Carnegie Mellon University
Research Showcase @ CMU

Human-Computer Interaction Institute | School of Computer Science

2008

More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing

Ryan S.J.d. Baker
Carnegie Mellon University

Albert T. Corbett
Carnegie Mellon University

Vincent Aleven
Carnegie Mellon University

Follow this and additional works at: http://repository.cmu.edu/hcii

Recommended Citation
.

This Article is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Human-Computer Interaction Institute by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing

Ryan S.J.d. Baker, Albert T. Corbett, Vincent Aleven

Human-Computer Interaction Institute, Carnegie Mellon University
{rsbaker, corbett, aleven} @cmu.edu

Abstract. Modeling students' knowledge is a fundamental part of intelligent tutoring systems. One of the most popular methods for estimating students' knowledge is Corbett and Anderson's [6] Bayesian Knowledge Tracing model. The model uses four parameters per skill, fit using student performance data, to relate performance to learning. Beck [1] showed that existing methods for determining these parameters are prone to the *Identifiability Problem:* the same performance data can be fit equally well by different parameters, with different implications on system behavior. Beck offered a solution based on Dirichlet Priors [1], but, we show this solution is vulnerable to a different problem, *Model Degeneracy*, where parameter values violate the model's conceptual meaning (such as a student being more likely to get a correct answer if he/she does not know a skill than if he/she does). We offer a new method for instantiating Bayesian Knowledge Tracing, using machine learning to make contextual estimations of the probability that a student has guessed or slipped. This method is no more prone to problems with Identifiability than Beck's solution, has less Model Degeneracy than competing approaches, and fits student performance data better than prior methods. Thus, it allows for more accurate and reliable student modeling in ITSs that use knowledge tracing.

1 Introduction

Modeling students' knowledge is a fundamental part of intelligent tutoring systems. Key aspects of modern intelligent tutoring systems, such as deciding which problems to give students [cf. 6], are reliant upon accurately estimating each student's knowledge state at any given time. Giving students appropriate amounts of practice on each skill promotes complete and efficient learning [4]; both over-practice and under-practice can be avoided through having student knowledge models that are as accurate and dependable as possible.

In recent years, Corbett & Anderson's Bayesian Knowledge Tracing model [6] has been used to model student knowledge in a variety of systems, including tutors for mathematics [9], computer programming [6], and reading skill [2], and is statistically equivalent to the two-node dynamic Bayesian network used in many other learning environments [10]. Bayesian Knowledge Tracing keeps a running assessment of the probability that a student currently knows each skill, continually updating that estimate based on student behavior. Cognitive Mastery Learning built on top of Bayesian Knowledge Tracing has been shown to significantly improve student learning [6].

However, a recent paper by Beck and Chang [2] gives evidence that current methods for developing Bayesian Knowledge Tracing models for specific skills are vulnerable to a statistical problem, the ***identifiability*** problem, where models with equal-

ly good statistical fit to performance data may make very different predictions about a student's knowledge state, and correspondingly may assign very different numbers of problems to a student. To address this problem, Beck and Chang [2] proposed constraining model parameters by finding a prior probability across all skills. However, as we will show in this paper, Beck and Chang's solution is vulnerable to a different statistical problem, which we term ***model degeneracy***, where it is possible to obtain model parameters which lead to paradoxical behavior, such as the probability the student knows a skill dropping after three correct answers in a row. In this paper, we propose both theoretical and empirical definitions of this problem.

These two problems, at their core, arise from how these models handle uncertainty – in particular, how these models address the possibility of a student slipping (knowing a skill, but giving a wrong answer) or guessing (giving a correct answer, despite not knowing the skill). Within this paper, we propose a new way to assess uncertainty within knowledge tracing, using machine learning to make contextual estimations of the probability that a student has guessed or slipped. We show that this method leads to a significantly closer fit between models and student performance than prior methods, has lower model degeneracy than these approaches, and that there is reason to believe that this method will not suffer from the identifiability problem.

