Is this Data for Real?

Rinat B. Rosenberg-Kima
University of California, Berkeley
rosenbergkima@berkeley.edu

Zachary Pardos
University of California, Berkeley
pardos@berkeley.edu

ABSTRACT
Simulated data plays a central role in Educational Data Mining and in particular in Bayesian Knowledge Tracing (BKT) research. The initial motivation for this paper was to try to answer the question: *given two datasets could you tell which of them is real and which of them is simulated?* The ability to answer this question may provide an additional indication of the goodness of the model, thus, if it is easy to discern simulated data from real data that could be an indication that the model does not provide an authentic representation of reality, whereas if it is hard to set the real and simulated data apart that might be an indication that the model is indeed authentic. In this paper we will describe initial analysis that was performed in an attempt to address this question. Additional findings that emerged during this exploration will be discussed as well.

Keywords
Bayesian Knowledge Tracing (BKT), simulated data, parameters space.

1. INTRODUCTION
Simulated data has been increasingly playing a central role in Educational Data Mining [1] and Bayesian Knowledge Tracing (BKT) research [1, 4]. For example, simulated data was used to explore the convergence properties of BKT models [5], an important area of investigation given the identifiability issues of the model [3]. In this paper, we would like to approach simulated data from a slightly different angle. In particular, we claim that the question,"*given two datasets could you tell which of them is real and which of them is simulated?",* is interesting as it can be used to evaluate the goodness of a model and may potentially serve as an alternative metric to RMSE, AUC, and others. We would like to start approaching this problem in this paper by comparing simulated data to real data with Knowledge Tracing as the model.

Knowledge Tracing (KT) models are widely used by cognitive tutors to estimate the latent skills of students [6]. Knowledge tracing is a Bayesian model, which assumes that each skill has 4 parameters: two knowledge parameters including initial (prior knowledge) and learn rate, and two performance parameters including guess and slip. KT in its simplest form assumes a single point estimate for prior knowledge and learn rate for all students, and similarly identical guess and slip rates for all students. Simulated data has been used to estimate the parameter space and in particular to answer questions that relate to the goal of maximizing the log likelihood (LL) of the model given parameters and data, and improving prediction power [7], [8], [9].

In this paper we would like to use the KT model as a framework for comparing the characteristics of simulated data to real data, and in particular to see whether it is possible to distinguish between the real and sim datasets.

2. DATA SETS
To compare simulated data to real data we started with 2 real dataset generated from the assisstment software[1] (specifically, datasets G6.207-exact.txt with 776 students and G6.259-exact.txt with 212 students) from a previous BKT study [10]. Both of the datasets consist of 6 questions in linear order where all students answer all questions. Next, we generated synthetic, simulated data using the best fitting parameters that were found for the real data as the generating parameters. By this we generated a simulated version of dataset G6.207 and a simulated version of dataset G6.259 that had the exact same number of questions, number of students, and was generated with what appears to be the best fitting parameters. The specific best fitting parameters that were found for each dataset and were used to generate the simulated data are presented in table 1.

Table 1. Best fitting parameters for each dataset. These parameters were used to generate the simulated datasets.

	N	Prior	Learn	Guess	Slip
G6.207	776	.453	.068	.270	.156
G6.259	212	.701	.044	.243	.165

3. METHODOLOGY
We are interested to find out whether it is possible to distinguish between the simulated data and the real data. The approach we took was to calculate LL for the gird of all the parameters space (prior, learn, guess, and slip). We hypothesized that the LL pattern of the simulated data and real data will be different across the parameters space. For each of the matrices we conducted a grid search with intervals of .04 that generated 25 intervals for each parameter and 390,625 total combinations of prior, learn, guess, and slip. For each one of the combinations LL was calculated and placed in a four dimensional matrix. We used fastBKT [11] to (a) calculate the best fitting parameters of the real datasets, (b) generate simulated data, and (c) calculate the LL of the parameters space. Additional code in Matlab and R was generated to put all the pieces together[2]. In particular, we calculated the LL for all the combinations of two parameters where the other two parameters were fixed to the best fitting value. In an additional analysis, we let all parameters be free and took the average LL for all combinations of two parameters, collapsed over the space of the other two parameters not visualized. The motivation for this was to visualize the error space interactions in the four dimensions of the model.

