Tracking student progress in a game-like learning environment with a Monte Carlo Bayesian Knowledge Tracing model

G.-H. Gweon
Department of Physics
University of California
Santa Cruz, CA, USA
1-831-459-1806
gweon@ucsc.edu

Hee-Sun Lee
Department of Physics
University of California
Santa Cruz, CA, USA
1-831-459-2326
hlee58@ucsc.edu

Chad Dorsey
The Concord Consortium
25 Love Lane
Concord, MA 01742
1-978-405-3100
cdorsey@concord.org

Robert Tinker
The Concord Consortium
25 Love Lane
Concord, MA 01742
1-978-405-3225
bob@concord.org

William Finzer
Concord Consortium West
6550 Vallejo St., Suite 101C
Emeryville, CA 94608
1-510-984-4380
wfinzer@concord.org

Daniel Damelin
The Concord Consortium
25 Love Lane
Concord, MA 01742
1-978-405-3242
ddamelin@concord.org

ABSTRACT

The Bayesian Knowledge Tracing (BKT) model is a popular model used for tracking student progress in learning systems such as an intelligent tutoring system. However, the model is not free of problems. Well-recognized problems include the identifiability problem and the empirical degeneracy problem. Unfortunately, these problems are still poorly understood and how they should be dealt with in practice is unclear. Here, we analyze the mathematical structure of the BKT model, identify a source of the difficulty, and construct a simple Monte Carlo BKT model to analyze the problem in real data. Using the student activity data obtained from the ramp task module at the Concord Consortium, we find that the Monte Carlo BKT analysis is capable of detecting the identifiability problem and the empirical degeneracy problem, and, more generally, gives an excellent summary of the student learning data. In particular, the student activity monitoring parameter M emerges as the central parameter.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; K.3.1 [**Computers and Education**]: Computer Uses in Education

General Terms

Algorithms, Measurement, Theory

LAK '15, March 16–20, 2015, Poughkeepsie, NY, USA

http://dx.doi.org/10.1145/2723576.2723608.

Keywords

Bayesian Knowledge Tracing, Monte Carlo, Educational data mining

1. INTRODUCTION

The Bayesian Knowledge Tracing (BKT) model [5] is widely used in the context of educational data mining [3, 4, 2, 6, 8]. It offers a simple model where student knowledge can be estimated as the student activity is scored in a structured online environment such as an intelligent tutoring system.

However, a major problem is that estimating student knowledge is often ambiguous. Known as the "identifiability problem" [3, 4], this problem means that completely different sets of model parameters may produce the same student performance curve, while estimating quite different knowledge.

Another problem is the "empirical degeneracy" [8] (or the "model degeneracy") problem—sometimes the model would predict low knowledge on high performance or high knowledge on low performance. Contextualizing certain parameters [1] or limiting the ranges of parameters are some adopted solutions to this problem in the literature.

These two problems are not understood well, and general solutions to these problems do not exist. Here, combining mathematical analysis and real data analysis, we show that these two problems can be dealt with in practice and are highly correlated. We suggest to use a new parameter (M; see Eq. 2 below) as an important detector of these problems as well as student learning. Our data analysis employs a new method, which we call the Monte Carlo BKT method.

2. THE BKT MODEL IN THEORY

The BKT model was originally developed by Corbett and Anderson [5]. This model involves four parameters, each of them having a numerical value from 0 to 1.

$p(L_1)$ This is the initial knowledge that a student has prior to taking on any learning activities.

$p(T)$ This is probability that the student will transition

from an unknowing state to a knowing state, as the result of using the knowledge during a unit of activity.

$p(G)$ This is the "guess parameter" that corresponds to the probability that the student will choose the correct answer in an activity, while the student has not acquired the required knowledge.

$p(S)$ This is the "slip parameter" that corresponds to the probability that the student will choose an incorrect answer in an activity, while the student has acquired the required knowledge.