1.1 Bayesian Knowledge Tracing

Corbett and Anderson's Bayesian Knowledge Tracing model [6] computes the probability that a student knows a given skill at a given time, combining data on the student's performance up to that point with four model parameters. In the model's canonical form, each problem step in the tutor is associated with a single cognitive skill. The model assumes that at any given opportunity to demonstrate a skill, a student either knows the skill or does not know the skill, and may either give a correct or incorrect response (help requests are treated as incorrect by the model). A student who does not know a skill generally will give an incorrect response, but there is a certain probability (called **G**, the Guess parameter) that the student will give a correct response. Correspondingly, a student who does know a skill generally will give a correct response, but there is a certain probability (called **S**, the Slip parameter) that the student will give an incorrect response. At the beginning of using the tutor, each student has an initial probability ($\mathbf{L_0}$) of knowing each skill, and at each opportunity to practice a skill the student does not know, the student has a certain probability (**T**) of learning the skill, regardless of whether their answer is correct.

The system's estimate that a student knows a skill is continually updated, every time the student gives a first response (correct, incorrect, or a help request) to a problem step. First, the system re-calculates the probability that the student knew the skill before the response, using the evidence from the response (help requests are treated as evidence that the student does not know the skill), using the first two equations of Figure 1. Then, the system accounts for the possibility that the student learned the skill during the problem step, using the third equation of Figure 1. Within the Cognitive Mastery algorithm used in most Cognitive Tutors [6], the student is assigned additional problems on skills that the system does not yet believe that the student has learned (e.g. skills that the student has less than 95% probability of knowing).

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * \big(1 - P(S)\big) + \ (1 - P(L_{n-1})) * (P(G))}$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + \ (1 - P(L_{n-1})) * (1 - P(G))}$$

$$P(L_n|Action_n) = \ P(L_{n-1}|Action_n) + \big((1 - P(L_{n-1}|Action_n)\big) * P(T))$$

Figure 1. The equations used to predict student knowledge from behavior in Bayesian Knowledge Tracing.

The four parameters in Bayesian Knowledge Tracing are fit, for each skill, using data from students using that skill within an intelligent tutor. The goal during parameter fitting is to figure out which combination of parameters best predicts the pattern of correct and incorrect responses in the existing data, and then to use that model to make predictions about new students' knowledge as they use the tutor.

Challenges in Estimating Parameters for Bayesian Knowledge Tracing Models

Recently, Beck and Chang [2] showed that common methods of fitting Bayesian Knowledge Tracing models suffer from the ***identifiability*** problem; different combinations of the four parameters can fit the same data equivalently well, but yield very different estimates of the probability that the student knows the skill at any given time. This is of practical importance, as different model parameters may require very different amounts of practice before inferring that the student has reached mastery.

A second challenge for fitting models in Bayesian Knowledge Tracing approach is what we term ***model degeneracy.*** The conceptual idea behind using Bayesian Knowledge Tracing to model student knowledge in intelligent tutors is that knowing a skill generally leads to correct performance, and that correct performance implies that a student knows the relevant skill; hence, by looking at whether a student's performance is correct, we can infer whether they know the skill. A model deviates from this theoretical conception, and thus is ***theoretically degenerate***, when its guess (**G**) parameter or slip (**S**) parameter is greater than 0.5. A slip parameter over 0.5 signifies that a student who knows a skill is more likely to get a wrong answer than a correct answer; similarly, a guess parameter over 0.5 implies that a student who does not know a skill is more likely to get a correct answer than a wrong answer.

It is also possible to conceive of empirical tests that show that a model violates the linkage between knowledge and performance; we term a model that fails such a test ***empirically degenerate.*** We propose two tests for empirical degeneracy. First, if a student's first N actions in the tutor are correct, the model's estimated probability that the student knows the skill should be higher than before these N actions. Second, if the student makes a large number M of correct responses in a row, the model should assess that the student has mastered the skill. The exact values of N and M are arbitrary – within this paper, we choose N=3 and M=10 as reasonable cut-off points for the two tests. In other words, if a student's first three actions in the tutor are all correct, but the model's estimated probability that the student knows the skill is lower than before these three actions, we say that the model failed the first test of empirical

degeneracy. If a student gets a skill correct ten times in a row without reaching skill mastery, we say that the model failed the second test of empirical degeneracy.