[1] Data can be obtained here: http://people.csail.mit.edu/zp/

[2] Matlab and R code will be available here:
[2] Matlab and R code will be available here:
http://myweb.fsu.edu/rr05/

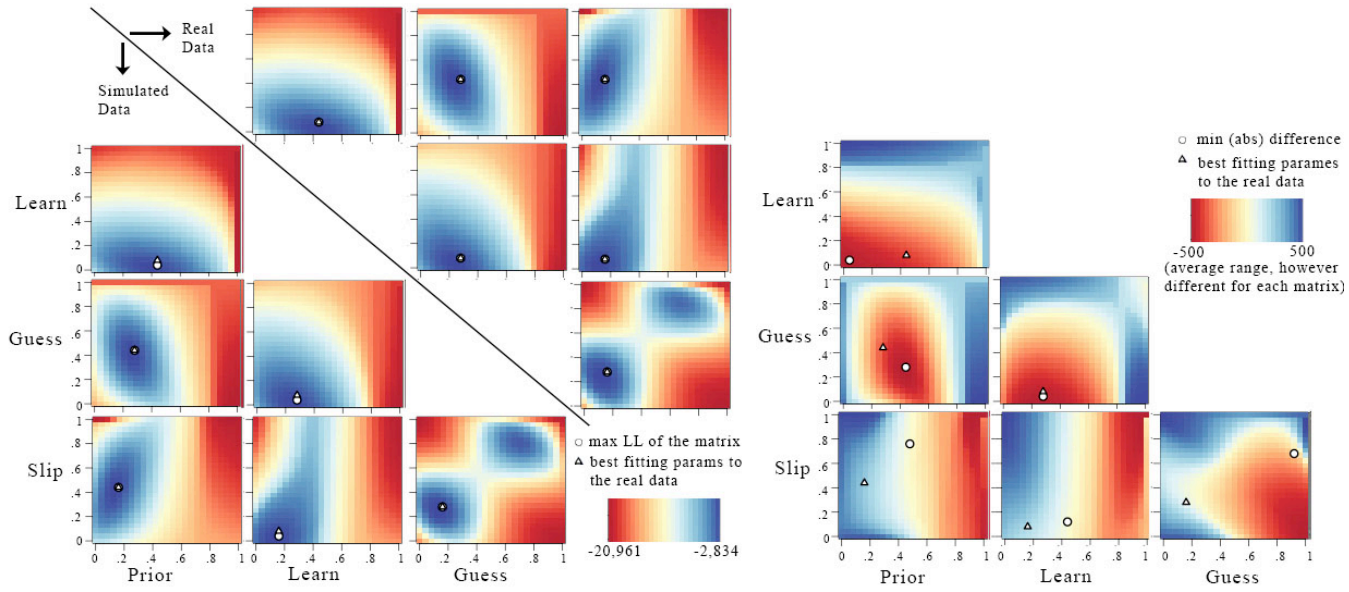

Figure 1.a (left). Heat maps of LL of real assistment dataset G6-207 (k=776 students) and a corresponding simulated data that was generated with the best fitting parameters of the real dataset. The two parameters not in each figure were fixed to the best parameters. Blue areas indicate high LL, and red areas indicate lower LL. Circles indicate maximum LL of the given matrix, and triangles indicate the best fitting parameters to the real data (that were also used to generate the simulated data). In this case the triangles and circles fit the same point.

Figure 1.b (right). Heat maps of delta LL between real dataset G6-207 and the corresponding simulated data that was generated with the best fitting parameters of the real dataset. The two parameters not in each figure were fixed to the best parameters. Blue areas indicate high difference between the real and sim LL, and red areas indicate lower difference. Circles indicate minimum absolute delta of the given matrix, and triangles indicate the best fitting parameters to the real data.

4. DOES THE LL OF SIM vs. REAL DATA LOOK DIFFERENT?

Our initial thinking was that as we are using a simple BKT model, it is not authentically reflecting reality in all its detail and therefore we will observe different patterns of LL across the parameters space between the real data and the simulated data. The LL space of simulated data in [5] was quite striking in its smooth surface but the appearance of real data was left as an open research question.

4.1 Does the LL of sim vs. real data looks different across two parameters grids?

First, we calculated the LL over all the combinations of two parameters for dataset G6.207 where the other two parameters were fixed to the best fitting value. For example, when we calculated LL for the combination of slip and prior (top right figure in figure 1.a), we fixed learn and guess to be .068 and .270 accordingly. To our great surprise, when we plotted heat maps of the LL matrices of the real data and the simulated data (Figure 1.a - real data is presented in the upper triangle and simulated (sim) data is presented in the lower triangle) we received what appears to be identical matrices (for example, the upper right heat map is the (slip x prior) LL matrix of the real data, whereas the lowest left heat map is the (slip x prior) LL matrix of the sim data).