Within the BKT model, these four parameters completely determine the latent knowledge $p(L_n)$ and the student performance curve $p(C_n)$, where n is the index of the activities in which the student has opportunities to apply the knowledge. $p(L_n)$ is the knowledge level estimated right before activity n (or, equivalently, after activity $n-1$; $n = 1, 2, ...$), and $p(C_n)$ is the probability that the student will get the correct answer on activity n. By fitting the actual student performance data with $p(C_n)$, we can obtain the estimates of the above four parameters, from which we can produce the student knowledge curve $p(L_n)$. As the student goes through activities, the typical outcome is that $p(L_n)$ increases; however, the model also allows, in principle, the opposite case in which $p(L_n)$ decreases as n increases.

2.1 The BKT model without measurement

Here, we envision a *purely theoretical* process, as originally considered by Beck and Chang [3, 4]. Imagine a student is carrying out learning activities, but withholds her/his answers. Therefore, there is no actual measurement of student learning. However, the student is learning.

$p(C_n)$ is the theoretical student performance curve. If measurements were made, then we would inevitably find that the actual performance curve is different from $p(C_n)$ due to the statistical nature of data.

The advantage of considering the BKT model without measurement is that such statistical noise can be ignored and we can use $p(C_n)$ as though it is the actual student performance data. In this case, it can be shown (see the next section; also, see Ref. [8]) that

$$p(L_{n+1}) = p(T) + (1 - p(T))\, p(L_n). \tag{1}$$

This implies a geometric series involving $p(L_n)$, and the series can be readily summed up, giving the following results (see the next section for more information on the derivation).

$$\begin{aligned} M &\equiv (1 - p(S) - p(G)) \cdot (1 - p(L_1)), && (2)\\ p(C_n) &= 1 - p(S) - M \cdot (1 - p(T))^{n-1}, && (3)\\ p(L_n) &= 1 - (1 - p(L_1))\,(1 - p(T))^{n-1}. && (4) \end{aligned}$$

If we define

$$n_T \equiv -\frac{1}{\log(1 - p(T))} \tag{5}$$

then we can rewrite our results for $p(C_n)$ and $p(L_n)$ as

$$\begin{aligned} p(C_n) &= 1 - p(S) - Me^{-(n-1)/n_T}, && (6)\\ p(L_n) &= 1 - (1 - p(L_1))\, e^{-(n-1)/n_T}. && (7) \end{aligned}$$

So, n_T tells us how fast or slow $p(C_n)$ and $p(L_n)$ approach their respective asymptotes, $1 - p(S)$ and 1. n_T is a **scale parameter** for the number of activities required in order for the learning to be perfected.

Our equations clearly explain the origin of the identifiability problem. While the theory has four parameters, $p(C_n)$ depends on *only three independent parameters*, M, $p(T)$, and $p(S)$; this provides the motivation for introducing the new symbol M. Any two different sets of values for $p(G)$ and $p(L_1)$ that give the same M value will give the same performance curve $p(C_n)$. Worse, there are infinite such sets in general.

2.2 The BKT model with measurement

In an actual BKT modeling, measurements *are* made, as student scores are available. The student score, denoted by s, may be Boolean (0 or 1) or continuous (0 to 1). Here, we consider the general case (the latter)

$$0 \le s \le 1, \qquad \text{student score.} \tag{8}$$

In this case, the posterior probability of the student knowledge, given the evidence of score s, is given by

$$\begin{aligned} p(L_n|s) &= s \cdot p(L_n|C_n) + (1 - s) \cdot p(L_n|I_n) \\ &= \left[\frac{s \cdot (1 - p(S))}{p(C_n)} + \frac{(1 - s)p(S)}{1 - p(C_n)}\right] p(L_n), \quad (9) \end{aligned}$$

where I_n means incorrect answer at step n. Given this posterior probability and the following two equations [5], $p(L_n)$ and $p(C_n)$ are determined completely by s values at each n and four parameters, $p(L_1)$, $p(T)$, $p(G)$, and $p(S)$.

$$\begin{aligned} p(L_{n+1}) &= p(L_n|s) + (1 - p(L_n|s))\, p(T), && (10)\\ p(C_n) &= p(L_n)\,(1 - p(S)) + (1 - p(L_n))\, p(G). && (11) \end{aligned}$$

The case of a continuum value of s has not been discussed in the literature to our knowledge, and so it is worth noting the following point. If we restrict the value of s to be Boolean, 0 or 1, then $p(L_n|s)$ becomes either $p(L_n|C_n)$ or $p(L_n|I_n)$, and so our model reduces to the more common BKT model employing Boolean student scores [2, 6]. As a side note, if we use $s = p(C_n)$, then all results in the previous section can be derived also.

