Improving Models of Slipping, Guessing, And Moment-By-Moment Learning with Estimates of Skill Difficulty

SUJITH M. GOWDA
Worcester Polytechnic Institute
JONATHAN P. ROWE
North Carolina State University
RYAN S.J.d. BAKER
Worcester Polytechnic Institute
MIN CHI
Stanford University
AND
KENNETH R. KOEDINGER
Carnegie Mellon University

Over the past several years, several extensions to Bayesian knowledge tracing have been proposed in order to improve predictions of students' in-tutor and post-test performance. One such extension is Contextual Guess and Slip, which incorporates machine-learned models of students' guess and slip behaviors in order to enhance the overall model's predictive performance [Baker et al. 2008a]. Similar machine learning approaches have been introduced in order to detect specific problem-solving steps during which students most likely learned particular skills [Baker, Goldstein, and Heffernan in press]. However, one important class of features that have not been considered in machine learning models used in these two techniques is metrics of item and skill difficulty, a key type of feature in other assessment frameworks [e.g Hambleton, Swaminathan, & Rogers, 1991; Pavlik, Cen, & Koedinger 2009]. In this paper, a set of engineered features that quantify skill difficulty and related skill-level constructs are investigated in terms of their ability to improve models of guessing, slipping, and detecting moment-by-moment learning. Supervised machine learning models that have been trained using the new skill-difficulty features are compared to models from the original contextual guess and slip and moment-by-moment learning detector work. This includes performance comparisons for predicting students' in-tutor responses, as well as post-test responses, for a pair of Cognitive Tutor data sets.

Categories and Subject Descriptors: I 2.7 **[Artificial Intelligence]**

General Terms: Educational Data Mining, Cognitive Tutor, Feature Engineering, Post-test prediction, Moment-by-Moment Learning, Contextual Guess, Contextual Slip, Bayesian Knowledge-Tracing

Additional Key Words and Phrases:

Authors' addresses: Sujith M. Gowda, Ryan S.J.d. Baker, Department of Social Science and Policy Studies, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA USA 01609. E-mail: sujithmg@wpi.edu, rsbaker@wpi.edu; Jonathan P. Rowe, Department of Computer Science, North Carolina State University, 890 Oval Drive, Raleigh, NC 27695. E-mail: jprowe@ncsu.edu; Min Chi, Developmental and Psychological Sciences, Stanford University, 485 Lasuen Mall, Stanford, CA 94305, E-mail: minchi@stanford.edu; Kenneth R. Koedinger, Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh PA USA 15213; E-mail : koedinger@cmu.edu

1. INTRODUCTION

Bayesian knowledge tracing (BKT) is a well-established approach for modeling student knowledge during interactions with an intelligent tutoring system. First deployed by Corbett and Anderson [1995] in the ACT Programming Tutor, BKT has since been applied successfully across a range of academic subjects and student populations, including elementary reading [Beck and Chang 2007], middle school mathematics [Koedinger 2002], and college-level genetics [Corbett et al. 2010]. In the approach, a two-node dynamic Bayesian network (DBN) is associated with each skill, and it is used to monitor student knowledge and problem-solving actions associated with that skill. The relationship between skills and steps during problem-solving is typically defined by a running cognitive model, although other forms of skill-item mappings can also be used. The probability values in BKT are defined using four parameters: p(T) is the Transition parameter, which is the probability of learning a skill immediately after an opportunity to apply it; p(L_0) is the Initial Learning parameter, which is the probability of knowing a skill prior to the first opportunity to apply it; p(G) is the guess parameter, which is the probability of providing a correct response despite not knowing an associated skill; and p(S) is the Slip parameter, which is the probability of providing an incorrect response despite knowing an associated skill. This concept of guess and slip is also seen in DINA IRT models [De La Torre, 2004]. The parameter values are typically machine-learned from data collected during prior students' tutorial interactions.

