A Comparison of Error Metrics for Learning Model Parameters in Bayesian Knowledge Tracing

Asif Dhanani
Seung Yeon Lee
Phitchaya Phothilimthana
Zachary Pardos

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-131
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-131.html

May 29, 2014

A Comparison of Error Metrics for Learning Model Parameters in Bayesian Knowledge Tracing

Asif Dhanani* Seung Yeon Lee* Phitchaya Mangpo Phothilimthana* Zachary Pardos
University of California, Berkeley
{asifdhanani, sy.lee, mangpo, pardos}@berkeley.edu

ABSTRACT

In the knowledge-tracing model, error metrics are used to guide parameter estimation towards values that accurately represent students' dynamic cognitive state. We compare several metrics, including log likelihood (LL), root mean squared error (RMSE), and area under the receiver operating characteristic curve (AUC), to evaluate which metric is most suited for this purpose. LL is commonly used as an error metric in Expectation Maximization (EM) to perform parameter estimation. RMSE and AUC have been suggested but have not been explored in depth. In order to examine the effectiveness of using each metric, we measure the correlations between the values calculated by each and the distances from the corresponding points to the ground truth. Additionally, we examine how each metric compares to the others. Our findings show that RMSE is significantly better than LL and AUC. With more knowledge of effective error metrics for estimating parameters in the knowledge-tracing model, we hope that better parameter searching algorithms can be created.

1. INTRODUCTION

Knowledge tracing, popularized by Corbett and Anderson is a well-known method for modeling student knowledge [6]. It has been used by many intelligent tutoring systems 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 a priori. 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 primary goal of knowledge tracing is to infer latent knowledge, which is unobservable. Thus, we check the accuracy of knowledge predictions by inspecting how well the model predicts students' performance; we pick the knowledge tracing parameters that best predict performance.

There are several methods to estimate the model parameters. In Bayesian Knowledge Tracing (BKT), the model can be fit to student performance data by using either a method which finds a best goodness-of-fit measure (e.g. sum of square errors, mean absolute error, RMSE, and AUC, etc.), or a method which finds maximum likelihood. For the goodness-of-fit measures method, grid search/brute force [1] is used to find the set of parameters minimizing errors. On the other hand, Expectation Maximization [5, 9] is often used to choose parameters maximizing the LL fit to the data. Many studies have compared different modeling approaches [1, 2, 7, 13]. However, the findings are varied across the studies, and it has still been unclear which method is the best at predicting student performance [3].

When learning model parameters, one of the essential elements is the error metric that is used. Choice of a type of error metric is crucial because the error metric takes a role of guiding the estimation method to the best parameters. Pardos and Yudelson compares different error metrics to investigate which one has the most accuracy of estimating the moment of learning [11]. Our work extends this comparison by looking closer into the relationship between the three popular error metrics: LL, RMSE, and AUC, and particularly elucidating the relationship to one another closer to the ground truth point.

Likelihood of a set of parameters is the probability of the observed student outcome given those parameter values. We use the logarithm of likelihood when maximizing its value. RMSE is a commonly used measure of the differences between the predicted values by a model and the true observed values. It is defined as the square root of the mean squared error. AUC is another frequently employed measure to assess the accuracy of predictive distribution models. It is defined as the area under the receiver operating characteristic (ROC) curve, and a greater value of AUC indicates a more accurate prediction.

With the simulated data, we examine the relationship between the values calculated by the three metrics and the ground truth parameters using grid search over the entire parameter space. Section 2 describes our data generation procedure. In section 3, we compare the error metrics graphically and numerically. In section 4, we examine heat maps of distributions of LL, RMSE, and AUC values. Next, in section 5, we compare LL and RMSE in depth using scatter plots of LL values and RMSE values with colors indicating distances from the ground truth. We conclude in section 6.

*Asif Dhanani, Seung Yeon Lee, and Phitchaya Mangpo Phothilimthana contributed equally to this work and are listed alphabetically.