Is the identifiability problem absent in the BKT model *with* measurement?

This question is a very important one. As we will show in this paper with real data, the answer is no, unless some other feature of the data places a strong constraint on $p(L_1)$ or $p(G)$.

From a theoretical point of view, also, it seems a bit too optimistic to conclude [8] that the identifiability problem does not exist in the BKT model with measurement just because the model now involves all four parameters in predicting the student performance. The reason why all four parameters are involved is only due to the statistical noise in student score s. It seems sensible to expect then, that, on average, *some* identifiability problem persists.

3. THE BKT MODEL IN PRACTICE

The BKT model is an ideal fit to use in a game-like learning environment, supposing that the following conditions are met. (1) Each level of activities must challenge students to learn one specific piece of knowledge. (2) Student must complete at least four activities at each level, thereby producing at least four score data points for the BKT model involving four parameters to fit.

3.1 The ramp game

The student score data analyzed in this paper were obtained from the "ramp task module" at the Concord Consortium. This game-like learning module is explained in more detail in another paper in this issue [7].

In the ramp task, students were asked to determine a height so that a car could land on a particular target. The ramp task consisted of five challenges, or five levels, requiring students to apply more and more sophisticated knowledge about the ramp system as follows.

Challenge 1: relationship between height and a fixed landing location.

Challenge 2: relationship between height and moving landing locations.

Challenge 3: relationship between height and moving landing locations when a friction value is changed from the previous challenge.

Challenge 4: relationship between height and moving landing locations when mass of the car is changed.

Challenge 5: relationship between friction and moving landing locations when starting height and mass are fixed.

Each challenge level was comprised of multiple steps: 3 steps for Challenges 1 and 4; 4 steps for Challenges 2 and 3; 6 steps for Challenge 5. Students' performances were scored automatically on a 0 to 100 scale based on how close to the target the car stopped. If students scored 67 points or higher, they progressed to the next step within that level. If students finished all required steps within a level, they moved to the first step of the next Challenge. Student scores were normalized to a 0 to 1 scale for our analysis.

Therefore, the data from the ramp game module were ideally suited for applying the BKT model analysis to continuous score data.

3.2 Monte Carlo BKT

Given the theoretical ideas discussed in Section 2, how might one extract BKT fit parameter values from the data? Clearly, our goal must be obtaining the *distributions* of fit parameter values for a given single data set. We achieve this goal by the following procedure, which we name a Monte Carlo BKT method, since the goal of the procedure is obtaining the probability distribution.

1. The standard Levenberg-Marquardt non-linear least squares fit algorithm is applied with randomly selected initial fit parameter values, $p(G)$, $p(L_1)$, $p(S)$, and $p(T)$.

2. At least 200 successful randomly initialized fits are collected, to ensure good statistics.

3. More fit results are collected, if necessary, until each parameter value converges within a set (found-to-be-sufficient) tolerance, 5×10^{-4} for this work. If this is achieved within a preset maximum number of fit trials, then the program is stopped and success is declared.

We place no restriction on parameter values: any parameter value can take any value between 0 and 1. After the program stops successfully, the average fit parameter values can be taken as representing the given data set, as we did in Ref. [7].

4. RESULTS AND DISCUSSION

Fig. 1 shows the Monte Carlo BKT fit results. Table 1 summarizes simple statistics for the fit parameters, including the total number of good fits N_{total} required for the convergence of the Monte Carlo iteration. Not surprisingly, data with practically no noise (data sets 1 and 4) converge very quickly, while others require many more iterations.

4.1 Robustness of $p(S)$

In all examined cases, $p(L_n)$ steadily increases over time. We interpret this as the tendency of the theory to follow **the positive eventual outcome scenario**.

$$p(L_n) \to 1, \quad \text{for } n \gg n_T, \tag{12}$$

$$p(C_n) \to 1 - p(S), \quad \text{for } n \gg n_T. \tag{13}$$

Here, the second equation follows from the first, through Eq. 11. Also, note that n_T is very small (Table 1), and so these asymptotic behaviors appear already for small n.

