problem in the HMM form can also explain the identifiablity problem in the BKT algorithm form.

In conclusion, neither EM with LL, $fminsearch$ with LL, or $fminsearch$ with RMSE perform any better at avoiding the false regions when an identifiability issue or label switching occurs. While we do not know under which condition these issues may occur, the generating values for prior and learn appear to have an influence. In the case of all three datasets, the generating prior is greater than 0.5. In two cases, the learn rate is very high, greater than 0.65, and in the third case it is zero. With a high prior or a zero learn rate, there are few training examples to observe possible guesses and slips. A more typical dataset, or skill, is one in which most students begin interacting without knowing the skill and learn gradually, as opposed to already knowing the skill at the beginning or not learning or learning immediately. These more normal types of learning scenarios may be more resilient to the identifiability issue and label switching.

To mitigate the issue, Dirichlet priors have been suggested [Beck and Chang 2007] to bias the parameters towards a particular region; this is similar to biasing the search

initial positions. This approach requires an assumption about which regions are plausible and which are not. Simply bounding the parameter search can accomplish a similar outcome if the same assumption is applied. Baker et al. [2008] suggested using contextual features to determine if an answer is more likely to be a slip (or guess). While this method can improve interpretation of individual student responses, it is based on regressing to an already trained BKT model, whose fitting procedure is prone to the same issues described in this section. BKT model extensions allowing for a different guess and slip per question have been robust to the label switching issue; however, the questions themselves must be given in a random order per student in order to be effective. While increasing the number of parameters is not an intuitive approach to solving theses issues, modeling the control for ordering effects might add constraints which ameliorate identifiability and label switching issues. Lastly, controlling for prior knowledge per student has shown promise in improving convergence properties [Pardos and Heffernan 2010], perhaps thanks to allowing for a greater variety of observations of guess and slip when using individual priors instead of a single point estimate.

8. REAL DATASETS

Up until this point, we have been working with simulated data. In this section, we would like to establish the relationship between simulated and real data. Ultimately, if simulated data and real data are similar, our findings in the previous sections should apply to real data as well. We also investigate the validity of the estimated parameters on the real data in this section.

8.1. Datasets

We obtained the datasets from three online educational platforms: ASSISTments, Cognitive Tutor, and Khan Academy. On each platform, we obtained three datasets for three skills: fraction, circle area, and exponent calculations. We only considered the first five responses of each student, discarding students with less than five responses. Table VII displays the numbers of students with at least five responses in the datasets.

Platform	Fraction	Circle Area	Exponent
ASSISTments	393	40	407
Cognitive Tutor	996	99	1221
Khan Academy	130	96	152

Table VII: Numbers of students in real datasets from ASSISTments, Cognitive Tutor, and Khan Acamedy on fraction, circle area, and exponent calculation skills.

8.2. Similarity Between Simulated and Real Data

Gradient Visualization. We hypothesized that the heat maps of the real data follow the similar patterns as those of the simulated data. Therefore, we visualized the guess-slip heat maps of the real datasets (with the same process as in Section 6.1) when fixing prior and learn to the best values according to RMSE since RMSE appears to be the most accurate metric in identifying ground truth. Figure 5 displays the heat maps of the exponent skill datasets, the biggest datasets, from the three platforms. We can observe that the heat maps of the real datasets exhibit the same patterns as those of the simulated datasets, shown in Figure 3.

Parameter Estimation. Besides visualizing heat maps, we ran parameter estimation using $fminsearch$ with the six error metrics (as in Section 7.1) on our simulated and real datasets. In each dataset, we calculated a parameter distance (Euclidean distance) between each pair of parameters estimated with two different error metrics.

