

Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor

Ryan S.J.d. Baker¹, Albert T. Corbett², Sujith M. Gowda¹, Angela Z. Wagner²,
Benjamin A. MacLaren², Linda R. Kauffman³, Aaron P. Mitchell³, Stephen Giguere¹

¹ Department of Social Science and Policy Studies, Worcester Polytechnic Institute
100 Institute Road, Worcester MA 01609, USA
rsbaker@wpi.edu, sujithmg@wpi.edu, sgiguere@wpi.edu

² Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue,
Pittsburgh, PA 15213, USA

corbett@cmu.edu, awagner@cmu.edu, maclaren@andrew.cmu.edu

³ Department of Biological Sciences, Carnegie Mellon University, 5000 Forbes Avenue,
Pittsburgh, PA 15213, USA
lk01@andrew.cmu.edu, apm1@andrew.cmu.edu

Abstract. Intelligent tutoring systems that utilize Bayesian Knowledge Tracing have achieved the ability to accurately predict student performance not only within the intelligent tutoring system, but on paper post-tests outside of the system. Recent work has suggested that contextual estimation of student guessing and slipping leads to better prediction within the tutoring software (Baker, Corbett, & Aleven, 2008a, 2008b). However, it is not yet clear whether this new variant on knowledge tracing is effective at predicting the latent student knowledge that leads to successful post-test performance. In this paper, we compare the Contextual-Guess-and-Slip variant on Bayesian Knowledge Tracing to classical four-parameter Bayesian Knowledge Tracing and the Individual Difference Weights variant of Bayesian Knowledge Tracing (Corbett & Anderson, 1995), investigating how well each model variant predicts post-test performance. We also test other ways to utilize contextual estimation of slipping within the tutor in post-test prediction, and discuss hypotheses for why slipping during tutor use is a significant predictor of post-test performance, even after Bayesian Knowledge Tracing estimates are controlled for.

Keywords: Student Modeling, Bayesian Knowledge Tracing, Intelligent Tutoring Systems, Educational Data Mining, Contextual Slip

1 Introduction

Since the mid-1990s, Intelligent Tutoring Systems have used Bayesian approaches to infer whether a student knows a skill, from the student's pattern of errors and correct responses within the software [6, 11, 18]. One popular approach, Bayesian Knowledge Tracing, has been used to model student knowledge in a variety of learning systems, including intelligent tutors for mathematics [10], genetics [7], computer programming [6], and reading skill [3]. Bayesian Knowledge Tracing has been shown to be statistically equivalent to the two-node dynamic Bayesian network used in many other learning environments [13]. Bayesian Knowledge Tracing keeps a running assessment of the probability that a student currently knows each skill. Each

time a student attempts a problem step for the first time, the software updates its probability that the student knows the relevant skill, based on whether the student successfully applied that skill. In the standard four-parameter version of Bayesian Knowledge Tracing described in Corbett & Anderson [6], each skill has two learning parameters, one for Initial Knowledge, and one for the probability of Learning the skill at each opportunity, and two performance parameters, one for Guessing correctly, and one for Slipping (making an error despite knowing the skill). By assessing the student's latent knowledge, it is possible to tailor the amount of practice each student receives, significantly improving student learning outcomes [5, 6].

Recent work has suggested that a new variant of Bayesian Knowledge Tracing, called Contextual-Guess-and-Slip, may be able to predict student performance within the tutoring software more precisely than prior approaches to Bayesian Knowledge Tracing [1,2]. The Contextual-Guess-and-Slip approach examines properties of each student response as it occurs, in order to assess the probability that the response is a guess or slip. However, while better prediction within the software is valuable, the real goal of Bayesian Knowledge Tracing is not to predict performance within the tutoring software, but to estimate the student's underlying knowledge – knowledge that should transfer to performance outside of the tutoring software, for example on post-tests.

Hence, in this paper, we investigate how well the Contextual-Guess-and-Slip model can predict student learning outside of the tutoring software, comparing it both to the canonical four-parameter version of Bayesian Knowledge Tracing, and to the Individual Difference Weights version of Bayesian Knowledge Tracing [6]. The Individual Difference Weights version finds student-level differences in the four parameters, and has been shown to improve the prediction of post-test performance for students who have reached mastery within the tutor. We also investigate other ways to utilize data on student slipping within the learning software, in order to study how to increase the accuracy of post-test prediction.