The positive eventual outcome is a rigorous mathematical property of the BKT model without measurement (Eq. 4) as long as $p(T) > 0$. We see that this scenario is also realized in all our examined cases that involve measurements.

According to Eq. 13, the student score data at large n values is determined by only one parameter, $p(S)$. Therefore, the $p(S)$ parameter must be determined without any ambiguity. Indeed, our $p(S)$ distribution is always sharp.

Since the latent knowledge approaches 1 in the end, should we say that all students acquired perfect knowledge? Clearly, this is not the case [7]. It seems that a sensible choice is to take $1-p(S)$ as the more practical *demonstrable* knowledge, which seems as important as, if not more important than, the latent knowledge.

4.2 Meaning of M

From Eqs. 2 and 11, we get

$$p(C_1) = 1 - p(S) - M. \tag{14}$$

This describes the initial value of the performance curve. In practice, there may be noise in the data, and $p(C_1)$ must be assessed with such noise filtered out. Looking at Eqs. 13 and 14, we see that M corresponds to the overall increase in performance. From Eq. 2, $-1 \le M \le 1$.

Thus, M makes it possible to *monitor student activities with one number*. If M is large and positive, say greater than 0.3, then the learning is progressing well. If M is close to zero, then the learning is stalling. If M is negative, then the learning is regressing. So, we are able to identify excellent learning (data sets 1, 6) and poor learning (data sets 5, 7) just by looking at the M values in Table 1.

4.3 The identifiability problem exists

The identifiability problem is diagnosed if any parameter distribution is broad. From Fig. 1, this is clearly the case for data sets 2, 4, and 7. **So, the identifiability problem exists even in the BKT model with measurement.**

More quantitatively, the identifiability problem is detected by large values of σ_G and σ_{L_1} (data sets 2, 5, and 7; and, to a lesser degree, data sets 3 and 4). We see that in most of these cases, there is a large negative correlation between $p(G)$ and $p(L_1)$, as shown in the last column of Table 1 and as expected from Eq. 2 and the discussion in Section 2.

The data that do not show the identifiability problem are characterized by the narrow distribution of fit parameter