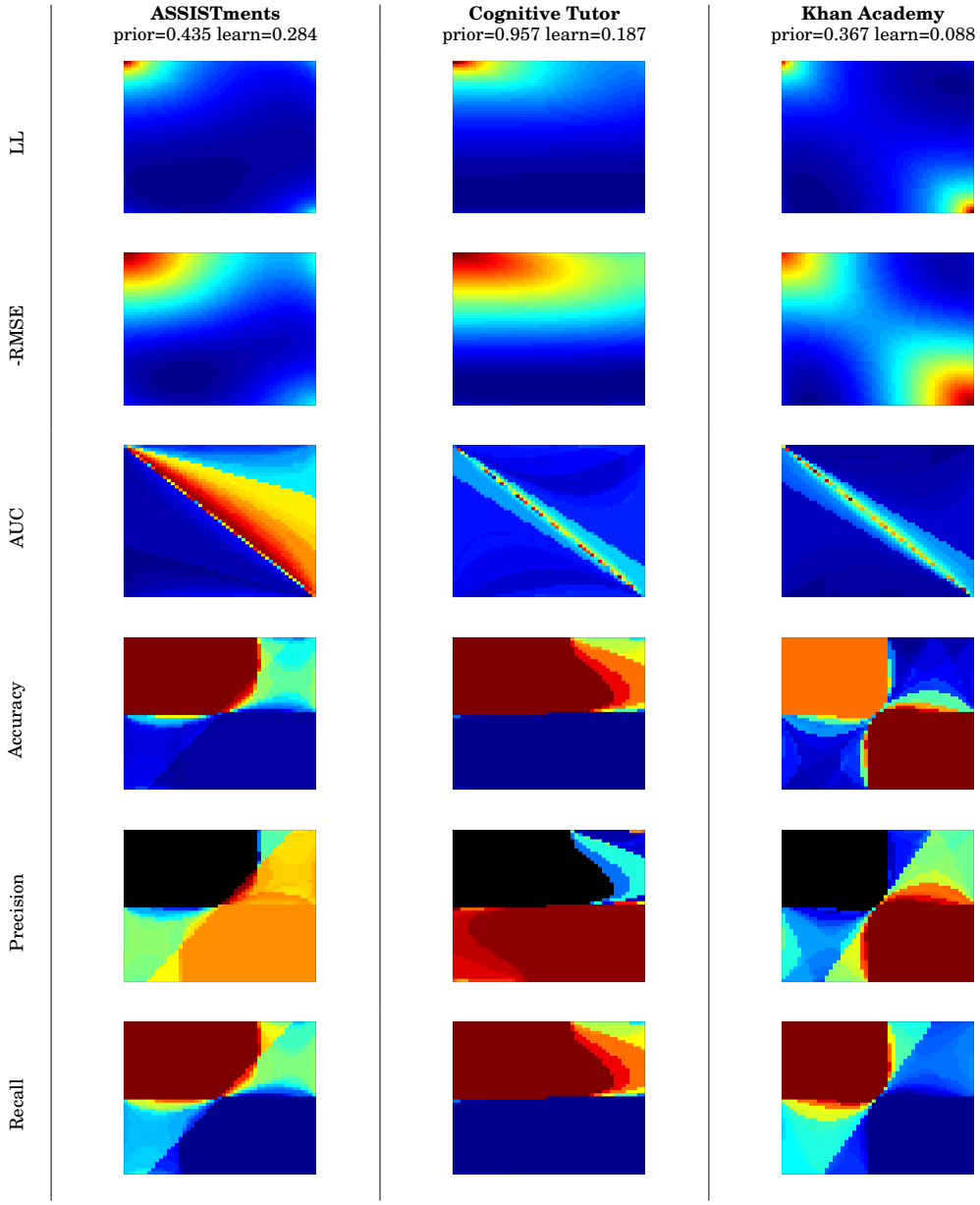

Fig. 5: Guess-slip heat maps for the skill of exponentiation from ASSISTments, Cognitive Tutor, and Khan Academy. We fixed prior and learn to best values found using RMSE. X-axis is guess and Y-axis is slip. The colors range from red to blue with blue representing high prediction accuracy or low error depending on the metric (blue is good, red is poor).

Distance	LL	RMSE	AUC	Precision	Recall	Accuary
LL	-	0.207	0.756	0.964	0.841	0.674
RMSE	-	-	0.762	0.870	0.860	0.577
AUC	-	-	-	0.738	0.870	0.789
Precision	-	-	-	-	0.840	0.598
Recall	-	-	-	-	-	0.439
Accuary	-	-	-	-	-	-

(a) Real datasets

Distance	LL	RMSE	AUC	Precision	Recall	Accuary
LL	-	0.004	0.492	0.676	0.654	0.480
RMSE	-	-	0.492	0.675	0.654	0.481
AUC	-	-	-	0.705	0.785	0.564
Precision	-	-	-	-	0.793	0.557
Recall	-	-	-	-	-	0.491
Accuary	-	-	-	-	-	-

(b) Simulated datasets

Table VIII: Average parameter distances. Each entry is the average distance between the parameters estimated by the row metric and the parameters estimated by the column metric.