Several attempts have been made to extend BKT in order to relax some of its assumptions about learning and performance, and improve predictive power. Corbett and Anderson [1995] proposed a technique that uses individualized weights to adjust the model's four parameters for each student. A key limitation of this method is that it cannot be used at run-time, as optimization is conducted after all data is obtained. Recent work by Pardos and Heffernan introduced the Prior Per Student individualization technique, which slightly modifies the standard two-node DBN structure [Pardos and Heffernan 2010]. In the technique, an individualization node is added with states for each student, and it is connected to nodes in the model to be individualized. Unlike Corbett and Anderson's technique, this approach uses data from the first action to individualize student priors, and can thus be used in a running tutor. An empirical evaluation found the Prior Per Student extension to be significantly more accurate in predicting student responses during tutoring sessions with the ASSISTment system. While both of these approaches improved upon the predictive accuracy of standard BKT, neither approach addressed the assumption that knowledge-tracing parameters are fixed over the course of a tutoring session.

An alternate extension to BKT is Contextual Guess and Slip, which focuses on relaxing the assumption that guess and slip probabilities are fixed [Baker, Corbett and Aleven 2008]. Contextual models of guessing and slipping leverage information about the current tutorial state in order to dynamically adjust the guess and slip parameters associated with each opportunity to apply a skill. Baker and colleagues [2008a] have demonstrated that contextual models of guessing and slipping improve the fit of knowledge tracing models to existing tutoring data, although later results contradict that finding [Baker et al. 2010]. However, Baker and colleagues [2010] have found that estimates of contextual slip, used in conjunction with standard Bayesian Knowledge-Tracing estimates of student knowledge, can be used to improve predictions of students' post-test performance outside of a tutor. A similar approach involving Bayesian analysis and supervised machine learning has been devised to detect how much learning occurs in each problem step (termed moment-by-moment learning) [Baker, Goldstein and Heffernan in press]. Accurately detecting the amount of learning, moment-by-moment, may enable intelligent tutoring systems to tailor practice opportunities more precisely, as well as shed light on the differences in learning rates between skills [e.g. Baker, Goldstein and Heffernan in press].

An integral step in constructing contextual models of guessing and slipping, as well as moment-by-moment learning detectors, is selecting appropriate predictor features for supervised machine learning. Baker and colleagues originally used a set of twenty three features to train linear regression models to predict guess and slip probabilities [Baker et al. 2008a], drawn in turn from earlier work detecting gaming the system [Baker et al. 2008b]. One important class of features that was not included in the original work is metrics of skill difficulty. This class of features is an important component in alternative student modeling techniques, such as Item Response Theory (IRT). In Item Response Theory, each test question is associated with an item characteristic curve, which estimates the probability that a student with a given ability will answer the question correctly [Baker 2001; Hambleton, Swaminathan and Rogers 1991]. A key parameter of IRT is question difficulty, which quantifies the level of ability necessary to answer the question correctly. Leveraging features that quantify skill and item difficulty is a promising direction for enhancing the performance of contextual models of guessing and slipping, as well as moment-by-moment learning detectors. For example, in the 2010 KDD Cup, an annual Data Mining and Knowledge Discovery competition where data mining teams across the world compete to solve a practical data mining problem, the goal was to predict student performance within a Cognitive Tutor. Several of the top performers in the 2010 KDD Cup leveraged skill difficulty features in their solutions [Pardos and Heffernan 2010b; Shen et al 2010; Yu et al. 2010]. Skill difficulty is also a component of Performance Factors Analysis, a competing approach to assessing student proficiency in intelligent tutors [Pavlik et al. 2009].

Within this paper, we investigate the addition of several skill-difficulty features for training models of guessing and slipping, as well as detecting moment-by-moment learning. We describe seven features that quantify different aspects of skill difficulty, and report findings about their impacts on the goodness of the resultant machine-learned models. Models that incorporate the new skill-difficulty features are compared to models that use only the original set of predictor features. We investigate whether the new models lead to improved predictions of students' guesses and slips, moment-by-moment learning, and post-test performance.

This paper is organized as follows. Additional background on contextual models of guessing and slipping is provided and details about detecting moment-by-moment learning are discussed. Afterwards, the data and method used in the current investigation are described. Results are presented about improvements in contextual models of guessing and slipping, moment-by-moment learning detectors, and predictions of post-test performance. The paper concludes with a discussion of implications for practice and future directions.