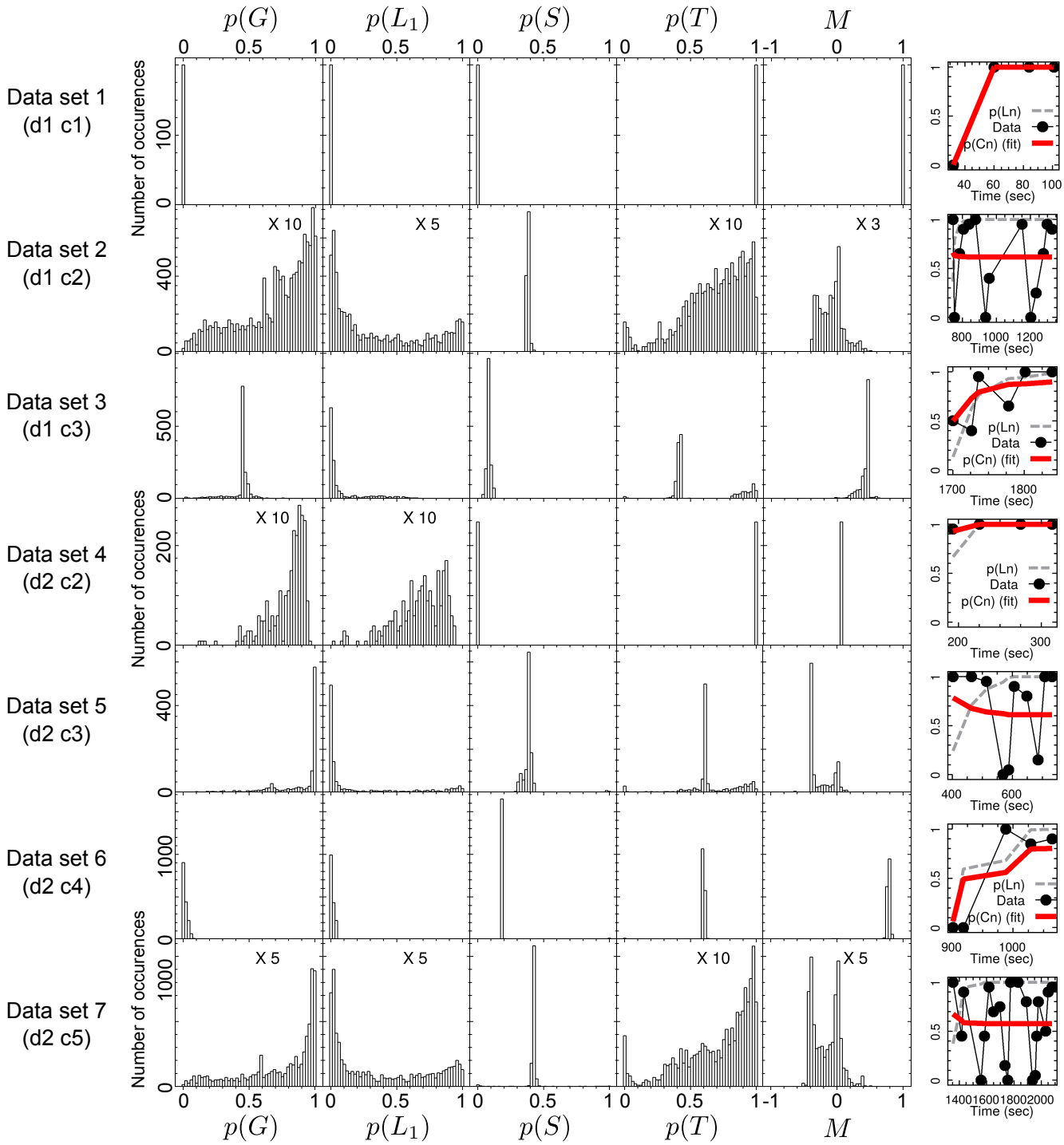

Figure 1: Monte Carlo samples of BKT parameters, $p(G)$, $p(L_1)$, $p(S)$, $p(T)$, and M. M is a derived "monitoring" parameter, defined by Eq. 2. Each data set is labeled as "d<l> c<n>," which stands for "day <l> challenge/level <n>." The overall behaviors of the complete collection of data sets have been discussed in Ref. [7]; here we focus on seven typical examples for our in-depth analysis. For each data set, we show the histograms for Monte Carlo samples of fit parameters. Some histograms were scaled up by the shown factors for clarity. On the rightmost column, data (connected black symbols) are shown with the fit (thick solid red lines) and the knowledge estimate (gray dashed lines), calculated using the averaged fit parameter values. The data are plotted as a function of time after session login by student. The sample sizes (N_{total}) that were required for convergence and other basic statistics are given in Table 1.

Data	N_{total}	$p(G)$ (σ_G)	$p(L_1)$ (σ_{L_1})	$p(S)$ (σ_S)	$p(T)$ (σ_T)	n_T	M (σ_M)	Corr(G, L_1)
1	200	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0	1.00 (0.00)	-0.61
2	1211	0.69 (0.27)	0.37 (0.35)	0.38 (0.03)	0.68 (0.24)	0.9	-0.09 (0.18)	-0.49
3	1564	0.44 (0.13)	0.13 (0.21)	0.10 (0.10)	0.57 (0.25)	1.2	0.40 (0.15)	-0.20
4	247	0.78 (0.15)	0.67 (0.19)	0.00 (0.00)	1.00 (0.00)	0	0.05 (0.00)	-0.79
5	1217	0.84 (0.24)	0.24 (0.34)	0.39 (0.09)	0.64 (0.21)	1.0	-0.24 (0.19)	-0.77
6	1725	0.03 (0.11)	0.04 (0.15)	0.20 (0.09)	0.59 (0.08)	1.1	0.76 (0.14)	0.89
7	1367	0.74 (0.28)	0.37 (0.36)	0.42 (0.07)	0.71 (0.26)	0.8	-0.17 (0.20)	-0.66

Table 1: Basic statistics for Monte Carlo BKT fit parameters presented in Fig. 1. Average values are presented along with standard deviations (σ's) in parentheses.

values. Data sets 1 and 6 are good examples. Here, $p(C_1) \approx 0$, which means, from Eq. 11, that $p(L_1) \approx 0$ and $p(G) \approx 0$. This constraint leads to no identifiability problem.