2 Data

The data used in the analyses presented here came from the Genetics Cognitive Tutor [7]. This tutor 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.

This study focuses on a tutor module that employs a gene mapping technique called *three-factor cross*. The tutor interface for this reasoning task is displayed in Figure 1. In this gene mapping technique a test cross is performed (in this case, of two fruit flies) that focuses on three genes. In Figure 1 the three genes are labeled G, H and F. In the data table on the left of the figure, the first column displays the eight possible offspring phenotypes that can result from this test cross and the second column displays the number of offspring with each phenotype. The problem solution depends on the phenomenon of “crossovers” in meiosis, in which the chromosomes in homologous pairs exchange genetic material. In Figure 1 the student has almost

finished the problem. To the right of the table, the student has summed the offspring in each of the phenotype groups and identified the group which reflects the parental phenotype (no crossovers), which groups result from a single crossover in meiosis, and which group results from two crossovers. The student has compared the phenotype groups to identify the middle of the three genes and entered a gene sequence below the table. Finally, in the lower right the student has calculated the crossover frequency between two of the genes, A and B, and the distance between the two genes. The student will perform the last two steps for the other two gene pairs.

In this study, 71 undergraduates enrolled in a genetics course at Carnegie Mellon University used the three-factor cross module as a homework assignment. Half the students completed a fixed curriculum of 8 problems and the other half completed between 6 and 12 problems under the control of Knowledge Tracing and Cognitive Mastery [6]. The 71 students completed a total of 19,150 problem solving attempts across 9259 problem steps in the tutor. Students completed a paper-and-pencil problem-solving pretest and posttest 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 of the two problems on each test form 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.

Student Teacher

7. In a student lab, a test cross was performed between a fruit fly that was heterozygous for three genes and one that was homozygous recessive. The offspring were scored for the three phenotypes. The student's data is shown below. Determine the gene order and the map distances for the three genes.

0. Frequency of Offspring Types

Type	Number	Group
G H f	3	I
g h F	6	I
g H f	52	II
G h F	59	II
C H F	32	III
g h f	39	III
g H F	388	IV
G h f	421	IV

1. Classify Offspring Groups

# in Group	Offspring Type of Group
9	DCO
111	SCO
71	SCO
809	Parental

Total 1000

2. Order Genes on the Chromosome

Gene 1	Gene 2	Gene 3
G	H	F

3. Compute Distance between each Gene Pair

Gene Pair	Frequency of Recombination	Map Units
G H	$(71 + 9) / 1000$	=> 8
		=>
		=>

Fig. 1. The Three-Factor Cross lesson of the Genetics Cognitive Tutor

After the study, Bayesian Knowledge Tracing and Contextual Guess and Slip models were fit and applied to data from students' performance within the tutor.

3 Bayesian Knowledge Tracing Variants

All of the models discussed in this paper are variants of Bayesian Knowledge Tracing, and compute the probability that a student knows a given skill at a given time. The Bayesian Knowledge Tracing 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 (**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 an initial response (a correct response, error, or help request) to a problem step. First, the system applies Bayes' Theorem to re-calculate the probability that the student knew the skill before making the attempt, using the evidence from the current step. Then, the system accounts for the possibility that the student learned the skill during the problem step. The equations for these calculations are:

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

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

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

Three variants on Bayesian Knowledge Tracing were applied to the data set. The first variant was the standard four-parameter version of Bayesian Knowledge Tracing described in [6], where each skill has a separate parameter for Initial Knowledge, Learning, Guessing, and Slipping. As in [6], the values of Guess and Slip were bounded, in order to avoid the "model degeneracy" problems [cf. 1] that arise when performance parameter estimates rise above 0.5 (When values of these parameters go above 0.5, it is possible to get paradoxical behavior where, for instance, a student who knows a skill is more likely to get it wrong than to get it right). In the analyses in this paper, both Guess and Slip were bounded to be below 0.3. However, unlike in [6], brute force search was used to find the best fitting parameter estimates – all potential parameter combinations of values at a grain-size of 0.01 were tried (e.g. 0.01 0.01 0.01 0.01, 0.01 0.01 0.01 0.02, 0.01 0.01 0.01 0.03... 0.01 0.01 0.02 0.01... 0.99 0.99 0.3 0.1). Recent investigations both in our group and among colleagues [e.g. 12, 14]

have suggested that the Bayesian Knowledge Tracing parameter space is non-convex [4] and that brute force approaches lead to better fit than previously-used algorithms such as Expectation Maximization [cf. 3], Conjugate Gradient Search [cf. 6], and Generalized Reduced Gradient Search [cf. 1]. These same investigations have suggested that brute force is computationally tractable for the data set and parameter set sizes seen in Bayesian Knowledge Tracing, since time increases linearly with the number of student actions but is constant for the number of skills (since only one skill applies to each student action, the number of mathematical operations is identical no matter how many skills are present). The four-parameter model's number of parameters is the number of cognitive rules * 4 – in this case, 32 parameters.