The extent of the similarity between the matrices was surprising and in order to get a better picture of the differences between them we plotted heat maps of the deltas between the real data and the simulated data (LL_RealData-LL_SimData) for each matrix. Even though the matrices appear to be identical, as can be seen in Figure 1.b, there is in fact a difference between the LL of the matrices although it is not a big difference compared to the values of LL. Another surprising finding was that the LL of the real data was in many cases higher than the LL of the sim data. We expected that the model would better explain the sim data as there should not be additional noise as expected in reality, and therefore the LL of the sim data should be higher, yet the findings were not consistent with this expectation.

Another interesting finding was that the location of the ground truth (the triangle) in most of the cases resulted in smaller delta between the real and the sim data although not in all cases (e.g., guess x slip). Note that the circles in Figure 1.b indicate the minimum absolute difference in LL between the real and the sim data, and this point is usually not located at the exact ground truth (except for learn x guess).

Another interesting finding can be seen in Figure 1.a - slip vs. guess. Much attention has been given to this LL space which revealed the apparent co-linearity of BKT with two primary areas of convergence, the upper right area being a false, or "implausible" converging area as defined by [3]. What is interesting in this figure is that despite what appears to be two global maxima, the point with the best LL in this dataset is in fact the lower region for both sim and real data.

Next we conducted the same analysis with the second dataset.

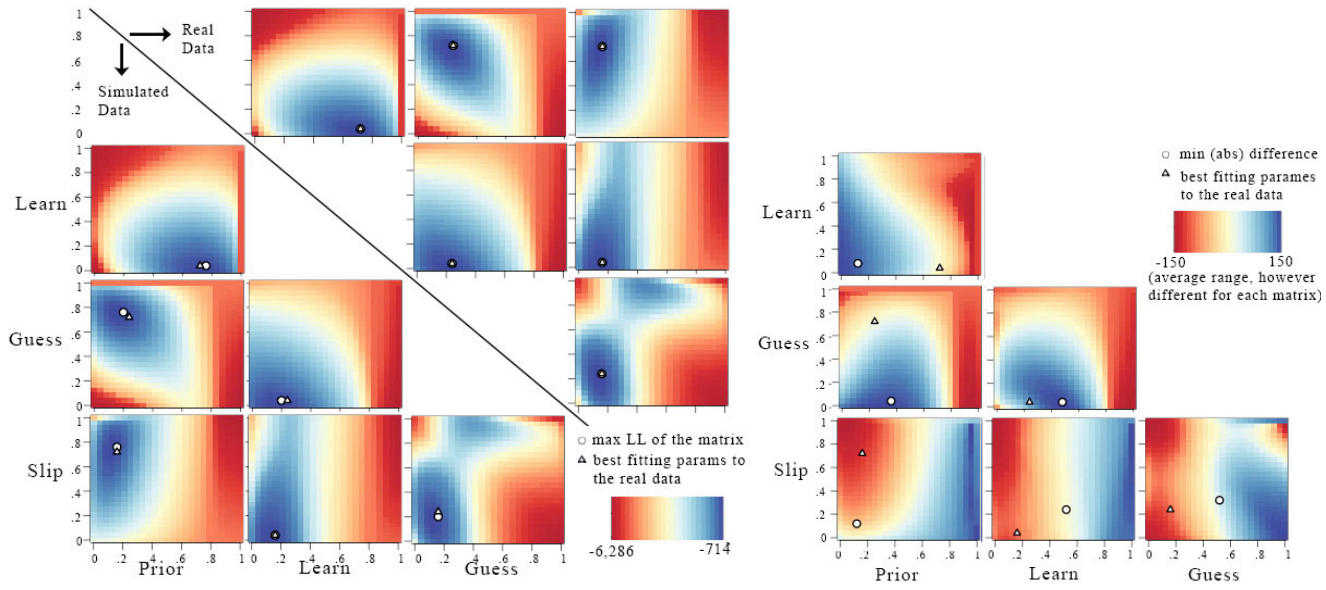

Figure 2.a (left) Heat maps of delta LL between real dataset G6-259 (k=212 students) and the corresponding simulated data that was generated with the best fitting parameters of the real dataset. The two parameters not in each figure were fixed to the best parameters. Blue areas indicate high difference between the real and sim LL, and red areas indicate lower difference. Circles indicate maximum LL of the given matrix, and triangles indicate the best fitting parameters to the real data.