Three Prior Approaches to Model Fitting in Bayesian Knowledge Tracing

The simplest ***baseline approach*** to fitting a Bayesian Knowledge Tracing model is to allow each of the four parameters to take on any value between 0 and 1. We fit parameters for this approach using Bayes Net Toolkit-Student Modeling (BNT-SM) [1].

An alternate approach is to bound the guess and slip parameters (the ***bounded guess and slip approach***). Generally, in existing tutors, the guess parameter is bounded to be between 0 and 0.3, and the slip parameter is bounded to be between 0 and 0.1, based on the most common number of candidate actions, and pragmatically, in order to err in the direction of requiring less practice for mastery. Though this approach was not explicitly designed to prevent model degeneracy, it makes theoretical degeneracy impossible. We fit parameters for this approach using Microsoft Excel.

A third way to fit a Bayesian Knowledge Tracing model is the ***Dirichlet Priors*** approach proposed in [1, 2]. A Gaussian probability distribution is found for how often different values of each parameter are seen across skills, and then the parameters of all skills are constrained by these prior probabilities. This approach biases all skills towards parameters that fit the whole data set well, with skills that have less data biased more strongly than skills that have large amounts of data. The prior probabilities lead to a single model always being the best-fitting model among the space of potential models, for each skill. We fit parameters for this approach using BNT-SM.

2 Analyzing Degeneracy in Previous Approaches

Prior work has already shown that the baseline and the bounded guess-and-slip approaches are vulnerable to the identifiability problem; the Dirichlet priors approach gives a single prediction, offering a response to the identifiability problem [1, 2]. In this section, we use data from the Middle School Tutor [9], an intelligent tutor which covers a wide span of mathematical topics covered by 6th-8th grade students (approximately 12-14 years old), to analyze whether these three model-fitting approaches are prone to problems with model degeneracy, and examine their accuracy. 232 students used the Middle School Tutor during the course of the 2002-2003 school year, making 581,785 transactions (either entering an answer or requesting a hint) on 171,987 problem steps covering 253 skills in 37 tutor lessons/units. 290,698 additional transactions were not included in either these totals or in our analyses, because they were not labeled with skills, information needed to apply Bayesian Knowledge Tracing.

Table 1 shows the level of theoretical and empirical model degeneracy for each of the three approaches. 76% of skills in the Dirichlet priors model and 75% of skills in the baseline model were theoretically degenerate; as a direct consequence of bounding guess and slip, 0% of skills in the bounded guess and slip model were theoretically degenerate. The difference between the bounded guess and slip model and each of the other two models was statistically significant, using the test of the significance of the difference between correlated proportions with McNemar's standard

Table 1. The number (proportion) of degenerate skills in each model.

Modeling Approach	Theoretically Degenerate Skills	Skills where a student who gets first three actions correct has lower P(**L**) afterwards (Empirical degeneracy test 1)	Skills where a student cannot reach mastery with 10 correct answers in a row (Empirical degeneracy test 2)
Baseline	189 (75%)	4 (2%)	57 (23%)
Bounded Guess and Slip	0 (0%)	0 (0%)	12 (5%)
Dirichlet Priors	192 (76%)	4 (2%)	57 (23%)

error estimate [7], Z=13.86, Z=13.75, two-tailed p<0.0001. The Dirichlet priors and baseline model were not significantly different from one another, Z=0.83, two-tailed p=0.41. Across models, failures of the first test of empirical degeneracy were rare; only 2% of the skills in the Dirichlet priors and baseline models failed this test, and 0% of skills in the bounded guess and slip model failed. However, 23% of the skills in the Dirichlet priors and baseline models failed the second test of empirical degeneracy. Fewer (5%) of the skills in the bounded guess and slip model failed the second test of empirical degeneracy, in both cases Z=5.58, two-tailed p<0.0001.