1.1 Contextual Models of Slipping and Guessing

The Contextual Guess and Slip (CGS) model of student knowledge is an extension of Corbett and Anderson's [1995] BKT model, proposed by Baker, Corbett and Aleven [2008]. Unlike Corbett and Anderson's approach, the CGS model estimates whether each individual student response is a guess, denoted P(G), or a slip, denoted P(S), based on contextual information. This is in contrast to approaches that utilize fixed guess and slip probability estimates throughout tutoring [e.g. Corbett & Anderson 1995].

In the BKT-CGS model, each skill has a separate parameter for Initial Knowledge and Transition, as in standard BKT. However, unlike in standard BKT, guess and slip probabilities are not estimated for each skill. Instead, they are computed each time a student attempts to answer a new problem step in the tutor, and they are based on machine-learned models of guessing and slipping behaviors within the current tutorial context (for example, longer responses and help requests are less likely to be slips). The Contextual Guess and Slip procedure involves five stages, and it proceeds as follows. First, a four-parameter model is obtained using the brute force variant of Bayesian knowledge tracing [Baker et al 2010]. Second, the four-parameter model is used to assign slip and guess probability labels to each action in the data set. Third, supervised machine learning is used to obtain models that predict the slip and guess labels. Fourth, the machine-learned models of guessing and slipping are incorporated into the BKT models in lieu of skill-by-skill labels for guessing and slipping. Last, parameters for Initial Knowledge, denoted $P(L_0)$, and Transition, denoted P(T), are fit. Additional details about this approach are available in [Baker et al. 2008a]. Baker et al. [2010] have found that two aggregations of contextual slip are significant predictors of post-test performance. While average contextual slip is a good predictor of post-test performance, the "certainty of slip" – the average contextual slip among actions thought to be slips (e.g. contextual slip > 0.5) – is an even stronger predictor of post-test performance.

1.2 Detecting Moment-by-Moment learning

Baker, Goldstein and Heffernan [in press] presents a model that predicts the probability that a student has learned a specific knowledge component at a specific problem step, termed P(J). The underlying architecture of this model is similar to the Contextual Guess and Slip model. First, training labels to detect moment-by-moment learning are generated for each problem step in a tutor data set. The labels are generated by applying Bayes' Rule to a combination of knowledge estimates from a traditional BKT model, as well as information about the correctness of two future problem-solving actions. Next, a set of predictor features is generated using past tutor data to form a training data set. Finally, these predictor features are used during supervised machine learning to train a model that predicts the moment-by-moment learning labels. The machine-learned model computes a learning probability for each problem-solving step using no data from the future. A recent investigation observed that a distillation of the variance in this model, termed "spikiness", is a good predictor of students' eventual learning within the tutor, as assessed by BKT. Additionally, spikiness aids in understanding the differences between gradual learning and learning via "eureka" moments, where a KC is understood suddenly [Baker, Goldstein, and Heffernan in press].

2. METHOD

The current work investigates whether contextual models of slipping, guessing, and moment-by-moment learning can be improved by incorporating skill-difficulty features in the machine-learned models. Data from two Intelligent Tutoring Systems, The Middle School Mathematics Cognitive Tutor [Koedinger 2002] and the Genetics Cognitive Tutor [Corbett et al. 2010], are considered. The Middle School Cognitive tutor, a precursor to the current curriculum Bridge to Algebra, covers a wide span of Mathematics topics in middle school mathematics. Topics covered by the Middle School Cognitive Tutor include fraction concepts, areas and perimeters with decimals, and simple histograms. The Middle School data set consists of an entire year's use of an intelligent tutor in suburban schools in the Northeastern USA. Within the data set, actions that were not labeled with skills were excluded, because skill information is necessary for the application of Bayesian knowledge tracing techniques.

The Genetics data set comes from the Genetics Cognitive Tutor [Corbett et al. 2010], which consists of 19 modules that support problem solving across a wide range of topics in genetics (Mendelian transmission, pedigree analysis, gene mapping, gene regulation and population genetics). Various subsets of the 19 modules have been piloted at 15 universities in North America. In the study that generated the current data set, 70 undergraduate students enrolled in a genetics course at Carnegie Mellon University used the three-factor cross module as a homework assignment. Students completed a paper-and-pencil problem-solving pre-test and post-test consisting of two problems. There were two test forms, and students were randomly selected to receive one version at pre-test and the other version at post-test in order to counterbalance test difficulty. Each problem on the tests consisted of 11 steps involving 7 of the 8 skills in the Three-Factor Cross tutor lesson, with two skills applied twice in each problem and one skill applied three times. Note that this data set is the same one used in [Baker et al. 2010]. The size of each data set is shown in Table 1.