The second variant was Contextual-Guess-and-Slip [1, 2]. In this approach, as above, each skill has a separate parameter for Initial Knowledge and Learning. However, Guess and Slip probabilities are no longer estimated for each skill; instead, they are computed each time a student attempts to answer a new problem step, based on machine-learned models of guess and slip response properties in context (for instance, longer responses and help requests are less likely to be slips). The same approach as in [1, 2] was used, where 1) a four-parameter model is obtained, 2) the four-parameter model is used to generate labels of the probability of slipping and guessing for each action within the data set, 3) machine learning is used to fit models predicting these labels, 4) the machine-learned models of guess and slip are substituted into Bayesian Knowledge Tracing in lieu of skill-by-skill labels for guess and slip, and finally 5) parameters for Initial Knowledge and Learn are fit. Greater detail on this approach is given in [1, 2]. The sole difference in the analyses in this paper is that brute force was used to fit the four-parameter model in step 1, instead of other methods for obtaining four-parameter models. In [1, 2], Contextual-Guess-and-Slip models were found to predict student correctness within three Cognitive Tutors for mathematics (Algebra, Geometry, and Middle School Mathematics) significantly better than four-parameter models obtained using curve-fitting [cf. 6] or Expectation Maximization [3]. As Contextual-Guess-and-Slip replaces two parameters *per skill* (G and S) with a smaller number of parameters across all skills (contextual G and contextual S), Contextual-Guess-and-Slip has a smaller total number of parameters, though only slightly so in this case, given the small number of skills; the parameter reduction is much greater when a larger number of skills are fit at once – in the mathematics tutors, the Contextual-Guess-and-Slip models only had 55% as many parameters as the four-parameter models.

The third variant was Individual Difference Weights on Bayesian Knowledge Tracing. With Individual Difference Weights, a four-parameter model is fit, and then a best-fitting weight for each student is computed for each parameter (e.g. student74 has one weight for L0 for all skills, one weight for T for all skills, one weight for G for all skills, and one weight for S for all skills, and parameter values are a function of the skill parameters and student weights). The student's individualized parameter values for a given skill are computed as a function of their individual difference weights and the skill-level parameters, using a formula given in [6]. That paper found that as students approached cognitive mastery, the Individual Difference Weights model was more accurate at predicting post-test scores than the four-parameter model. As the Individual Difference Weights approach has four parameters for each skill and

four parameters for each student, it has substantially more parameters than the other approaches – in this case, 316 parameters.

4 Modeling Student Performance

4.1 Predicting Performance in the Tutor

While knowledge tracing models the student's knowledge, the underlying assumptions also yield an accuracy prediction in applying a rule:

$$P(\text{correct}_n) = P(L_{n-1}) * (1 - P(S)) + (1 - P(L_{n-1})) * P(G)$$

However, applying this rule to compare the three models' fit to the tutor data biases in favor of the Contextual-Guess-and-Slip model, since that model examines properties of each student response in order to generate a contextualized estimate of $p(S)$ and $p(G)$ for that response. To compare the three models' fit to student tutor performance in an unbiased fashion, we predict student correctness at time N just from the model's knowledge estimate at time $N-1$. This approach underestimates accuracy for all models (since it does not include the probability of guessing and slipping when answering), but does not bias in favor of any type of model.

We evaluate model goodness using A' [8], the probability that the model can distinguish correct responses from errors, because A' is an appropriate metric when the predicted value is binary (correct or not correct), and the predictors are numerical (probability of knowing the skill, or probability of getting the skill correct). We determine whether a model is statistically significantly better than chance by computing the A' value for each student, comparing differences between A' values and chance [cf. 8] (giving a Z value for each student), and then using Stouffer's method [15] to aggregate across students. We determine whether the difference between two models is statistically significant by computing A' values for each student, comparing differences between A' values with a Z test [8], and then aggregating across students using Stouffer's method [15]. Both of these methods account for the non-independence of actions within each student.

