Improving Contextual Models of Guessing and Slipping with a Truncated Training Set

Ryan S.J.d. Baker, Albert T. Corbett, Vincent Aleven
{rsbaker, corbett, aleven}@cmu.edu
Human Computer Interaction Institute, Carnegie Mellon University

Abstract. A recent innovation in student knowledge modeling is the replacement of static estimates of the probability that a student has guessed or slipped with more contextual estimation of these probabilities [2], significantly improving prediction of future performance in one case. We extend this method by adjusting the training set used to develop the contextual models of guessing and slipping, removing training examples where the prior probability that the student knew the skill was very high or very low. We show that this adjustment significantly improves prediction of future performance, relative to previous methods, within data sets from three different Cognitive Tutors.

1 Introduction

Developing accurate models of students' knowledge as they use educational software is valuable for many goals. First, it enables learning systems to respond more accurately to differences in student knowledge, optimizing the amount of practice each student receives on each skill [cf. 9]. Second, estimates of student knowledge are often a useful component in the development of models of more complex behavioral constructs, such as gaming the system [3], which are in turn increasingly used in analyses of learning and motivation [cf. 4,10]. As such, assessments of student knowledge are one of the key building-blocks of educational data mining.

One popular and validated method for modeling students' knowledge is Corbett and Anderson's [9] Bayesian Knowledge Tracing model, which has been used within Cognitive Tutors [cf. 1] for mathematics [1], computer programming [9], and reading skill [7]. This model is statistically equivalent to the two-node dynamic Bayesian network used in many other learning environments [14].

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 and algorithms derived from Bayes Theorem. In order to do this, Bayesian Knowledge Tracing uses four parameters for each skill, including a probability that the student will "guess" and obtain a correct answer without knowing the skill (**G**) and a probability that the student will "slip" and obtain an incorrect answer even though the student knows the skill (**S**). In Corbett and Anderson [9], these parameters are estimated from data for each skill and are invariant across context (i.e. for a given skill, any student will always have the same probability of slipping, no matter what the situation is).

Recent work has attempted to improve on Corbett and Anderson's original approach. Beck and Chang [6,7] noted that multiple sets of parameters fit performance data equally well within Corbett and Anderson's approach, and introduced a method for selecting a single best set of parameters, using Dirichlet Priors fit across skills. The Dirichlet Priors

method significantly improved fit to data from the Geometry Cognitive Tutor[6], but did not improve fit to data from a Cognitive Tutor for middle school mathematics [2].

In other recent work, Baker, Corbett, and Aleven [2] presented a method for estimating the guess (**G**) and slip (**S**) model parameters contextually. In this approach the model's estimate of the probability that an action is a guess or slip is no longer invariant – instead, it depends on details of the action (such as how much time the action took, and how often the student requested help on the skill in the past). Though the Contextual Guess and Slip approach used about half as many parameters as Corbett and Anderson's original approach and Beck and Chang's Dirichlet Priors approach, it was significantly more successful at predicting student performance within an intelligent tutoring system for middle school mathematics [2] than either of these two earlier approaches.

In this paper, we propose an extension to the Contextual Guess and Slip method. Specifically, we refine the training set used to generate the contextual models of guessing and slipping, in order to make the training set more representative of the situations where these estimates will affect predictions of future student performance. We study this new model both in the same data set as [2] and replicate its effectiveness in two new data sets, from Cognitive Tutors on Geometry and Algebra. This paper's contribution is both in showing that this extension significantly improves the model, and in showing that the Contextual Guess and Slip method improves prediction in multiple intelligent tutors.

2 Bayesian Knowledge-Tracing

As previously mentioned, the Contextual Guess and Slip model of student knowledge is an extension of Corbett and Anderson's [9] Bayesian Knowledge Tracing model. Both models compute the probability that a student knows a given skill at a given time, interpreting data on student performance with a four-parameter model.

In the models' 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.

However, the parameter values are different between models. In Corbett and Anderson's approach and Dirichlet Priors, each skill has four parameter values, used in all situations. In the Contextual Guess and Slip model, each skill has two parameter values used in all situations ($\mathbf{L_0}$**, T**), and two parameter values which vary according to context (**G, S**).

The system's estimate that a student knows a skill is continually updated, every time the student gives a first response (a correct response, error, or help request) to a problem step. First, the system re-calculates 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:

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

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

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

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. Corbett and Anderson [9] instead used a bounded approach, where the guess and slip parameters are not allowed to rise above pre-chosen thresholds. Beck and Chang [7] showed that both of these approaches are prone to the "identifiability problem", where multiple models can fit the data equally well. They proposed that models be chosen using Dirichlet Priors, which chooses a single best model by biasing parameters towards values that fit the whole data set well. Within this paper, we fit parameters for the Dirichlet Priors approach using Bayes Net Toolkit-Student Modeling (BNT-SM) [6].