4.4 Empirical degeneracy can be detected

The empirical degeneracy problem can be diagnosed in the rightmost column of Fig. 1. Data sets 2, 5, and 7 clearly show this problem, since $p(L_n)$ increases while $p(C_n)$ decreases over time. In our work, empirical degeneracy is regarded as something that we can detect by analysis, rather than something that we avoid, e.g., by artificially limiting the values of $p(S)$ and $p(G)$ to small values. A theoretical analysis of the BKT inference iteration [8] shows that the empirical degeneracy condition corresponds to $p(S)+p(G) > 1$. According to this condition, our data sets 2, 5, and 7 show empirical degeneracy, in good agreement with our visual diagnosis.

We find that M is the detector of empirical degeneracy as well. M and $p(S) + p(G)$ have an extremely high negative correlation ($r = -0.99$), which is not surprising given Eq. 2. So, empirical degeneracy is detected by a negative value of M (Table 1). **Therefore, empirical degeneracy is a sign of poor learning** (cf., Section 4.2).

Going further, we also find that **the empirical degeneracy problem and the identifiability problem are also highly correlated**. The identifiability problem necessarily leads to large values of σ_G and σ_{L_1}. These values and the value of M again show a very high negative correlation ($r = -0.93$), showing that M can also detect the identifiability problem.

5. CONCLUSIONS

In this paper, we presented data from student online learning activities and their unbiased analysis using a Monte Carlo BKT model. The outcome of the analysis shows a problematic learning can be detected through the low M parameter value, which in turn indicates the empirical degeneracy problem and the identifiability problem. The entire numerical procedure that starts from reading the raw log data from a database and ends with complete Monte Carlo fits takes about 2 seconds to 20 seconds per data set, depending on tolerance setting. Therefore, our procedure is amenable to real time implementation in educational settings.

6. ACKNOWLEDGMENTS

We are grateful to Trudi Lord and Cynthia McIntyre for giving valuable feedback on the manuscript. This material is based upon work supported by the National Science Foundation under grants REC-1147621 and REC-1435470. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors acknowledge Edmund Hazzard who developed curriculum activities for the ramp task, and students and teachers who participated in this study.

7. REFERENCES

[1] R. S. 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*, pages 52–63. Springer, 2010.

[2] R. S. J. d. Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. In B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, editors, *Intelligent Tutoring Systems*, number 5091 in Lecture Notes in Computer Science, pages 406–415. Springer Berlin Heidelberg, Jan. 2008.

[3] J. E. Beck. Difficulties in inferring student knowledge from observations (and why you should care). In *Educational Data Mining: Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education*, pages 21–30, 2007.

[4] J. E. Beck and K.-m. Chang. Identifiability: A fundamental problem of student modeling. In C. Conati, K. McCoy, and G. Paliouras, editors, *User Modeling 2007*, number 4511 in Lecture Notes in Computer Science, pages 137–146. Springer Berlin Heidelberg, Jan. 2007.

[5] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, Dec. 1994.

[6] J. D. Gobert, M. Sao Pedro, J. Raziuddin, and R. S. Baker. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4):521–563, Sept. 2013.

[7] H.-S. Lee, G.-H. Gweon, C. Dorsey, R. Tinker, W. Finzer, D. Damelin, N. Kimball, A. Pallant, and T. Lord. How does Bayesian Knowledge Tracing model emergence of knowledge about a mechanical system? In *LAK15 Conference Proceedings*, LAK15, 2015.

[8] B. v. d. Sande. Properties of the Bayesian Knowledge Tracing model. *JEDM - Journal of Educational Data Mining*, 5(2):1–10, July 2013.