2. DATASETS

To assess whether LL, RMSE, or AUC is the best error metric to use in parameter searching for the BKT model, we needed datasets with known parameter values in order to compare these with the parameter values predicted by using different error metrics. In this paper, we evaluate the error metrics on the basic BKT model with four parameters: prior, learn, guess, and slip. To synthesize a dataset, we ran a simulation to generate student responses based on predefined known ground truth parameter values similar to data generation in [10]. We constructed a KT model using functions from MATLAB's Bayes Net Toolbox [8]. Each dataset contains data points of N students answering M questions. Each data point indicates whether the student's answer to the question is correct or not.

We generated 26 datasets with diverse parameter values. 15 datasets contain data for 3,000 students, and 11 datasets contain data for 30,000 students. 19 of the datasets have responses for five questions per student, and 7 of them have responses for ten. Figure 1 shows the distribution of prior, learn, guess, and slip parameter values in our datasets. Most of our datasets have low guess ($guess \leq 0.5$) which is true in most tutoring systems. However, some problem sets such as exercises in the Reading Tutor have high guess [4], so we generated some datasets with high guess as well.

Property	True	False
$prior \leq 0.5$	16	10
$learn \leq 0.5$	19	7
$guess \leq 0.5$	17	9
$slip \leq 0.5$	15	11

Figure 1: Distribution of datasets' parameter values. True and False mean number of datasets that have and do not have the specified property respectively.

3. CORRELATIONS TO THE GROUND TRUTH

In this section, we evaluate the accuracy of LL, RMSE, and AUC by analyzing the correlations between the values calculated by the different error metrics and the distances from the corresponding points to the ground truth.

3.1 Methodology

For each dataset, we evaluated LL, RMSE, and AUC values on all points over the entire prior/learn/guess/slip parameter space with a 0.05 interval. Each point P is defined as

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

On each point P, we used the MLE-hmm library [12] to calculate students' predicted responses (probability that students will answer questions correctly). We then used these predicted responses with the actual responses to calculate LL, RMSE, and AUC for all points.

To determine which error metric is the best for this purpose, we looked at the correlations between LL, RMSE, AUC values and the euclidian distances from the points to the ground truth. The distance from point P to the ground truth R is:

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

We plotted LL values vs distances, -RMSE values vs distances, and AUC values vs distances. Note that we used -RMSE instead of RMSE to standardize our convention across different error metrics. With this change, for all error metrics, higher error metric values indicate smaller error (closer to the ground truth). In addition to visualizing the results, we calculated (1) correlation between LL values and logarithm of distances, called *LL correlation*, (2) correlation between logarithm of -RMSE values and logarithm of distances, called *RMSE correlation*, and (3) correlation between logarithm of AUC values and logarithm of distances, called *AUC correlation*. We applied logarithm to all error metrics other than LL in order to compare everything in the same scale as LL. We will refer to correlations of logarithm simply as correlations.

Finally, we tested whether the correlation between the values calculated by any particular error metric and the distances is significantly stronger than the others'. We tested this by running one-tailed paired t-tests on (1) LL correlations vs RMSE correlations, (2) LL correlations vs AUC correlations, and (3) RMSE correlations vs AUC correlations on all 26 datasets.

3.2 Results

In order to determine the accuracy of the error metrics, we plotted values calculated by each error metric against distances from the ground truth in order. Figure 2 shows the graphs of dataset 2, which contains 3,000 students answering 5 questions. These graphs were generated with prior = 0.2, learn = 0.444, guess = 0.321, and slip = 0.123. As shown in the figure, both LL and -RMSE show a general pattern of being larger when the distance from the ground truth is lower. This pattern is common in all datasets. This indicates LL and RMSE as fairly good measures of distances from the ground truth. In the case of AUC, we were not able to discern any observable relation between distances from the ground truth and AUC values. In certain cases, AUC exhibited a similar pattern to that of RMSE and LL, but the appearance of this pattern was extremely inconsistent.