Table 1. The size of each data set (after exclusion of actions not labeled with skills)

Data set	**Actions**	**Problem Steps**	**Skills**	**Students**
Middle school	581,785	171,987	253	232
Genetics	19,150	9,259	8	71

In the original work on contextual models of guessing and slipping [Baker, Corbett, and Aleven, 2008] and detecting moment-by-moment learning [Baker, Goldstein, and Heffernan 2010], four primary categories of predictor features were used during the supervised machine learning stages of the procedures. The original features date back to the development "gaming the system" detectors for Cognitive Tutors [Baker et al. 2008b]. The four categories include:

- **Action Correctness.** This category includes features that characterize whether a problem-solving action is correct, incorrect, constitutes a known misconception, or a help request.
- **Step Interface Type.** This category includes features that are based on the type of interface widget involved in the problem-solving action. For example, a particular problem step may involve typing a string or specifying a number.
- **Response Times**. This category includes features that are derived from the amount of time taken to complete problem-solving steps. Example features include the amount of time (seconds) taken on a particular step, time-taken expressed as standard deviations from the mean, and total time spent during the last three problem-solving actions.
- **Problem-Solving History.** This category includes features that characterize the student's problem-solving history in the tutor. Examples include: the number of past five actions that were involved in the current step, and the number of times that the student asked for help in the past eight problem-solving actions.

Within this paper, we add seven new predictor features that quantify key aspects of the skill associated with each problem-solving step. As discussed in the introduction, the new features are inspired by Item Response Theory concepts of skill difficulty, which were prominent in many of the successful entries in the KDD cup. The features are calculated by determining average metric values using other students' problem-solving histories; this procedure is cross-validated at the student level during evaluation of this approach (see below for details). The newly engineered features include the following:

- Features based on item difficulty:
  In Cognitive Tutors, a student's action is represented using four predefined categories:
  - Right – when the user inputs correct answer

  - Help - when the user requests help
  - Bug - when the tutor gives a bug message, indicating a misconception
  - Wrong - when the user inputs incorrect answer

  We assessed average performance among all students for each of these categories: “Average-right”, “Average-bug” and “Average-help”. These features were calculated by computing the proportion of each of category across all the students for a given skill.
- Features based on user input:
  In the data distillation, a user’s input type is categorized as either a number or a string. Based on user input information we derived two features “Average-number” and “Average-string” by computing the proportion of "inputs which are a number" and "inputs which are a string" respectively across all the students for a given skill. Students will have a greater proportion of actions of one type, relative to other students completing the same curriculum, when they have more difficulty with that type of action. Consequently, this feature indirectly measures the relative difficulty of skill groups.
- Feature based on time:
  The Cognitive tutor records the time taken by the user to answer a step. Using this time we derived the feature “Average-time” by calculating the average time taken to answer problem steps across all students for a given skill. Students who take longer can, in the aggregate, are assumed to be having greater difficulty.
- Feature based on number of opportunities to practice a skill:
  This feature is a distillation of the feature “optoprac” from [Baker et al. 2008b], which computes the number of times that the same skill has been seen by the student prior to this action. Using optoprac, the feature “Average-optoprac” was calculated by aggregating across actions and students for each skill.

The new set of predictor features provides several aggregate assessments of the skills involved in the problem-solving process. These new features are considered for machine learning contextual models of guessing and slipping behaviors, as well as models for detecting moment-by-moment learning. The next section describes the supervised machine learning analysis that has been conducted to investigate the impact of including these features, and the results in terms of model goodness.

3. RESULTS

In discussing the results, we will first present the linear regression analyses we conducted to develop each type of model, and then discuss improvements in the overall model from incorporating the new features. We will discuss model improvements in terms of fit to training labels, as well as external measures such as predicting post-test performance. We use linear regression for consistency with the original analyses [Baker, Corbett, and Aleven 2008; Baker, Goldstein, and Heffernan, in press], where linear regression was found to lead to acceptable cross-validated performance.