Each model's effectiveness at predicting student performance within the tutor is shown in Figure 2. The four-parameter model achieved A' of 0.758 in predicting student performance at time N from the students' knowledge estimate at time $N-1$. The Contextual-Guess-and-Slip model achieved A' of 0.755. The Individual Difference Weights model performed more poorly, with an A' of 0.734. All three models were significantly better than chance, $Z=48.69$, $Z=44.85$, $Z=45.33$, $p<0.0001$. The difference between the four-parameter model and the Individual Difference Weights model was significant, $Z=-2.12$, $p=0.03$, but the other two differences were not significant, $Z=-1.36$, $p=0.17$, $Z=-0.78$, $p=0.43$.

If we instead predict student performance at time N using the guess and slip parameters, the four-parameter model and the individual difference weights models achieve much closer performance (As previously mentioned, it is not valid to use this approach with the Contextual-Guess-and-Slip model). In this case, the four-parameter

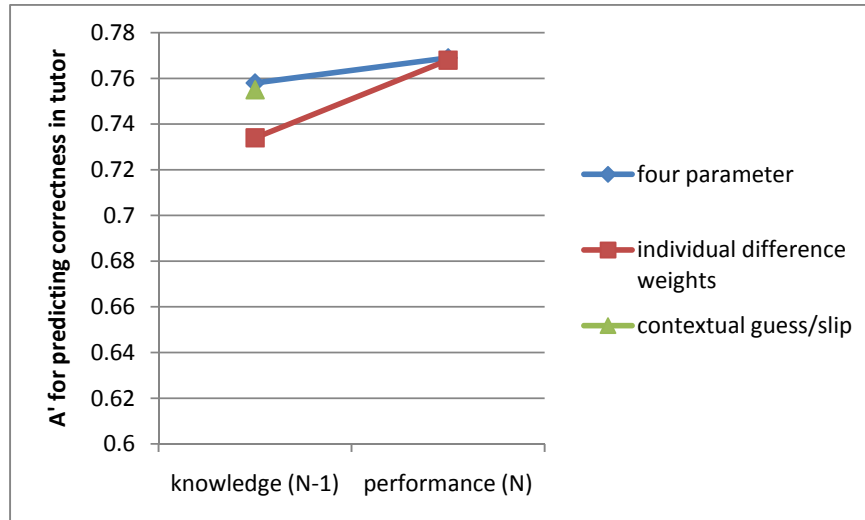


Fig. 2. The ability of each model to predict performance within the tutor

model achieves an A' of 0.769, and the Individual Difference Weights model achieves an A' of 0.768. Both models are significantly better than chance, $Z=50.44$, $Z= 51.12$, $p<0.001$. These two models are not statistically significantly different from each other, $Z=0.127$, $p=0.89$.

4.2 Predicting Post-Test Solely From Final Knowledge Estimates

Beyond predicting performance within the tutor, it is important to see how well the different methods predict student performance outside of the tutor. If any method sees significant degradation outside of the tutor, it may be over-fit to student behavior within the tutor, rather than capturing indicators of learning that will persist even outside of the tutor.

Again, the simplest way to use tutor estimates of student knowledge to predict the post-test, is simply to look at the correlation between just the models' estimate of student knowledge and the student's post-test performance. This approach is unlikely to be the most precise approach, as it ignores the possibility that the student will guess or slip on the test. However, it is equally feasible for all three approaches.

In predicting the post-test, we account for the number of times each skill will be utilized on the test (assuming perfect performance). Of the eight skills in the tutor lesson, one is not exercised on the test, and is eliminated from the model predicting the post-test. Of the remaining seven skills, four are exercised once, two are exercised twice and one is exercised three times, in each of the two posttest problems. These first two skills are each counted twice and the latter skill three times in our attempts to predict the post-test. We utilize this approach in all attempts to predict the post-test in this paper (including in later sections). As post-test scores represent the average

correct on the test, we average the estimates of student skill together rather than multiplying them.

The full pattern of results for each model's ability to predict the post-test, based on the different assumptions in sections 4.1, 4.2, and 4.3, is shown in Figure 3. We predict the post-test using each model's estimate of student knowledge of each skill, assessing the goodness of prediction with correlation, since the model estimates and the actual test scores are both numerical.