The correlations we calculated between the error metrics and the distances from the ground truth support the findings from our visual representation. Both suggest RMSE as the best indicator. The average LL correlation, RMSE correlation, and AUC correlation were 0.4419, 0.4827, and 0.3983 respectively. We define that an error metric A is *better* than B if the correlation between values calculated by an error metric A and the distances to the ground truth are higher than that of B. Figure 3 shows the results of correlation comparison of the three metrics. It appears that RMSE was consistently better than LL on all 26 datasets. However, RMSE or LL correlations are not always higher than AUC correlations—RMSE is better than AUC on 18 out of 26 datasets, and LL is better than AUC on 15 datasets.

Nevertheless, the result from the one-tailed paired t-test shown in Figure 4 reveal that RMSE is better, on average,

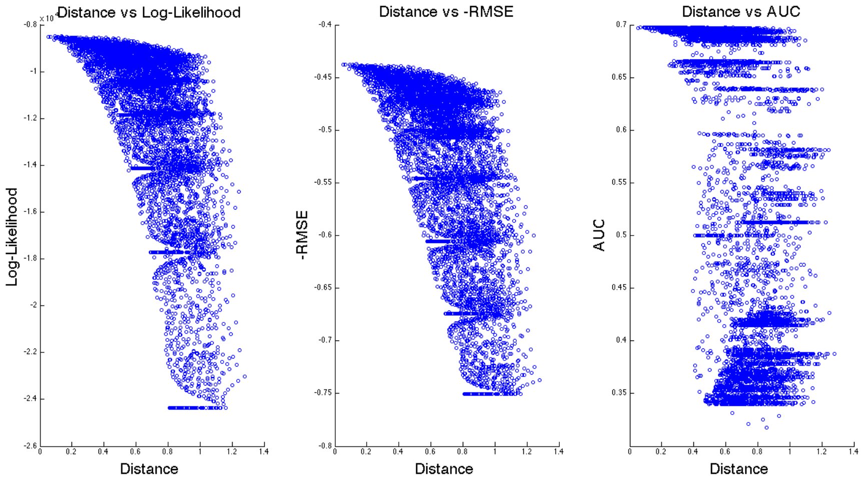


Figure 2: Values calculated by different error metrics vs distances to the ground truth

Comparison		Number of datasets	Total
RMSE vs LL	RMSE > LL	26	26
	RMSE < LL	0	
RMSE vs AUC	RMSE > AUC	18	26
	RMSE < AUC	8	
LL vs AUC	LL > AUC	15	26
	LL < AUC	11	

Figure 3: Correlation comparisons of error metrics

Comparision	Δ of correlations	t	p-value
RMSE > LL	0.0408	8.9900	<< 0.0001
RMSE > AUC	0.0844	2.7583	0.0054
LL > AUC	0.0436	1.4511	0.0796

Figure 4: T-test statistics of RMSE correlation > LL correlation, RMSE correlation > AUC correlation, and LL correlation > AUC correlation

than both LL and AUC. This difference is statistically significant. Also, LL is better than AUC on average but the difference is not statistically significant.

4. DISTRIBUTIONS OF VALUES

In this section, we examine further why RMSE seemed to be better than LL and AUC by looking at the distributions of LL, RMSE, and AUC values over the range of parameters represented by our datasets.

4.1 Methodology

We visualized the values of LL, -RMSE, and AUC of all points over the 2 dimensional guess/slip space with a 0.02 interval. We fixed prior and learn parameter values to the actual ground truth values. We created heat maps of LL, -RMSE, and AUC. Using the guess and slip parameters as the axes, we visualize LL, -RMSE, and AUC values by color. The colors range from dark red to dark blue corresponding to the values ranging from low to high. The range of colors is shown in Figure 5(d).

4.2 Results

Figure 5 shows the heat maps of LL, -RMSE, and AUC on the same dataset used in the previous section. The white dot in each graph represents the location of the ground truth.