3 Contextual Estimation of Guess and Slip

In this section, we will discuss a new approach to Bayesian Knowledge tracing, which removes one of the framework's assumptions, to address these modeling issues. In all three prior approaches, each of the four parameters is held constant across contexts, for any given skill. (One other prior approach changed parameter values depending on whether help was used or not [5], and anticipates our approach, although that approach's contextualization was far simpler than what is proposed here).

In the new approach we propose, we contextually estimate whether each individual student response is a guess or a slip, rather than using fixed guess and slip probability estimates across all situations. In this section, we describe our method for predicting whether individual actions are guesses or slips. We will then discuss how these predictions are integrated into a new approach to Bayesian Knowledge Tracing. Our method is as follows:

- We take a set of correct responses in the log files. For each correct student response, we apply a Bayesian analysis to estimate the probability the student knew the applicable rule or guessed, based on the student's performance on successive opportunities to apply the rule. A similar procedure is used to assess whether each non-correct response stemmed from the student not knowing the skill, or from knowing the skill but slipping.
- We use machine learning to identify features of an action that characterize whether that action was a guess or a slip. These features do not use any information from subsequent actions; hence, they can be used to predict whether an action is a guess or a slip immediately after it occurs.

- In modeling student problem-solving, we use the machine learned models to dynamically estimate the probability that a response is a guess or a slip. We employ these dynamic performance estimates in the Bayesian Knowledge Tracing algorithm to update the probability that the student knows the skill.

The first step is to label a set of existing student actions with the probability that these actions involve guessing or slipping, to serve as inputs to a machine learning algorithm. The set of student actions to be labeled is drawn from the set of first actions on the 64 skills for which the Dirichlet Priors model is not theoretically degenerate. We chose to use skills that are not theoretical degenerate to avoid training our models to include model degeneracy, and used Dirichlet Priors in order to avoid creating an equivalence class of potential models (i.e. the identifiability problem). We then use estimates from this model in order to create the contextual guess and slip models.

We label student actions (***N***) with the probability that they represented a guess or slip, using information about the two actions afterwards (***N+1***, ***N+2***). Using information about future actions gives considerable information about the true probability that a student's action at time ***N*** was due to knowing the skill – if actions ***N, N+1***, and ***N+2*** are all correct, it is relatively unlikely that ***N***'s correctness was due to guessing.

The probability that the student guessed or slipped at time ***N*** (i.e., the action at time ***N,*** which we term $\boldsymbol{A_n}$) is directly obtainable from the probability that the student knew the skill at time ***N***, given knowledge about the action's correctness:

$P(A_n \text{ is guess} \mid A_n \text{ is correct}) = 1 - P(L_n)$ $\qquad$ $P(A_n \text{ is slip} \mid A_n \text{ is incorrect}) = P(L_n)$

We can calculate the probability that the student knew the skill at time ***N***, given information about the actions at time ***N+1*** and ***N+2*** (which we term $\boldsymbol{A_{+1+2}}$). We do so by using Bayes' Rule to combine 1) the probability of the actions at time ***N+1*** and ***N+2*** ($\boldsymbol{A_{+1+2}}$), given the probability that the student knew the skill at time ***N*** ($\mathbf{L_n}$); 2) the prior probability that the student knew the skill at time ***N*** ($\mathbf{L_n}$); and 3) the initial probability of the actions at time ***N+1*** and ***N+2*** ($\boldsymbol{A_{+1+2}}$).

In equation form, this gives: $P(L_n \mid A_{+1+2}) = \frac{P(A_{+1+2} \mid L_n) * P(L_n)}{P(A_{+1+2})}$

The probability of the actions at time ***N+1*** and ***N+2*** is computed as

$$P(A_{+1+2}) = P(L_n) * P(A_{+1+2} \mid L_n) + (1 - P(L_n)) * P(A_{+1+2} \mid \sim L_n)$$

The probability of the actions at time ***N+1*** and ***N+2,*** in the case that the student knew the skill at time ***N*** ($\mathbf{L_n}$)***,*** is a function of the probability that the student guessed or slipped at each opportunity to practice the skill. C denotes a correct action; ~C denotes an incorrect action (an error or help request).