Metrics	Guess > 0.5	Slip > 0.5
LL	1	1
RMSE	2	2
AUC	5	6
Precision	9	9
Recall	0	0
Accuracy	3	3

Table IX: Number of datasets (out of nine datasets) whose estimated parameters contain guess > 0.5 or slip > 0.5 when estimating using the different error metrics

We then calculated the average parameter distances on all simulated datasets and on all real datasets. We then used the average parameter distances as the characteristics of datasets in our comparison between the simulated and real datsets.

Tables VIIIa and VIIIb display the average parameter distances on the real and simulated datasets respectively. The correlation of the parameter distances between real and simulated datasets is 0.8599 with p-value < 0.001. Thus, we conclude that the real and simulated datasets are significantly similar and that our findings so far are valid to not only simulated data but also real data. Notice that all average parameter distances from the real datasets are larger than from the simulated datasets. This may imply that the choice of error metric is even more important when used in the real setting.

8.3. Parameter Validity

We further examined the resulting parameters estimated with different metrics from Section 8.2. Although we do not have the ground truth for the real datasets, we know that in general guess and slip values of most skills should be less than 0.5. Thus, we can use this criteria for checking parameter validity. Table IX reports numbers of datasets in which parameters estimated by different error metrics are degenerated (guess > 0.5 or slip > 0.5).

At the first glance, recall seemed to perform best at estimating the real datasets. However, when we examined the parameters esitmated using recall, we found that the estimated parameters are exactly the same (prior = 0.2551, learn = 0.8407, guess

= 0.4733, and slip = 0.3804) for all datasets as well as for the simulated datasets. This is in fact not totally surprising. According to Equation (1), $recall = TP/(TP + FN)$. Therefore, there can exist parameters that predict all students' answers to be correct; $FP = 0$, and $TP/(TP + FN) = 1$. The parameters estimated by recall have this characteristic because learn is very high, and slip is low; thus, the probability that a student gives a correct answer is always more than 0.5 for any attempt, so recall considers all the responses to be correct (positive). In fact, there are many more sets of parameters that have this characteristic. When we evaluated recall at all points in the entire parameter space with interval 0.1, there are between 2,400 and 3,200 points of parameters out of 10,000 points that have recall = 1. We can also observe this characteristic in the recall heat maps in Figure 3 and 5 as the highest value regions (dark blue region) are widely span across the parameter space.

Apart from recall, according to the result, LL was the best, estimating guess > 0.5 in only one dataset and slip > 0.5 in only one dataset. RMSE estimated almost identical parameters as did LL, except for the circle area calculation skill from Khan Academy. In that particular dataset, the estimated parameters using LL (P_{LL}) had prior = 0.2009, learn = 0.0491, guess = 0.05629, and slip = 0.1276, while the estimated parameters using RMSE (P_{RMSE}) had prior = 0.7183, learn = 0.186, guess = 0.9929, and slip = 1. However, when we evaluated LL and RMSE of P_{LL} and P_{RMSE}, we actually found that P_{LL} had better LL and RMSE values than did P_{RMSE}; unfortunately, $fminsearch$ failed to discover P_{LL} when using RMSE. Therefore, it was not because RMSE was inferior to LL, but $fminsearch$ might not be the right algorithm for estimating parameters. Accuracy appeared to be better than AUC according to the result, and precision still appeared to be the worst metric.

Once again, the result from analyzing the real data suggests that LL and RMSE are superior indicators of parameter validity.

9. CONCLUSION

According to all experiments we performed, the results consistently showed that when evaluating Bayesian Knowledge Tracing models, RMSE and log-likelihood were superior indicators of parameter validity (best representing ground truth). The other metrics, on the other hand, were poor indicators and sometimes anti-correlated with the closeness to ground truth. Our survey of EDM literature from 2010 to 2016 revealed that RMSE was the most popular metric to evaluate BKT models. Our results validate this standard choice of evaluation metric. However, the close second and third most frequently-used error metrics to evaluate BKT models—AUC and accuracy, respectively—were shown to be particularly poor indicators of the closeness of a model's parameters to the ground truth in our studies. Metrics other than RMSE and LL should therefore be steered clear of when evaluating BKT models except in the cases where the model would be used purely for its prediction of observables and not interpretation of its parameters or inferences on knowledge. For parameter fitting, both RMSE and LL were again superior and provided a smoother gradient allowing for better convergence than the other metrics. These results validate the existing standard practices of using EM with LL or grid search with RMSE (or other residual based error metrics such as SSE and MSE).