Figure 2.b (right). Heat maps of delta LL between real assistment dataset G6-259 and the corresponding simulated data that was generated with the best fitting parameters of the real dataset. The average is across the two parameters not in each figure. Blue areas indicate high difference between the real and sim LL, and red areas indicate lower difference. Circles indicate minimum absolute delta of the given matrix, and triangles indicate the best fitting parameters to the real data.

Even though the G6-259 dataset was significantly smaller than the first dataset, we received very similar results to the first dataset with surprisingly similar heat maps for the sim and real data (see Figure 2.a). Like in the first dataset, notice that even though the LL heat maps look very similar, there is a difference in the delta heat maps (see Figure 2.b). Nevertheless, there is an interesting difference between the two datasets. Concretely, unlike the bigger dataset (G6-207), in G6-259 the LL of the sim data was actually higher than the real data in most cases.

4.2 What if we average LL over 2 parameters across all the combinations of the other 2 parameters?

We were interested to find out how will the heat maps look like if we do not fix the other two parameters to be best fit, but rather average the LL across the entire space of the other two parameters. For example, to calculate the matrix of guess and slip we practically calculated a matrix of guess and slip LL for each combination of learn and prior (25 x 25 = 625 matrices) instead of only one matrix for the best fit learn and prior. Then, we took the average of all these matrices for each combination of guess and slip (see Figure 3.a). The results are both surprising and interesting. As far as (guess x slip), we no longer receive the two maximum (global and local) that we received when learn and prior where fixed to best fit parameters. Another interesting finding is the relationship between the average maximum across the other two parameters and the overall best fit parameters for given two parameters. For example, if we look at the heat map of matrix (learn x prior) we can see that there is not a big difference between the average maximum point (white circle) and the overall best fit parameters (white triangle). This may indicate that changing guess and slip will not affect the value of learn and prior that maximizes the LL, therefore might suggest independency. If we look at (guess x learn), we see that changes in prior and slip will again not have an impact on the best fit value of guess, however, they will affect the value of learn. Then again, if we look at the heat map of (prior x guess), we will see that both prior and guess are sensitive to changes in learn and slip. Yet again, the extremely surprising part of these results is that the sim data appear to be almost identical to the real data. It is possible to see from Figure 3.b though that indeed there are differences between the simulation data and the real data and like before, the LL of the real data is higher than that of the sim data in the larger dataset.

Like for the fixed matrices, we received similar LL matrices for the smaller dataset (G6-259) (see table 4.a). In addition, as before, the LL of the sim data for this dataset was higher than that of the real data (the opposite direction of the larger dataset G6-207). Another interesting finding for this dataset can be seen in the (guess x slip) matrices (4.b). Notice that while the sim data converged to the lower point of the blue area, the real data converged to the higher point. Nevertheless, this only happened in the averages matrices and not in the fixed ones.