$P(A_{+1+2} = C, C \mid L_n) = P(\sim S)^2$ $\qquad$ $P(A_{+1+2} = C, \sim C \mid L_n) = P(G)P(\sim S)$

$P(A_{+1+2} = \sim C, C \mid L_n) = P(G)P(\sim S)$ $\qquad$ $P(A_{+1+2} = \sim C, \sim C \mid L_n) = P(G)^2$

The probability of the actions at time ***N+1*** and ***N+2,*** in the case that the student did not know the skill at time ***N*** ($\mathbf{L_n}$)***,*** is a function of the probability that the student

learned the skill between actions ***N*** and ***N+1***, the probability that the student learned the skill between actions ***N+1*** and ***N+2***, and the probability of a guess or slip.

$$P(A_{+1+2} = C, C| \sim L_n) = P(T)P(\sim S)^2 + P(\sim T)P(T)P(G)P(\sim S) + P(\sim T)^2 P(G)^2$$
$$P(A_{+1+2} = C, \sim C| \sim L_n) = P(T)P(\sim S)P(S) + P(\sim T)P(T)P(G)\big(P(S)\big) + P(\sim T)^2 P(G)P(\sim G)$$
$$P(A_{+1+2} = \sim C, C| \sim L_n) = P(T)P(S)P(\sim S) + P(\sim T)P(T)P(\sim G)P(\sim S) + P(\sim T)^2 P(\sim G)P(G))$$
$$P(A_{+1+2} = \sim C, \sim C| \sim L_n) = P(T)P(S)^2 + P(\sim T)P(T)P(\sim G)P(S) + P(\sim T)^2 P(\sim G)^2$$

Once the actions are labeled with estimates of whether they were guesses or slips, we use these labels to create machine-learned models that can accurately predict at run-time whether a given action is a guess or slip. The original labels were developed using future knowledge, but the machine-learned models predict guessing and slipping using only data about the action itself (no future data).

For each action, we distilled a set of 23 features describing that action; the features used in the final models are shown in Table 2. We then used Linear Regression, within Weka [11], to create 2 models predicting the probability of guessing or slipping. Linear Regression gave slightly better performance under 10-fold cross-validation than a Support Vector Machine or Multilayer Perceptron – r=0.44 within the guess model, and r=0.38 within the slip model.

Then, when we have models that can predict the probability that any action was a guess or a slip, we can label the first action of each opportunity to use a skill with predictions as to how likely it is to be a guess and slip. Then, parameter values can be fit for P(**T**) and P($\mathbf{L_0}$), for each skill, using curve-fitting. At this point, we have a model that makes predictions about student knowledge each time they attempt to use a skill for the first time on a given problem step. This model also involves considerably fewer parameters than previous models – whereas all three prior models had exactly 4 parameters per skill, this model has 2 parameters fit per skill, and 27 parameters fit across all skills, for an average of 2.11 parameters per skill.

Table 2. The machine learned models of guessing (left) and slipping (right). In the unusual case where output values fall outside the range {0,1}, they are bounded to 0 or 1.

Feature	**P(G)=**	**P(S)=**
Action is a help request		+ 0.066
Percent of past opportunities where student has requested help on this skill		- 0.047
Percent of past opportunities where student has made errors on this skill		- 0.004
Response is a string	+ 0.049	- 0.02
Time taken	+ 0.002	- 0.0002
Time taken (SD faster (-) or slower (+) than average across all students)	- 0.024	+ 0.01
Time taken in last 5 actions (calculated in SD off average across students)	- 0.003	+ 0.002
Total number of times student has gotten this skill wrong on the first try	- 0.002	+ 0.0002
Total time taken on this skill so far (across all problems)	+ 0.001	- 0.001
Number of last 5 actions which involved same interface element	+ 0.014	- 0.026
Number of last 8 actions which involved help request	+ 0.042	- 0.019
Number of last 5 actions which were wrong	+ 0.036	- 0.033
At least 3 of last 5 actions involved same interface element & were wrong	+ 0.067	+ 0.013
Number of opportunities student has already had to use current skill	+ 0.003	- 0.001
Constant term	+ 0.066	+ 0.442

4 Evaluating the Contextual Guess and Slip Model

Are models created by the contextual guess and slip method identifiable? The initial skill models used to create the labels for the linear regression process were generated by the Dirichlet priors method, and thus represent an optimal and unique parameter set for that method [2]. Linear Regression itself has a single optimal solution [3]. Finally, with only two parameters to fit, the contextual guess and slip model of each skill has a unique best solution for each pair of parameters P(**T**) and P($\mathbf{L_0}$). Hence, since each step in the model fitting process has a single optimal solution, the contextual guess and slip method is as identifiable as the Dirichlet Priors method.