There are a few limitations of our study. First, we conducted our analyses on the standard four parameter BKT model used in current intelligent tutoring systems. However, researchers have extended the model to include additional parameters and our studies did not include these new variants of the BKT model. Second, we conducted most of our analyses on simulated datasets due to the lack of the ground truth in real world settings. When we conducted the analysis on the real data, we could only use the heuristic criterion that guess and slip should be less than 0.5 for checking param-

eter validity but this criterion may not be appropriate for all settings. The domain of reading tutors, for example, often have high guess and slip values associated with words.

BKT and its extension models are being used to make pedagogical discoveries and inform instructional practice in classrooms and within education technologies. The original BKT model relies on accurate and interpretable parameter values in order to make valid inferences on skill mastery. Our findings indicate that the accuracy of these inferences is sensitive to the metric chosen to fit and select the underlying model. The selection of the parameters for this student-centric model and broader development of the model among the learning analytics field is strongly impacted by the chosen error metric. It is incumbent upon the field to therefore choose the appropriate criteria to guide model selection and development that improves the chances of realizing more effective interactive learning environments for future learners.

REFERENCES

Ryan Baker, Albert Corbett, and Vincent Aleven. 2008. In *Intelligent Tutoring Systems*. Lecture Notes in Computer Science, Vol. 5091.

Joseph E. Beck and Kai-Min Chang. 2007. Identifiability: A Fundamental Problem of Student Modeling. In *Proceedings of the 11th International Conference on User Modeling (UM '07)*. Springer-Verlag, Berlin, Heidelberg, 137–146. DOI:http://dx.doi.org/10.1007/978-3-540-73078-1_17

Joseph E Beck, Kai-min Chang, Jack Mostow, and Albert Corbett. 2008. Does help help? Introducing the Bayesian Evaluation and Assessment methodology. In *International Conference on Intelligent Tutoring Systems*. Springer, 383–394.

Joseph E. Beck and June Sison. 2006. Using Knowledge Tracing in a Noisy Environment to Measure Student Reading Proficiencies. *Int. J. Artif. Intell. Ed.* 16, 2 (April 2006), 129–143. http://dl.acm.org/citation.cfm?id=1435344.1435347

Joseph E. Beck and Xiaolu Xiong. 2013. Limits to Accuracy: How Well Can We Do at Student Modeling?. In *Proceedings of the 6th International Conference on Educational Data Mining*.

Behzad Beheshti and Michel C Desmarais. 2015. Goodness of Fit of Skills Assessment Approaches: Insights from Patterns of Real vs. Synthetic Data Sets. *International Educational Data Mining Society* (2015).

Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher* 13, 6 (1984), 4–16.

Gilles Celeux. 1998. Bayesian Inference for Mixture: The Label Switching Problem. In *COMPSTAT*, Roger Payne and Peter Green (Eds.). Physica-Verlag HD, 227–232. DOI:http://dx.doi.org/10.1007/978-3-662-01131-7_26

Kaimin Chang, Joseph Beck, Jack Mostow, and Albert Corbett. 2006. A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*.

AlbertT. Corbett and JohnR. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4 (1994). http://dx.doi.org/10.1007/BF01099821

Corinna Cortes and Mehryar Mohri. 2004. AUC optimization vs. error rate minimization. *Advances in neural information processing systems* 16, 16 (2004), 313–320.

Asif Dhanani, Seung Yeon Lee, Phitchaya Mangpo Phothilimthana, and Zachary Pardos. 2014. A comparison of error metrics for learning model parameters in bayesian knowledge tracing. In *Workshop Approaching Twenty Years of Knowledge Tracing (BKT20y)*. Citeseer, 8–9.

César Ferri, José Hernández-Orallo, and R Modroiu. 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters* 30, 1 (2009), 27–38.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning* (2 ed.). Springer.

Yun Huang, José P González-Brenes, Rohit Kumar, and Peter Brusilovsky. 2015. A Framework for Multifaceted Evaluation of Student Models. *Proceedings of the 8th International Conference on Educational Data Mining* (2015), 84–91.

Kenneth R. Koedinger, John R. Anderson, William H. Hadley, and Mary A. Mark. 1997. Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education* (1997), 30–43.