The four-parameter approach achieves a correlation of 0.430 to the post-test. The contextual-guess-and-slip approach achieves a correlation of 0.289 to the post-test. The individual difference weights approach achieves a correlation of 0.412 to the post-test. All three of the correlations are statistically significant ($p = .02$ for the contextual-guess-and-slip approach, and $p < 0.01$ for the other two), while none of the differences among the correlations were statistically significant, although the difference between the four-parameter model and the contextual-guess-and-slip approach approached significance, $t(68) = 1.63$, $p = 0.12$, for a two-tailed test of the significance of the difference between two correlation coefficients for correlated samples.

4.3 Predicting Post-Test From Final Knowledge Estimates and Non-Contextual Guess/Slip Estimates

One limitation to the approach above is that performance, whether in the tutor or on a paper post-test, is not simply a function of the student's knowledge. It is also a function of the probability that the students guesses (giving a correct answer despite not knowing the skill), or slips (making an error despite knowing the skill), as described in section 4.1. Determining appropriate guess and slip rates for the paper post-test is not a trivial problem, since the students are working in a different environment. For instance, they may be more or less prone to physical slips (such as mis-typing or mis-writing) on paper than in the tutor, and they may be more or less cautious when the tutor is not providing immediate accuracy feedback. However, the performance parameter estimates derived from tutor behavior with both the standard four-parameter model and the Individual Difference Weights version have been shown to predict test data quite accurately [6, 7].

Within the four-parameter and Individual Difference Weights approaches, we can compute the probability that the student will get each answer right on the post-test based on both the final knowledge estimates, and the model's parameters for guess and slip for each skill (or for Individual Difference Weights, the parameters for each skill and student):

$$P(\text{correct}_n) = P(L_{n-1}) * (1 - P(S)) + (1 - P(L_{n-1})) * P(G)$$

When we do this, the correlation between the estimates of the probability of getting the skill correct in the four-parameter model, and the post-test score rises very slightly, from 0.430 to 0.434. The difference between this model, and the earlier fit where four-parameter model estimates of final knowledge are used, is not statistically

significant, $t(68)=0.63$, $p=0.53$, for a two-tailed test of the significance of the difference between two correlation coefficients for correlated samples.

By contrast, the correlation between the estimates of the probability of getting the skill correct in the Individual Difference Weights model, and the post-test score appears to drop within this approach, from 0.412 to 0.352. But as above, this model does not differ significantly in correlation from the model using the estimates of the probability of knowledge in the individual difference weights model, $t(68)=0.66$, $p=0.51$, for a two-tailed test of the significance of the difference between two correlation coefficients for correlated samples. In addition, there is not a statistically significant difference between the two models, $t(68)= 1.19$, $p=0.23$, for a two-tailed test of the significance of the difference between two correlation coefficients for correlated samples.

4.4 Predicting Post-Test From Final Knowledge Estimates and Contextual Guess/Slip Estimates

Contextual models of guess and slip assess the probability that a student slipped at a specific time, within the tutoring software. These models cannot be used as-is to predict guess and slip on a paper post-test, as behavioral indicators such as timing are not available. Instead, the contextual estimates of guess and slip from within the tutor can be used as an indicator of how much each student guessed and slipped while using the tutor. The most straightforward way to do this is to average the contextual slip and guess values at each problem step.

Hence, one option for using these estimates is to use the average guess and slip for each student and skill in lieu of the non-contextual parameter estimates, within equation 1 above. A model which does this achieves a poor correlation to post-test score, 0.181. It is not statistically significantly worse than the earlier fit using the contextual guess-and-slip model's estimates of final knowledge, $t(68)=-0.70$, $p=0.48$, for a two-tailed test of the significance of the difference between two correlation coefficients for correlated samples. It is, however, marginally statistically significantly worse than the four-parameter model's prediction of performance, $t(68)=-1.74$, $p=0.09$.

Although the contextual guess-and-slip model's estimates of performance, used in this fashion, led to poor prediction of the post-test score, there may still be useful information in the contextual estimates of guess and slip themselves. The correlation between average contextual slip and post-test score, $r=0.272$, is statistically significantly higher than chance, $F(1,69)=5.521$, $p=0.02$, and the correlation between average contextual guess and post-test score, $r=-0.325$, is also statistically significantly higher than chance, $F(1,69)=8.17$, $p<0.01$. In addition, if we restrict analysis to values of contextual slip over 0.5 (e.g., where there is a probability over 50% that the action is a slip), the correlation to post-test score is particularly strong, $r=0.453$, and is statistically significantly higher than chance, $F(1,69)=17.801$, $p<0.001$.