In all heat maps, the white dot (ground truth) is located in the darkest blue region (the area with the highest values for each metric). The heat maps of LL and -RMSE are fairly similar, while the heat map of AUC is very different from the other two. The heat maps of LL and -RMSE have a very concentrated region of high LL and -RMSE values (the darkest blue region), while the high AUC values region is very spread out. This pattern is consistent throughout all datasets. Hence, we conclude that AUC is a fairly poor indicator for how close parameters are to the ground truth.

The heat maps also provide further support to using RMSE instead of LL. If we follow the gradient from the lowest value to the highest value in the LL heat map, the gradient is very high at the beginning (far from the ground truth) and is very low at the end (close to the ground truth). Whereas in the -RMSE heat map, the change in the gradient is low. Additionally, notice that the darkest blue region in -RMSE heat map is smaller than that in LL heat map. This suggests that we may be able to refine the proximity of the ground truth better with RMSE.

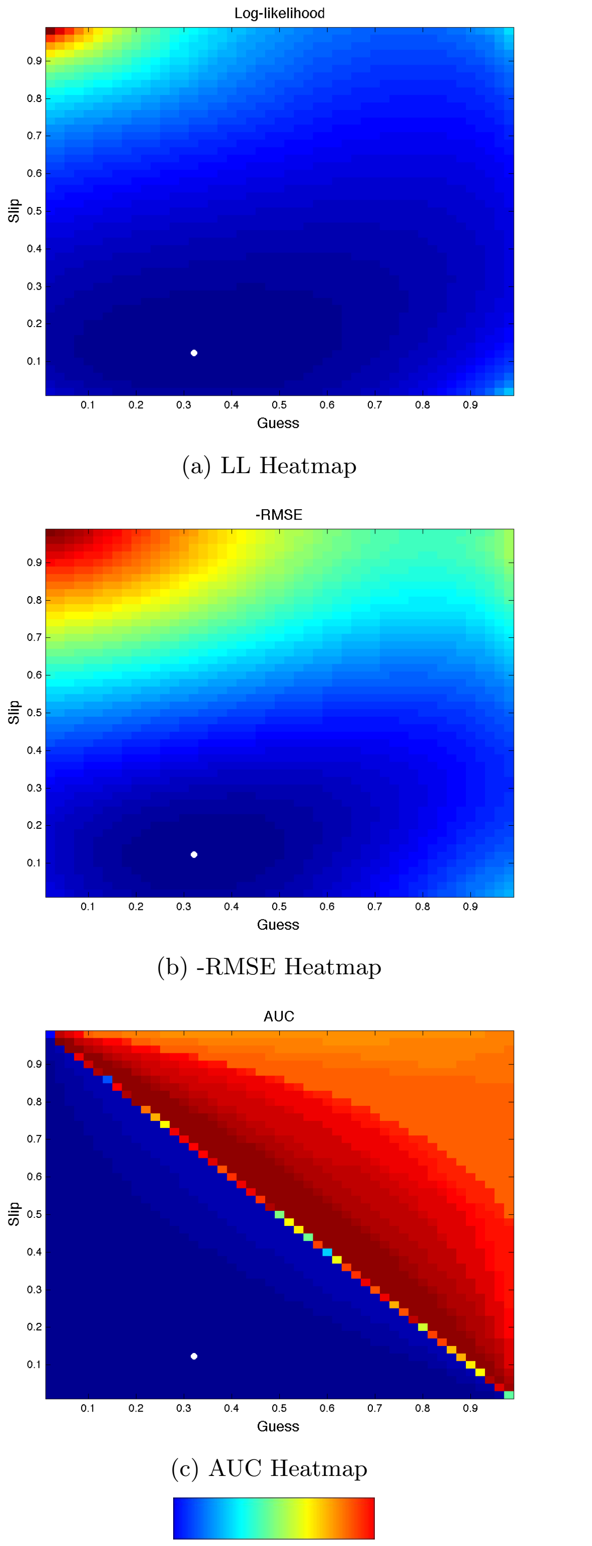
Because AUC proved to be a poor indicator of the distance to the ground truth, we excluded it as a metric for the remainder of the comparisons.