Chen Lin and Min Chi. 2016. Intervention-BKT: Incorporating Instructional Interventions into Bayesian Knowledge Tracing. In *International Conference on Intelligent Tutoring Systems*. Springer, 208–218.

Ran Liu, Elizabeth A McLaughlin, and Kenneth R Koedinger. 2014. Interpreting model discovery and testing generalization to a new dataset. In *Educational Data Mining 2014*.

Kevin Murphy. 2001. The bayes net toolbox for matlab. *Computing Science and Statistics* (2001).

J. A. Nelder and R. Mead. 1965. A Simplex Method for Function Minimization. *Comput. J.* 7, 4 (1965), 308–313. DOI:http://dx.doi.org/10.1093/comjnl/7.4.308

Juraj Niznan, Jan Papousek, and Radek Pelanek. 2014. Exploring the Role of Small Differences in Predictive Accuracy Using Simulated Data.. In *AIED Workshop on Simulated Learners*.

Zachary Pardos and Neil Heffernan. 2009. Determining the Significance of Item Order in Randomized Problem Sets. In *Proceedings of the 2nd International Conference on Educational Data Mining*.

Zachary Pardos and Neil Heffernan. 2010. Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In *Proceedings of the 3rd International Conference on Educational Data Mining*.

Zachary Pardos and Michael Yudelson. 2013. Towards Moment of Learning Accuracy. In *AIED Workshop on Simulated Learners*.

Zachary A Pardos, Matthew D Dailey, and Neil T Heffernan. 2011. Learning what works in ITS from non-traditional randomized controlled trial data. *International Journal of Artificial Intelligence in Education* 21, 1-2 (2011), 47–63.

Radek Pelánek. 2015. Metrics for evaluation of student models. *Journal of Educational Data Mining* (2015).

Richard A. Redner and Homer F. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* (1984), 195–239.

Joseph Rollinson and Emma Brunskill. 2015. AFrom Predictive Models to Instructional Policies. In *Proceedings of the 8th International Conference on Educational Data Mining*.

Rinat B. Rosenberg-Kima and Zachary A. Pardos. 2014. Is this model for real? Simulating data to reveal the proximity of a model to reality. In *AIED Workshop on Simulated Learners*.

Brett van de Sande. 2013. Properties of the Bayesian Knowledge Tracing Model. *Journal of Educational Data Mining* (2013).

Michael Yudelson and Kenneth Koedinger. 2013. Estimating the benefits of student model improvements on a substantive scale. In *Educational Data Mining 2013*.

Online Appendix to:
Dueling Metrics: Choosing the Appropriate Error Metric for Models of Cognition in the Learning Analytics Field

Phitchaya Mangpo Phothilimthana, University of California, Berkeley
Seung Yeon Lee, University of California, Berkeley
Zachary A. Pardos, University of California, Berkeley

A. PARAMETERS OF THE 26 DATASETS

Table X shows numbers of student (N), numbers of questions (Q), and the model parameters (prior, learn, guess, and slip) used in generating the 26 datasets.

data	N	Q	parameters			
			prior	learn	guess	slip
1	300	5	0.500	0.444	0.321	0.123
2	3,000	5	0.200	0.444	0.321	0.123
3	3,000	5	0.800	0.444	0.321	0.123
4	3,000	5	0.500	0.000	0.321	0.123
5	30,000	5	0.500	0.000	0.123	0.321
6	3,000	5	0.485	0.236	0.395	0.173
7	3,000	5	0.146	0.525	0.225	0.351
8	3,000	5	0.622	0.015	0.734	0.497
9	3,000	5	0.674	0.356	0.813	0.590
10	3,000	5	0.145	0.542	0.356	0.825
11	3,000	5	0.745	0.683	0.570	0.735
12	3,000	10	0.135	0.356	0.013	0.223
13	3,000	10	0.872	0.145	0.514	0.014
14	3,000	10	0.175	0.375	0.015	0.532
15	3,000	10	0.256	0.714	0.614	0.520
16	3,000	10	0.618	0.154	0.389	0.820
17	30,000	5	0.245	0.385	0.012	0.001
18	30,000	5	0.734	0.002	0.726	0.555
19	30,000	10	0.164	0.393	0.032	0.375
20	30,000	10	0.724	0.155	0.726	0.830
21	30,000	5	0.200	0.250	0.650	0.700
22	30,000	5	0.300	0.350	0.750	0.500
23	30,000	5	0.400	0.450	0.450	0.750
24	30,000	5	0.464	0.700	0.250	0.200
25	30,000	5	0.564	0.800	0.350	0.400
26	30,000	5	0.643	0.900	0.250	0.600