One way to determine whether this information is potentially useful is to generate a post hoc prediction of post-test performance with a combination of the average contextual slip, and the four-parameter prediction of post-test performance (e.g. the

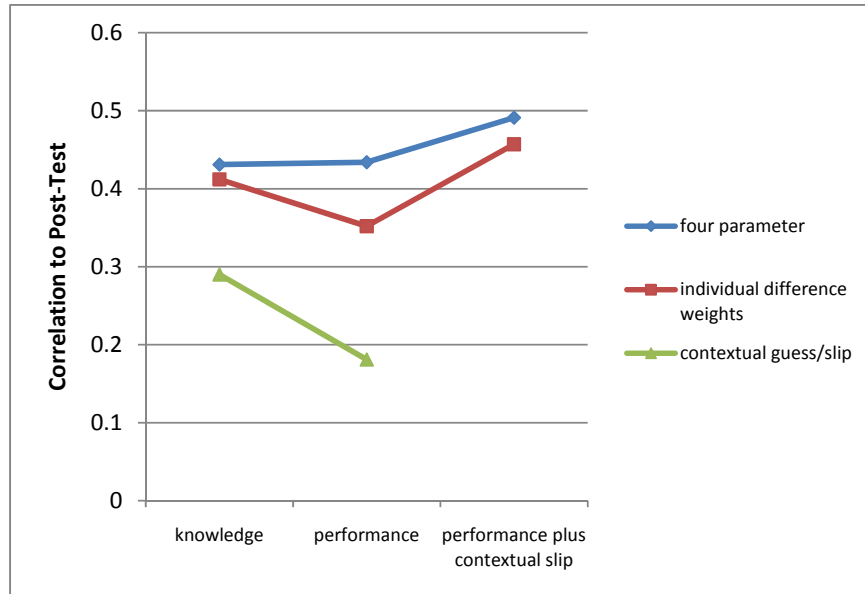


Fig. 3. The correlation of each model to the post-test

model including skill-level estimates of guess and slip, from section 4.3), using linear regression. While the average contextual slip and average contextual guess are not statistically significant predictors in a model already containing the four-parameter model prediction, average contextual slip over 0.5 is statistically significant in a model already containing the four-parameter model prediction, $F(2,68)=10.81$, $p<0.001$. In other words, a linear regression model including both average contextual slip over 0.5 and the four-parameter prediction, has statistically significantly better fit than the model just containing the brute-force prediction, achieving a correlation to post-test of 0.491.

This finding indicates that slipping within the tutor is an indicator of some aspect of student learning that is associated with the failure to transfer knowledge to a cognitively identical problem in a different setting (outside the tutor).

A similar pattern is seen with the Individual Difference Weights model. Again, average contextual slip over 0.5 is statistically significant in a model already containing the Individual Difference Weights model prediction, $F(2,68)=9.00$, $p<0.01$.

A clear implication can be seen from this pattern of results. Though the current formulation of Contextual-Guess-and-Slip does not port to the post-test, there is clear evidence that future models integrating evidence on contextual slip have the potential to do better at predicting the post-test than the current generation of Bayesian Knowledge Tracing prediction. Determining how to integrate contextual slip information in a replicable fashion will be an important area of future work.

5 Discussion and Conclusions

Overall, the findings here suggest that the Contextual-Guess-and-Slip approach, in its current form, does a fine job of predicting performance within the tutoring system, performing comparably to or slightly better than the four-parameter approach and Individual Difference Weights approach. However, the Contextual-Guess-and-Slip approach predicted performance much more poorly outside of the tutor than within the tutor. One possible explanation is that Contextual-Guess-and-Slip is over-fit to aspects of student performance within the tutor; that said, given that Contextual-Guess-and-Slip has fewer parameters, it is unlikely that it is over-fit in general, compared to the other models [cf. 9]. Another explanation is that by allowing a student who slips a great deal to still be assessed as having mastery, Contextual-Guess-and-Slip discards evidence of incomplete or non-robust knowledge.

Note that the Individual Difference Weights approach also failed to predict the post-test better than the standard four-parameter approach. In Corbett and Anderson's earlier work [6], the advantage of the Individual Difference Weights model emerged when students reached very high levels of $P(L_n)$ estimates, higher than students reached in this study. The results of this study provides converging evidence that the benefit of individual different weights only emerges for high $P(L_n)$ levels.