5. DIRECT COMPARISON: LL AND RMSE

After looking at the correlations between distances from the ground truth and the values calculated by the various error metrics, and the distribution of the values calculated by each metric, we further compare LL and RMSE to investigate RMSE's apparent outperformance of LL.

5.1 Methodology

We plotted LL values and RMSE values of all points against each other in order to observe the behavior of the two metrics



(a) LL Heatmap

(b) -RMSE Heatmap

(c) AUC Heatmap

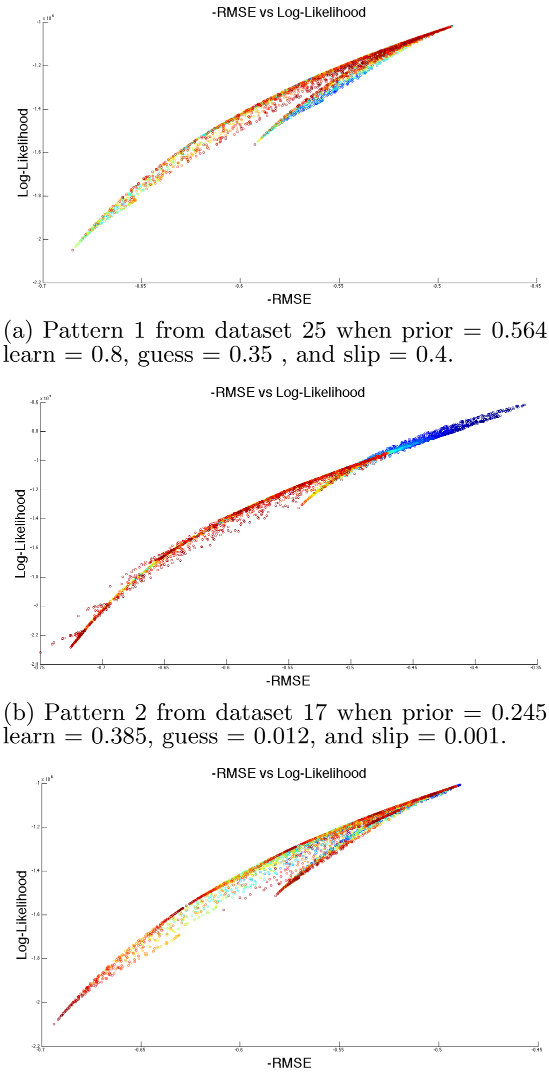
(d) Range of colors.

Figure 5: LL, -RMSE, and AUC values when fixing prior and learn parameter values and varying guess and slip parameter values. Red represents low values, while blue represents high values. The white dots represent the ground truth.

in detail. We then labeled each data point by its distance to the ground truth with a color. The range of colors is the same as used for the heat maps.

5.2 Results

In our graphical comparison of LL and RMSE, we were able to observe a general pattern for the distribution of data points. Figure 6 shows the LL vs -RMSE graphs from 3 different datasets. The graphical results from these three were the most representative of our overall datasets. As expected, LL values and RMSE values correlate logarithmically. In addition to the main logarithmic curve, there is an additional curve, which we will refer to as the *hook*, which exists for a portion of the graph. The length of the hook varies from dataset to dataset, ranging from being almost unobservable to existing for the entirety of the graph. The hook converges with the main curve when the -RMSE and LL values are both sufficiently high— the points are very



(a) Pattern 1 from dataset 25 when prior = 0.564, learn = 0.8, guess = 0.35 , and slip = 0.4.

(b) Pattern 2 from dataset 17 when prior = 0.245, learn = 0.385, guess = 0.012, and slip = 0.001.

(c) Pattern 3 from dataset 22 when prior = 0.3, learn = 0.35, guess = 0.75, and slip = 0.5.

Figure 6: LL vs -RMSE.

close to the ground truth. Before the convergence, a set of parameters with an RMSE value may have multiple LL values and vice versa. The 3 different graphs in Figure 6 show different types of hooks.