3.1 Model Development

Using both the original features and new features, we used RapidMiner [Mierswa, et. al. 2006] to develop linear regression models, using default settings (e.g. M5-prime feature selection). Leave-one-out cross-validation (LOOCV) was conducted, at the student level, to evaluate the models. By cross-validating at the student level rather than the action level, we can have greater confidence that the resultant models will generalize to new groups of students. The LOOCV procedure is as follows: Before the generation of folds, labels were generated for contextual slip and guess and for moment-by-moment learning using the same labeling process as in [Baker et al. 2008a; Baker, Goldstein and Heffernan in press]. Special care needed to be taken to ensure that skill-level features did not violate independence during cross-validation. Values for all of the skill-level features were calculated using only data from the training set. After calculating the skill-level feature values for each fold, the features were added to both the training set and the test set (e.g. features generated on training set data were used within the test set). This corresponds to what would occur in real-world applications of this approach, where skill-level features might be generated on data from one cohort and applied to students in a later cohort.

For each training set, three linear regression models were machine-learned: one for guessing, one for slipping, and one for detecting learning moment-by-moment. By applying these models to the corresponding test set, model predictions for guess, slip and P(J) were obtained. Correlations were next calculated between the labels and guess, slip, and P(J) model predictions for each student and then averaged across the students.

After obtaining correlation values for the new models, the same cross-validation procedure was applied to obtain training and test sets that included only the original predictor features. These models did not include any of the new skill-difficulty features, and were intended to serve as a comparison group. The statistical significance of the difference between the new feature model and the old feature model was computed by computing the correlation coefficient for each fold (student), converting the correlation coefficients to Z values, and then aggregating across folds (students) using Stouffer's Z.

The cross-validated correlation of each model to the training labels is shown in Table 2. As can be seen, all models involving the new features achieve statistically significantly better fit to the training labels in the test folds, at the $p < 0.001$ level, than the models that use only the original feature set. These significant differences are observed for all three types of models—guessing, slipping, and detecting moment-by-moment learning—and they are observed in both the Middle School and Genetics data sets.

In the Middle School data set, the slip model using the new features achieves a 32% improvement over the model with the original features, the guess model achieves a 46% improvement, and the P(J) model achieves a 17% improvement. In the Genetics data set, the slip model with the new features achieves a 22% improvement over the original model. The P(J) model with the new features achieves a 3% improvement over the model with the original features. The guess model with the new features achieves a very large improvement in correlation, from 0.029 to 0.422. Of the seven newly engineered features, five are found in all three regression models in both data sets: Average-right, average-help, average-string, average-time and average-optoprac. Of the three types of models, the strong performance for the slip models in both data sets is of particular interest, because slipping within a Cognitive Tutor has been shown to be a significant predictor of post-test performance [Baker et al. 2010], even after controlling for student knowledge.

Table 2: Correlation coefficients between labels and predictions

Data-Set	**Model**	**OldFeatures-*r***	**NewFeatures-*r***	**Z**	***P***
Middle School	Slip	0.322	0.424	23.91	< 0.001
	Guess	0.112	0.163	5.42	< 0.001
	P(J)	0.476	0.558	84.16	< 0.001
Genetics	Slip	0.456	0.582	6.73	< 0.001
	Guess	0.029	0.422	33.92	< 0.001
	P(J)	0.764	0.790	21.09	< 0.001

3.2 Post-test Prediction Using Final Knowledge and Contextual Slip Estimates

Baker, et al. [2010] showed that a linear combination of average contextual slip over 0.5 (termed "certainty of slip") and BKT estimates of post-test performance predicts students' post-tests statistically significantly better than BKT estimates alone. We replicate this analysis using the same Genetics data set and the new models with skill-difficulty features. We also analyze regression models combining average slip (average of Contextual slip for each student across all the actions) and BKT. The analysis is not replicated for the Middle School data, because that data set lacks post-test scores. Unlike Baker et al. [2010], we use cross-validated values of contextual slip in this analysis. By using the cross-validated contextual slip values, we can have greater confidence that slip models can be generalized for new data sets.