What about model degeneracy? We can test for the two types of empirical model degeneracy using the student log data. A model fails the first test of empirical model degeneracy when a student gets the first three actions correct on a specific skill but then has lower P(**L**) afterwards. There were 2558 cases in the data where a student got the first three actions correct on a specific skill; in only 1 of the 2558 cases did the student have a lower P(**L**) afterwards – and, in that case, the student got the skill incorrect on the next 7 opportunities. The proportion of failure of the first test (1/2558 = 0.0004%) is significantly lower than the proportion of failure of this test in the baseline or Dirichlet Priors models, in each case t(251)= -2.00, two-tailed p=0.05, for a paired t-test (comparing model performance within each skill), but is not significantly higher than the proportion of failure of the first test for the bounded model, t(251)= 1.00, two-tailed p=0.32.

A model fails the second test of empirical degeneracy if a student gets ten actions correct in a row but does not reach mastery. There were 758 cases in the data where a student got the first ten actions correct on a specific skill; in 13 of the 758 cases the student afterwards had a P(**L**) below mastery (0.95). This proportion (1.7%) is significantly lower than the baseline, Dirichlet Priors, and bounded models, respectively, t(251)= -8.42, t(251)= -8.42, t(251) = -3.37, in all three cases two-tailed p<0.001.

Hence, there is evidence for limited degeneracy in the Contextual Guess and Slip model, but this model is substantially less degenerate than the baseline or Dirichlet Priors models, and appears to be less degenerate than the bounded model as well.

There are two ways to measure the accuracy of the four knowledge tracing models. The first is to compare actions at time ***N*** to the models' predictions of the probability that actions at time ***N*** will be correct – P($\mathbf{L_n}$)*P(**~S**)+ P(~$\mathbf{L_n}$)*P(**G**). This method accurately represents exactly what each model predicts; however, this method biases in favor of the Contextual Guess and Slip model, since that model uses information associated with the answer being predicted to estimate the probability of guessing and slipping. An alternate measure which is appropriate for all four models

Table 3. Each model's accuracy across the 171,989 first actions. Comparisons use model prediction of knowledge state after previous attempt at skill. Standard errors given in parentheses.

Modeling Approach	A'	r
Baseline	0.66 (0.001)	0.29
Bounded Guess and Slip (Corbett's method)	0.61 (0.001)	0.25
Dirichlet Priors (Beck's method)	0.65 (0.001)	0.26
Contextual Guess and Slip	0.75 (0.001)	0.43

is to compare actions at time ***N*** to the models' predictions of the probability that the student knew the skill at time ***N***, before the student answered. This method underestimates accuracy for all models (since it does not include the probability of guessing and slipping when answering), but is preferable because it does not favor any model. We use A' (the probability that the model can distinguish a correct response from an incorrect response [8]) and correlation as the measures of model accuracy.

The full pattern of results is shown in Table 3. The Contextual Guess and Slip method achieves the highest value of A', 0.75. The second-best model is the Baseline model, with A' of 0.66. The Contextual Guess and Slip model's A' achieves 27% of the possible improvement over the Baseline model, a statistically significant difference in fit, $Z=2.86$, two-tailed $p<0.01$ (an adjusted standard error is used for A' to control for type II error stemming from non-independence). The Contextual Guess and Slip method also achieves the highest correlation, 0.43, 48% higher than the second-best model, again the Baseline model ($r=0.29$), $t(171984)=69.12$, two tailed $p<0.0001$. (This test is under-conservative, as it assumes independence; if we collapse across students, an overly conservative test, the result remains significant, $t(231)=7.95$, two tailed $p<0.0001$). Hence, for both measures of model accuracy, Contextual Guess and Slip performs substantially better than prior knowledge tracing methods.