Table X: Numbers of student (N), numbers of questions (Q), and the model parameters used in generating the 26 datasets

B. CORRELATION COEFFICIENTS

Tables XI and XII display the correlation coefficients between the error metric values and the distances to the ground truth of the 26 datasets. Table XI shows the correlation coefficients on the *entire space*, while Table XII shows the correlation coefficients on the *nearby space*.

dataset	LL	-RMSE	AUC	precision	recall	accuracy
1	0.6662	0.6938	0.5714	N/A	0.5439	0.5630
2	0.6042	0.6347	0.4785	N/A	0.4645	0.5092
3	0.5313	0.5533	0.5099	N/A	0.4519	0.4542
4	0.4845	0.5167	0.3036	N/A	0.3835	0.4189
5	0.2840	0.2761	0.3836	N/A	0.0559	0.1046
6	0.6237	0.6450	0.4847	N/A	0.4987	0.5134
7	0.4158	0.4547	0.4613	N/A	0.1065	0.3779
8	0.3615	0.3570	0.4686	N/A	0.1505	0.1713
9	0.3472	0.3650	0.5112	N/A	0.0500	0.1722
10	0.6080	0.6234	0.0879	N/A	-0.5421	0.5394
11	0.4355	0.4600	0.3234	N/A	-0.3600	0.3680
12	0.4839	0.5361	0.5697	N/A	0.2123	0.3900
13	0.6326	0.6472	0.3609	N/A	0.5472	0.5461
14	0.4135	0.4283	0.4411	N/A	-0.1880	0.3262
15	0.2473	0.2601	0.1955	N/A	0.0291	0.1031
16	0.6486	0.6601	0.2873	N/A	-0.5498	0.5495
17	0.6049	0.6596	0.5930	N/A	0.3743	0.5056
18	0.3106	0.3145	0.4753	N/A	0.0472	0.1315
19	0.4146	0.4674	0.5335	N/A	0.0391	0.4968
20	0.5033	0.5184	0.5370	N/A	-0.3533	0.3811
21	0.4536	0.4776	0.3741	N/A	-0.1901	0.3678
22	0.4662	0.4720	0.3559	N/A	0.2158	0.2689
23	0.7133	0.7343	0.2384	N/A	-0.5875	0.5931
24	0.4779	0.5081	0.5441	N/A	0.3636	0.3963
25	0.2418	0.2601	0.3218	N/A	0.1080	0.1355
26	0.2586	0.2833	0.2234	N/A	-0.2507	0.2671

Table XI: Correlation coefficients on all points over the entire parameter space with a 0.05 interval.

dataset	LL	-RMSE	AUC	precision	recall	accuracy
1	0.5114	0.5851	0.1453	-0.2114	0.2181	0.2238
2	0.4996	0.5914	0.3978	0.0826	-0.0273	0.3116
3	0.4247	0.5226	0.3088	-0.1019	0.0953	0.0903
4	0.6377	0.7231	0.5152	0.2327	-0.0706	0.4749
5	0.6890	0.7477	0.3739	0.2535	-0.1145	0.4316
6	0.5422	0.5615	0.3208	-0.2351	0.2478	0.2576
7	0.5665	0.5683	0.2997	-0.0369	0.0692	0.3062
8	0.5008	0.4965	0.1199	-0.2961	0.3073	0.3165
9	0.5101	0.5125	0.0688	N/A	-0.0729	0.3801
10	0.4887	0.5526	0.1385	N/A	N/A	N/A
11	0.4769	0.4976	-0.0080	N/A	N/A	N/A
12	0.7097	0.7816	0.2569	0.1845	0.0439	0.3970
13	0.5255	0.5510	0.4504	-0.1644	0.1591	0.1450
14	0.6120	0.5859	0.3918	N/A	-0.1631	0.1637
15	0.5219	0.5211	0.2294	N/A	-0.0304	0.1100
16	0.5090	0.5292	0.3716	N/A	N/A	N/A
17	0.7337	0.8758	0.1852	0.3194	-0.3089	0.3083
18	0.5070	0.5070	0.3835	N/A	0.0419	0.3496
19	0.6644	0.6578	0.3423	0.1512	0.2177	0.4596
20	0.5444	0.6261	0.4221	N/A	0.0192	0.4127
21	0.5257	0.5242	0.2536	N/A	0.0453	0.3269
22	0.5125	0.5137	0.3232	-0.2132	0.2115	0.2520
23	0.4954	0.5070	0.1458	N/A	N/A	N/A
24	0.5225	0.5530	0.2738	-0.0144	0.0144	0.0144
25	0.5106	0.5117	0.0728	N/A	0.2215	0.2651
26	0.5041	0.5047	0.2411	N/A	-0.1346	0.1345