Several models were compared in terms of correlations with students' post-test performance after interacting with the Genetics Tutor. The four-parameter brute force variant of Bayesian Knowledge Tracing (BKT-BF) model achieves a cross-validated correlation of r = 0.426 to students' post-tests, which is statistically significantly better than chance, F(1,69) = 15.27, p < 0.001. For the model generated using the original features, we find that a combination of cross-validated certainty of slip (average of slip above 0.5) and BKT-BF estimates achieves a cross-validated correlation of 0.437. The certainty of slip term in this model is not statistically significant, when added to a model containing BKT-BF, t(68) = 0.92, p = 0.36, for a two-tailed test. When the same model is generated using the new feature set, the model achieves a cross-validated correlation of 0.448. Again, the certainty of slip term in the combined model is not statistically significant, t(68) = 1.29, p=0.20.

However, when average slip is taken rather than certainty of slip, a different pattern emerges. The original feature version of average slip, when combined with BKT-BF, achieves a cross-validated correlation of 0.479. In this model, the average slip term is statistically significant, t(68)=2.06, p=0.04, signifying that the model containing average slip is significantly better than the model containing BKT-BF alone. The new feature version of average slip, when combined with BKT-BF, achieves a cross-validated correlation of 0.485. In this model, the average slip term is also statistically significant, t(68)=2.19, p=0.03.

The difference between the four models combining variants on contextual slip can be compared using BiC', the Bayesian Information Criterion for Linear Regression Models (Raftery, 1995). Within these four models, the model containing the new feature version of average slip achieves the best BiC', but does not perform significantly better than any of the other three models (differences of 6 between model values for BiC' are considered equivalent to statistical significance at the p<0.05 level).

Hence, the new features significantly improve cross-validated fit to the test-fold training labels for each of these models, but do not appear to significantly improve post-test prediction, although there is some trend in that direction. Nonetheless, as these model fits are cross-validated, there is evidence for improvement (as conducting significance tests or computing BiC' on cross-validated data is doubly-stringent).

Table 3. Cross-validated correlations between post-test and t-test scores of 2nd parameter using regression analysis and BiC' values

Model	**Correlation**	**t-test of 2nd Param**	**P-value of 2nd Param**	**BiC'**
BKT-BF-Predictions Only	0.426			-9.763
BKT-BF-Preds + Old_Certainty_Slip	0.437	t(68)= -0.920	0.361	-6.333
BKT-BF-Preds + New_Certainty_Slip	0.448	t(68)= -1.285	0.203	-7.180
BKT-BF-Preds + Old_Avg_Slip	0.479	t(68)= -2.059	0.043	-9.742
BKT-BF-Preds + New_Avg_Slip	0.485	t(68)= -2.186	0.032	-10.269

4. CONCLUSION

In this paper, we present a new set of features related to metrics of skill difficulty, which when combined with original features [Baker et al 2008a], have been shown to statistically significantly improve the predictive capabilities of guessing, slipping and moment-by-moment learning models. The newly engineered features are inspired by Item Response Theory concepts of item difficulty, and they aggregate empirical estimates of skill performance, types of user input, problem-solving action time, and average practice opportunities for each skill. We find that guess, slip, and moment-by-moment learning models that use the new skill-difficulty features outperform the original-feature models under cross validation for two Cognitive Tutor data sets. Therefore, the findings suggest that the new features can be computed using previous years' class data, and they can then be directly incorporated into models for new students.

Of the three models that were investigated, the slip model is of particular interest because certainty of slip has been shown to be a significant predictor of post-test performance [Baker et al. 2010]. In this paper, we replicated the tests in that paper, using cross-validation. We found that a model predicting the post-test using both average contextual slip and Bayesian Knowledge Tracing had significantly better goodness than a model using Bayesian Knowledge Tracing alone, even when cross-validating the data. However, the evidence for improvement stemming from the new features appeared to be weak. While the new-feature models achieved slightly better cross-validated performance than the original-feature models, the difference in BiC' values was small. One possible explanation for the small improvements in post-test prediction could be imprecision in the guess and slip labels. The new models with skill-difficulty features are performing more effectively at predicting the existing guess and slip labels, but any inaccuracies or lack of precision in these original labels could degrade post-test prediction performance. Another possible explanation could be that the contextual slip estimates may have reached a ceiling in their capacity to predict the post-test. In this case, additional features would be necessary to account for the remaining differences between post-test responses and model predictions.