Despite the failure of Contextual-Guess-and-Slip to predict performance on the post-test, estimates of Contextual Slip appear to be a valuable addition to the knowledge and performance prediction obtained in the four-parameter approach. A post-hoc model combining average contextual slip among actions where $P(S)$ was over 0.5, and the performance predictions from Bayesian Knowledge Tracing, performs significantly better than the performance predictions alone. This finding indicates that slipping during the tutor is an indicator of some aspect of student learning that is not captured by Bayesian Knowledge Tracing. However, $P(S)$ can have several meanings, including indicating shallow knowledge or general carelessness during tutor usage. It may be possible to disentangle these possibilities with measures of robust learning [cf. 16, 17], where shallow learning is likely to compromise performance to a greater degree, and with questionnaire assessments of carelessness.

Hence, it appears that potential remains for utilizing contextual estimates of slipping in predicting student performance outside of intelligent tutoring systems. This is important, because better prediction of post-test scores is likely to lead to more effective adaptation within intelligent tutoring systems – in particular, understanding **why** slip predicts post-test will determine which type of adaptation is most appropriate for a student who appears to know the skill within the software, but who has frequently slipped during the process of knowledge acquisition.

Acknowledgements. This research was supported by the National Science Foundation via grant “Empirical Research: Emerging Research: Robust and Efficient Learning: Modeling and Remediating Students’ Domain Knowledge”, award number DRL0910188.

References

1. Baker, R.S.J.d., Corbett, A.T., Aleven, V.: 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 (2008).
2. Baker, R.S.J.d., Corbett, A.T., Aleven, V.: Improving Contextual Models of Guessing and Slipping with a Truncated Training Set. In: *Proceedings of the 1st International Conference on Educational Data Mining*, 67-76 (2008).
3. Beck, J.E., Chang, K.-m.: Identifiability: A Fundamental Problem of Student Modeling. In: *Proceedings of the 11th International Conference on User Modeling (UM 2007)*.
4. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge, UK: Cambridge University Press (2004).
5. Corbett, A.: Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. Gmytrasiewicz, & J. Vassileva (Eds.), *UM2001, User Modeling: Proceedings of the Eighth International Conference* (pp. 137-147). Berlin: Springer (2001)
6. Corbett, A.T., Anderson, J.R.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278 (1995).
7. Corbett, A., Kauffman, L., Maclaren, B., Wagner, A., Jones, E.: A Cognitive Tutor for Genetics Problem Solving: Learning Gains and Student Modeling. *Journal of Educational Computing Research*, 42, 219-239 (2010).
8. Fogarty, J., Baker, R., Hudson, S.: Case Studies in the use of ROC Curve Analysis for Sensor-Based Estimates in Human Computer Interaction. In: *Proceedings of Graphics Interface*, 129-136 (2005).
9. Hawkins, D.M.: The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, 44 (1), 1-12 (2004).
10. Koedinger, K.R., Corbett, A.T.: Cognitive tutors: Technology bringing learning sciences to the classroom. In: *The Cambridge handbook of the learning sciences*, R.K. Sawyer, Editor. Cambridge University Press: New York, NY, 61-77 (2006).
11. Martin, J., VanLehn, K.: Student Assessment Using Bayesian Nets. *International Journal of Human-Computer Studies*, 42, 575-591 (1995).
12. Pavlik, P.I., Cen, H., Koedinger, J.R.: Performance Factors Analysis – A New Alternative to Knowledge Tracing. In: *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 531-540 (2009).
13. Reye, J.: Student Modeling based on Belief Networks. *International Journal of Artificial Intelligence in Education*, 14, 1-33 (2004).
14. Ritter, S., Harris, T., Nixon, T., Dickinson, D., Murray, R.C., Towle, B.: Reducing the Knowledge Tracing Space. In: *Proceedings of the 2nd International Conference on Educational Data Mining*, 151-160 (2009).
15. Rosenthal, R., Rosnow, R.L.: *Essentials of Behavioral Research: Methods and Data Analysis* (2nd Edition). Boston, MA: McGraw-Hill (1991).
16. Schmidt, R.A., Bjork, R.A.: New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3 (4), 207-217 (1992).
17. Schwartz, D. L., Martin, T.: Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22, 129-184 (2004).
18. Shute, V.J.: SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction*, 5 (1), 1-44 (1995).