Figure 6(a) represents the most common pattern among all datasets. For a portion of the graph before the convergence, when we look at a fixed LL value with varied RMSE values, most points in the hook have higher -RMSE values and are closer to the ground truth than do the points in the main curve. This pattern further supports our argument that RMSE values and distances to the ground truth correlate strongly. However, when we look at a fixed RMSE value with varied LL values, the points in certain parts of the hook have lower LL values but are closer to the ground truth than do the points in the main curve. The result shows that in the divergent area, LL values and distances do not correlate. This evidence explains why the RMSE correlations are higher than the LL correlations as seen in the previous section. As both the curve and the hook converge, we can infer that after this point, both RMSE and LL will give similar estimates of the ground truth. However, for a portion of the graph before this point, RMSE is a better predictor of ground truth values.

Figure 6(b) displays the pattern in which the hook is almost aligned with the main curve. In this pattern, when we fix an RMSE value, LL values only vary slightly, and vice versa. Hence, both LL and RMSE seem to give similar estimates of the ground truth in this case.

The last pattern, shown in Figure 6(c), is characterized by points close to the ground truth appearing between the main curve and the hook. More specifically, these points are closer to the ground truth than the surrounding points on the main curve and the hook. Neither LL nor RMSE seem to produce an optimal representation in this case. This demonstrates that RMSE may not always be the best measure of ground truth. Further investigation is necessary to distinguish what causes this shape and how to find the best parameter values in this type of situation.

6. CONCLUSION

In our comparison of LL, RMSE, and AUC as metrics for evaluating the closeness of estimated parameters to the true parameters in the knowledge tracing model, we discovered that RMSE serves as the strongest indicator. RMSE has a significantly higher correlation to the distance from the ground truth on average than both LL and AUC. Additionally, AUC appears to be a poor metric for fitting BKT parameters, similar to the finding in[11]. Our detailed comparison of LL and RMSE revealed that when the estimated parameter value is not very close to the ground truth, RMSE is the best indicator of distance to the ground truth. We would recommend further studies to examine which error metrics are best to use for the region after the hook and main curve converge. The effectiveness of teaching systems without human supervision relies on the ability of the systems to predict the implicit knowledge states of students. We hope that our work can help advance the parameter learning algorithms used in the knowledge tracing model, which in turn can make these teaching systems more effective.

7. REFERENCES

[1] R. Baker, A. Corbett, S. Gowda, A. Wagner, B. MacLaren, L. Kauffman, A. Mitchell, and S. Giguere. Contextual slip and prediction of student performance after use of an intelligent tutor. In *User Modeling, Adaptation, and Personalization*, volume 6075 of *Lecture Notes in Computer Science*. 2010.

[2] R. S. Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 2008.

[3] R. S. Baker, Z. A. Pardos, S. M. Gowda, B. B. Nooraei, and N. T. Heffernan. Ensembling predictions of student knowledge within intelligent tutoring systems. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, 2011.

[4] J. Beck and K.-m. Chang. Identifiability: A fundamental problem of student modeling. In *User Modeling 2007*, volume 4511 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2007.

[5] K. Chang, J. Beck, J. Mostow, and A. Corbett. A bayes net toolkit for student modeling in intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 2006.

[6] A. Corbett and J. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 1994.

[7] Y. Gong, J. Beck, and N. Heffernan. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Intelligent Tutoring Systems*, volume 6094 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010.

[8] K. Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 2001.

[9] Z. Pardos and N. Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*. 2010.

[10] Z. A. Pardos and N. T. Heffernan. 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*, 2010.

[11] Z. A. Pardos and M. V. Yudelson. Towards moment of learning accuracy. In *Proceedings of the 1st AIED Workshop on Simulated Learners*, 2013.

[12] Z. A. Pardos et al. *Scaling Cognitive Models in the Massive Open Environment*, in preparation.

[13] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis –a new alternative to knowledge tracing. In *Proceedings of Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, 2009.