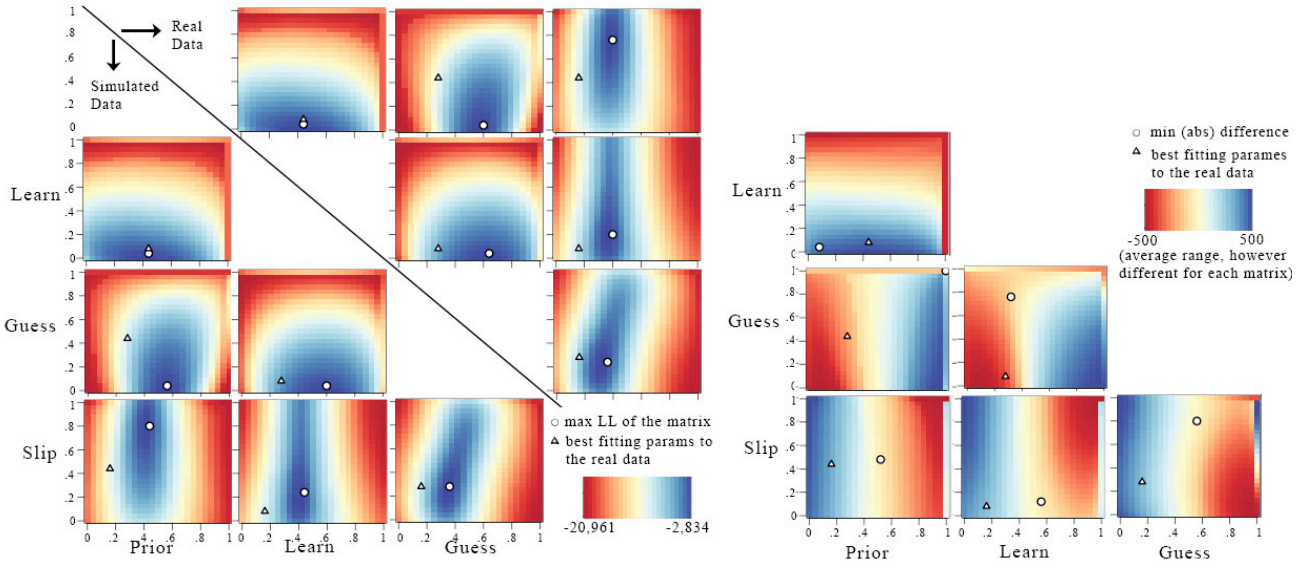

Figure 3.a (left). Heat maps of average LL of real assistment dataset G6-207 (k=776 students) and a corresponding simulated data that was generated with the best fitting parameters of the real dataset. The average is across the two parameters not in each figure. Blue areas indicate high LL, and red areas indicate lower LL. Circles indicate maximum LL of the given matrix, and triangles indicate the best fitting parameters to the real data (that were also used to generate the simulated data).

Figure 3.b (right). Heat maps of delta LL between real assistment dataset G6-207 and the corresponding simulated data that was generated with the best fitting parameters of the real dataset. The average is across the two parameters not in each figure. Blue areas indicate high difference between the real and sim LL, and red areas indicate lower difference. Circles indicate minimum absolute delta of the given matrix, and triangles indicate the best fitting parameters to the real data.

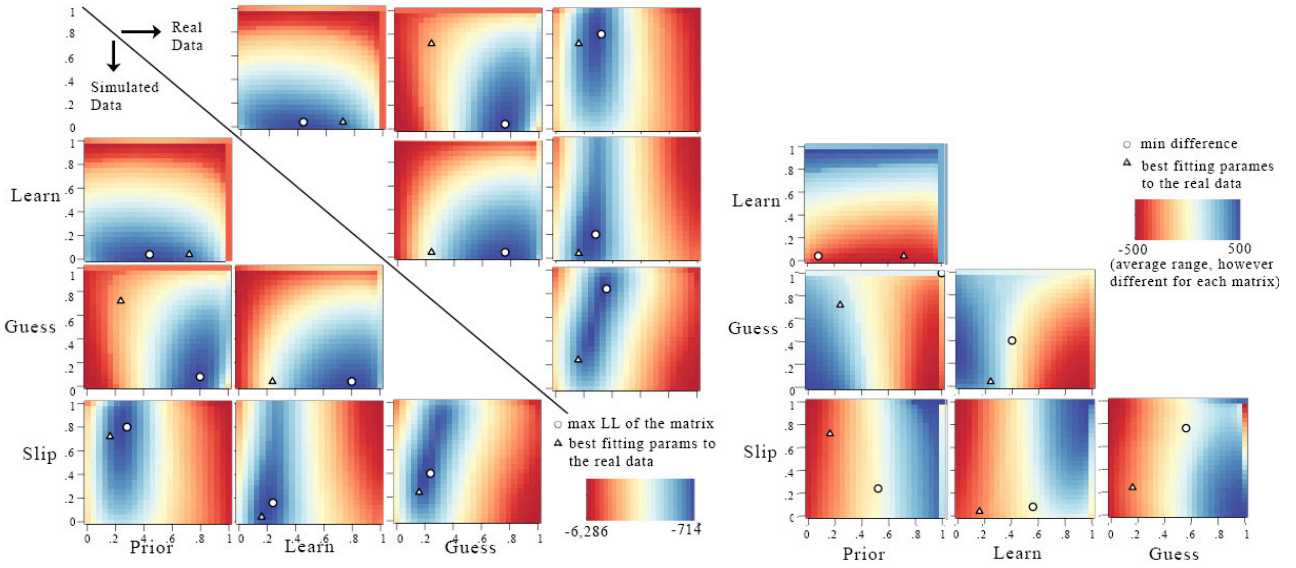

Figure 4.a (left). Heat maps of average LL of real assistment dataset G6-259 (k=212 students) and a corresponding simulated data that was generated with the best fitting parameters of the real dataset. The average is across the two parameters not in each figure. Blue areas indicate high LL, and red areas indicate lower LL. Circles indicate maximum LL of the given matrix, and triangles indicate the best fitting parameters to the real data (that were also used to generate the simulated data).