5 Conclusions

In this paper, we have proposed a new way to contextually estimate the probability that a student obtained a correct answer by guessing, or an incorrect answer by slipping, within Bayesian Knowledge Tracing. The method we propose is less vulnerable to model degeneracy than previous methods of student knowledge modeling, and is as good as the best of previous approaches at dealing with challenges to identifiability. In addition, our method leads to substantially higher accuracy than prior methods – improving A' by 27% of potential gain, and improving correlation by 48%. Plus, the method seems quite generalizable; the machine-learned models of guess and slip was trained on only 64 skills, but functioned effectively within all 253 skills it was tested on. An interesting area for future research will be studying how widely the guess and slip models can be transferred with no re-training at all, and still function effectively. Similarly, it will be important to replicate this result in data from another Cognitive Tutor, and in other intelligent tutors [cf. 5].

Though the contextual estimation of guess and slip has proven more successful than earlier student knowledge modeling, we see this paper as just the beginning of a new, more contextually sensitive approach to student modeling. First of all, it is probably possible to increase the accuracy of the contextual estimates of slip even further, by incorporating data about second and subsequent attempts within a given opportunity to practice a skill (an error followed very rapidly by the correct answer is probably much more likely to be a slip than an error followed by three more slow errors and a help request). Second, it is possible to combine overall estimation of the probability of guesses and slips (as used here) with information about individual skills, potentially raising accuracy further still. Third, it is possible to estimate the probability of learn-

ing a skill P(**T**) at any given time in the same contextual fashion as used here. We look forward to new possibilities for substantially more sensitive and accurate estimation of student knowledge. And, in the long term, more sensitive and accurate estimation of student knowledge will have considerable pay-offs: it will enable more accurate assignment of learning materials to students, optimize the amount of practice on each skill [cf. 4], and may even enable different types of remediation for different types of errors (such as giving specific remediation for slips).

Acknowledgements

We would like to thank Project LISTEN and Joseph Beck for offering the BNT-SM toolkit which made this comparison of models feasible. This work was funded by NSF grant REC-043779 to "IERI: Learning-Oriented Dialogs in Cognitive Tutors: Toward a Scalable Solution to Performance Orientation", and by the Pittsburgh Science of Learning Center, National Science Foundation award SBE-0354420.

References

1. Beck, J. (2007) Difficulties in inferring student knowledge from observations (and why you should care). *Educational Data Mining: Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education,* 21-30.
2. Beck, J.E., Chang, K.-m. (2007) Identifiability: A Fundamental Problem of Student Modeling. *Proceedings of the 11th International Conference on User Modeling (UM 2007).*
3. Boyd, S., Vandenberghe, L. (2004). *Convex Optimization.* Cambridge, UK: Cambridge University Press.
4. Cen, H., Koedinger, K.R., Junker, B. (2007). Is Over Practice Necessary? Improving Learning Efficiency with the Cognitive Tutor. *Proceedings of the 13th International Conference on Artificial Intelligence and Education.*
5. Chang, K., Beck, J., Mostow, J., Corbett, A.T. (2006) Does Help Help? A Bayes Net Approach to Modeling Tutor Interventions. *Proceedings of the AAAI2006 Workshop on Educational Data Mining*.
6. Corbett, A.T., Anderson, J.R. (1995) Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction, 4,* 253-278.
7. Ferguson, G.A. (1971) *Statistical Analysis in Psychology and Education*. New York: McGraw-Hill.
8. Hanley, J.A., McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology,* 143, 29-36.
9. Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. *Proceedings of PME-NA XXXIII (the North American Chapter of the International Group for the Psychology of Mathematics Education).*
10. Reye, J. (2004) Student Modeling based on Belief Networks. *International Journal of Artificial Intelligence in Education, 14,* 1-33.
11. Witten, I.H., Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques.* San Francisco, CA: Morgan Kaufmann.