Table XII: Correlation coefficients on all points over the nearby space with a 0.02 interval.

C. DISTANCES BETWEEN ESTIMATED AND TRUE PARAMETERS

Table XIIIa shows the distances between the true parameters and the parameters estimated using different error metrics. Table XIIIb shows the distances between the true parameters and the parameters estimated using LL as an error metric with different optimization algorithms: the Nelder-Mead simplex direct search ($fminsearch$) and EM.

	Distances to ground truth			
dataset	LL	-RMSE	AUC	Accuracy
1	0.16046	**0.13851**	0.35401	0.55390
2	**0.01870**	0.02018	0.42327	0.49986
3	**0.05339**	0.05379	0.25588	0.73739
4	**0.79249**	0.79318	0.83887	0.98154
5	0.00936	**0.00912**	0.86006	0.97494
6	**0.08051**	0.08259	0.24353	0.68375
7	0.05039	**0.04757**	0.62954	0.13820
8	0.08243	**0.08241**	0.76614	0.94761
9	0.05520	**0.05513**	0.20995	0.25054
10	**0.06052**	0.06425	0.42118	0.41955
11	0.70889	**0.68040**	0.81159	0.81159
12	**0.01549**	0.01618	0.38466	0.24073
13	0.00708	**0.00684**	0.18687	1.00025
14	**0.00946**	0.00958	0.67632	0.39972
15	**0.07440**	0.07469	0.72818	0.62698
16	0.03552	**0.03002**	0.42083	0.42235
17	0.00205	**0.00176**	0.66953	0.34508
18	0.02349	**0.02332**	0.35834	0.51167
19	**0.00376**	0.00398	0.46414	0.20174
20	**0.00778**	0.00835	0.14700	0.18727
21	0.00554	**0.00546**	0.43619	0.45631
22	**0.01813**	0.01817	0.62902	0.57764
23	0.02575	**0.02484**	0.40507	0.40135
24	0.03385	**0.03254**	0.55523	0.38189
25	**0.16129**	0.16217	0.96108	0.33565
26	**0.65286**	0.65288	0.86217	0.87407
mean	0.1211	0.1192	0.5269	0.5370
min	0.0020	0.0018	0.1470	0.1382
max	0.7925	0.7932	0.9611	1.0002
best	13	13	0	0

(a) Comparing error metrics using $fminsearch$

	Distances to ground truth	
dataset	fminsearch	EM
1	**0.160456**	0.160505
2	**0.018700**	**0.018700**
3	0.053392	**0.053379**
4	**0.792490**	**0.792490**
5	**0.009361**	0.009364
6	0.080506	**0.080482**
7	**0.050389**	0.050432
8	**0.082431**	0.082566
9	**0.055203**	0.055236
10	**0.060524**	0.066500
11	0.708886	**0.560981**
12	**0.015491**	**0.015491**
13	**0.007084**	**0.007084**
14	**0.009462**	0.009463
15	**0.074401**	0.107410
16	**0.035518**	0.035571
17	**0.002048**	**0.002048**
18	**0.023489**	0.025221
19	**0.003764**	**0.003764**
20	**0.007785**	**0.007785**
21	0.005540	**0.005517**
22	0.018134	**0.016804**
23	0.025751	**0.016068**
24	**0.033850**	0.033875
25	**0.161295**	0.168621
26	0.652855	**0.422812**
mean	0.1211	0.1080
min	0.0020	0.0020
max	0.7925	0.7925
best	19	14

(b) Comparing search algorithms using LL as an error metric

Table XIII: Distances between the estimated parameters and the ground truth. Bold indicates the best estimation when comparing error metrics (left) and when comparing the Nelder-Mead search and EM (right).