There are several promising future directions for this work. First, continued exploration of new predictor features may yield further improvements in the accuracies of guess, slip, and moment-by-moment learning models. Second, further investigation is necessary to determine whether imprecision in guess and slip labels are responsible for the relatively modest gains observed in post-test prediction performance. This requires further analysis of students' guessing and slipping behaviors in order to assess the quality of guess, slip, and P(J) labels generated by Bayesian analysis procedures. Finally, additional work should be conducted to determine how enhanced guess, slip, and moment-by-moment learning models can be most effectively incorporated back into run-time Bayesian knowledge tracing models in order to improve the effectiveness of intelligent tutoring systems.

ACKNOWLEDGEMENTS

This research was supported via grant "Empirical Research: Emerging Research: Robust and Efficient Learning: Modeling and Remediating Students' Domain Knowledge", National Science Foundation award number #DRL-0910188, and by the Pittsburgh Science of Learning Center, NSF award number #SBE-0836012. We would also like to thank April Galyardt and Zachary Pardos for helpful suggestions at an early stage of this project.

REFERENCES

BAKER, F.B. 2001. *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

BAKER, R.S.J.D., CORBETT, A.T., ALEVEN, V. 2008a. More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 406-415.

BAKER, R.S.J.D., CORBETT, A.T., GOWDA, S.M., WAGNER, A.Z., MACLAREN, B.M., KAUFFMAN, L.R., MITCHELL, A.P., GIGUERE, S. 2010. Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization*, 52-63.

BAKER, R.S.J.D., CORBETT, A.T., ROLL, I., KOEDINGER, K.R. 2008b. Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18, 3, 287-31

BAKER, R.S.J.D., GOLDSTEIN, A.B., HEFFERNAN, N.T. in press, Detecting Learning Moment-by-Moment. To appear in *International Journal of Artificial Intelligence in Education.*

BECK, J. AND CHANG, K. 2007, Identifiability: A Fundamental Problem of Student Modeling. *Proceedings of the 11th International Conference on User Modeling*. 137-146.

CORBETT, A., KAUFFMAN, L., MACLAREN, B., WAGNER, A., JONES, E. 2010. A Cognitive Tutor for Genetics Problem Solving: Learning Gains and Student Modeling. *Journal of Educational Computing Research, 42, 219-239.*

CORBETT, A.T., ANDERSON, J.R. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

DE LA TORRE, J. 2004, Higher-Order Latent Trait Models for Cognitive Diagnosis. *Psychometrika*, 69, 3, 333-353.

HAMBLETON, R.K. SWAMINATHAN, H., ROGERS, H.J. 1991. *Fundamentals of Item Response Theory.* Sage Publications, Newbury Park, CA.

KOEDINGER, K.R. 2002. Toward Evidence for Instructional Principles: Examples from Cognitive Tutor Math 6. In: *Proceedings of PME-NA XXXIII (the North American Chapter of the International Group for the Psychology of Mathematics Education).*

MIERSWA, I., WURST, M., KLINKENBERG, R., SCHOLZ, M., EULER, T. 2006. YALE: Rapid Prototyping for Complex Data Mining Tasks. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2006), 935-940.

PARDOS, Z. AND HEFFERNAN, N. 2010a. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization*, 255-266.

PARDOS, Z. AND HEFFERNAN, N.T. 2010b. Using HMMs and Bagged Decision Trees to Leverage Rich Features of User and Skill from an Intelligent Tutoring System Dataset. *Proceedings of the KDD Cup 2010 Workshop held as part of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

PAVLIK, P.I., CEN, H., KOEDINGER, K.R. 2009. Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models. *Proceedings of the 2nd International Conference on Educational Data Mining,* 121-130.

RAFTERY, A.E. 1995. Bayesian Model Selection in Social Science Research. *Sociological Methodology, 28,* 111-163.

SHEN, Y., CHEN, Q., FANG, M., YANG, Q., AND WU, T. 2010 Predicting Student Performance : A Solution for the KDD Cup 2010 Challenge. *Proceedings of the KDD Cup 2010 Workshop held as part of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

YU, H.-FU, LO, H.-YI, HSIEH, H.-PING, et al. 2010 Feature Engineering and Classifier Ensemble for KDD Cup 2010. *Proceedings of the KDD Cup 2010 Workshop held as part of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.