Figure 4.b (right). Heat maps of delta LL between real assistment dataset G6-259 and the corresponding simulated data that was generated with the best fitting parameters of the real dataset. The average is across the two parameters not in each figure.

5. DISCUSSION AND FUTURE WORK

The initial motivation of this paper was to find whether it is possible to discern a real data from a sim data. If for a given model it is possible to tell apart a sim data from a real data then the authenticity of the model can be questioned. This line of thinking is in particular typical of simulation use in Science context, where different models are used to generate simulated data, and then if a simulated data has a good fit to the real phenomena at hand, then it may be possible to claim that the model provides an authentic explanation of the system [12]. We believe that it may be possible to generate a new matric for evaluating the goodness of a model by comparing a simulated data from this model to real data.

In this work we explored similarities between simulated and real data. Nevertheless, we are yet to answer the question "is this data for real?". In other words, what we still did not do in this work is come up with an algorithm that can take a dataset and determine whether it is real or simulated. Another way to think of it is to come out with an algorithm that can tell us whether it is possible to discern real and simulated data and use it as an indication of the goodness of the model. We found differences between the real and sim data, but are they strong enough to be noticed by such algorithm in a consistent way? In future work we plan to further investigate this question by creating a training set of multiple real datasets and sim datasets and use machine learning techniques to extract a learning algorithm from this training dataset that can take as input a dataset and determine whether it is real or sim. We argue that if such algorithm can be found, it is an indication that the underlying model can be improved. In future work we also plan to compare different variations of the KT model and contrast their resulting simulated data with real data. In particular we plan to generate a more complex set of simulated data that is based on a more complex model (e.g., different learning rate for different types of questions), and then use it as "real" data with the (wrong) assumption that the model is simple (standard BKT model) to simulate a scenario where the real data is indeed grounded in more complex model than our assumptions and see what results would a learning algorithm that uses this "real" data in comparison to a sim data will yield.

In addition, this paper raises interesting questions that we did not think of while trying to answer our initial question. For example, it seems like there is potential to dive deeper into the average LL (Figures 3&4) and find more about the relationships and dependencies between the different parameters. Another question that emerged is how could it be that the simulated data had lower LL than the real data in the bigger dataset yet lower in the smaller dataset? Further analysis is needed to answer these questions.

Last but not least, given the remarkable resemblance between the sim data and the real data, these initial findings provide an indication that the BKT model is a model with a very strong hold in reality.

6. REFERENCES

[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.

[2] M. C. Desmarais and I. Pelczer, "On the Faithfulness of Simulated Student Performance Data.," in *EDM*, 2010, pp. 21–30.

[3] J. E. Beck and K. Chang, "Identifiability: A fundamental problem of student modeling," in *User Modeling 2007*, Springer, 2007, pp. 137–146.

[4] Z. A. Pardos and M. V. Yudelson, "Towards Moment of Learning Accuracy," in *AIED 2013 Workshops Proceedings Volume 4*, 2013, p. 3.

[5] 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 *EDM*, 2010, pp. 161–170.

[6] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253–278, 1994.

[7] S. Ritter, T. K. Harris, T. Nixon, D. Dickison, R. C. Murray, and B. Towle, "Reducing the Knowledge Tracing Space.," *Int. Work. Group Educ. Data Min.*, 2009.

[8] R. S. d Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere, "Contextual slip and prediction of student performance after use of an intelligent tutor," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 52–63.

[9] 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 *Intelligent Tutoring Systems*, 2008, pp. 406–415.

[10] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 255–266.

[11] Z. A. Pardos and M. J. Johnson, "Scaling Cognitive Modeling to Massive Open Environments (in preparation)," *TOCHI Spec. Issue Learn. Scale*.

[12] U. Wilensky, "GasLab—an Extensible Modeling Toolkit for Connecting Micro-and Macro-properties of Gases," in *Modeling and simulation in science and mathematics education*, Springer, 1999, pp. 151–178.