However, the baseline and Dirichlet Priors approaches may result in parameters which are "theoretically degenerate" [2]. The conceptual idea behind using Bayesian Knowledge Tracing to model student knowledge is that knowing a skill generally leads to correct performance, and that correct performance implies that a student knows the relevant 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 answer incorrectly than correctly; similarly, a guess parameter over 0.5 signifies that a student who does not know a skill is more likely to answer correctly than incorrectly.

3 The Contextual Guess and Slip Model of Student Knowledge

Baker, Corbett, and Aleven [2] proposed a new way of fitting parameters: estimating whether each individual student response is a guess or a slip based on contextual information (such as prior history and the speed of response), rather than using fixed guess and slip probability estimates across situations. This modeling approach was tested within a data set from an intelligent tutor for middle school mathematics, and significantly reduced the degree of model degeneracy. This approach was significantly better at predicting student performance than models developed using the Dirichlet Priors, bounded, and baseline methods, despite using substantially fewer parameters.

The first step of the Contextual Guess and Slip method is to label a set of existing student actions with the probability that these actions involve guessing or slipping, using the Dirichlet Priors skill estimates. The set of student actions to be labeled is drawn (in this

approach) from the set of first actions on each problem step, on the set of skills for which the Dirichlet Priors model is not theoretically degenerate. This set of skills was used, rather than all skills, in order to avoid training the models to include model degeneracy. Each student action (***N***) is labeled with the probability that it represents 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 (in most cases) 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 information about the action's correctness:

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

Next, the probability that the student knew the skill at time ***N*** can be calculated, given information about the actions at time ***N+1*** and ***N+2*** (which we term $\boldsymbol{A_{+1+2}}$). This is done 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 times ***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.

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

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

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

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

Once the set of actions is labeled with estimates of whether each action was a guess or slip, the labels are used to train models that can accurately predict at run-time the probability that 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 and events before the action (i.e. no future data is used).

For each action, a set of 23 features are distilled to describe that action, including information on the action itself (time taken, type of interface widget) and the action's historical context (for instance, how many errors the student had made on the same skill in past problems). Linear Regression is then used, within Weka [16], to create 2 models predicting the probability of guessing (model 1) and slipping (model 2).

Finally, these 2 models are used within Bayesian Knowledge Tracing to dynamically estimate the probability that each response is a guess or a slip. The first action of each opportunity to use a skill is labeled (using the machine-learned models) with predictions as to how likely it is to be a guess or slip, and parameter values are fit for P(**T**) and P($\mathbf{L_0}$), for each skill. At this point, this model – like the earlier work – can make a prediction about student knowledge each time a student attempts to use a skill for the first time on a given problem step. It is worth noting that this model involves considerably fewer parameters than previous models – whereas the Dirichlet Priors and baseline models had exactly 4 parameters per skill, this model fits just over 2 parameters per skill (parameters for **T** and $\mathbf{L_0}$ for each skill, with parameters for **G** and **S** amortizes across all skills).

4 Choice of Data Set Used to Train Contextual Models

In the version of the Contextual Guess and Slip method published in [2], the data set used to train a knowledge model is the set of first actions on each problem step, on the set of skills for which the Dirichlet Priors model is not theoretically degenerate. However, there are potential drawbacks to using this data set. Specifically, if the data set involves significant amounts of over-practice, there may be a large number of actions for which a student has a probability close to 1 of knowing the relevant skill. On these actions, the estimated probability that any incorrect response is due to a slip may be very close to 1, and the probability that any correct response is due to a guess may be very close to 0.

To give an example: Let us consider a skill which has Dirichlet Prior values of P(**G**) = 0.3 P(**S**) = 0.2, P(**T**) = 0.1, and at the current opportunity to practice the skill P($\mathbf{L_{n-1}}$) = 0.99. If the current action is incorrect (**~C**), and the following two actions are not correct (**~C,~C**), it is reasonable to assume that the current incorrect action is due to not knowing the skill, rather than a slip. However, the probability that the current action was a slip will be very high, 97.6%, according to the equations above, because of the very high value of P($\mathbf{L_{n-1}}$). This may be the correct prediction in this context; but if the model trains on this prediction and then uses it in different contexts when P($\mathbf{L_{n-1}}$) is further from 1, the probability that those actions are slips may be overestimated. (One explanation for why three errors in a row could occur on a skill with very high P($\mathbf{L_{n-1}}$) is that the mapping between actions and skills may have errors [cf. 8,9]; fixing such errors is a research topic in its own right [cf. 5,8]).

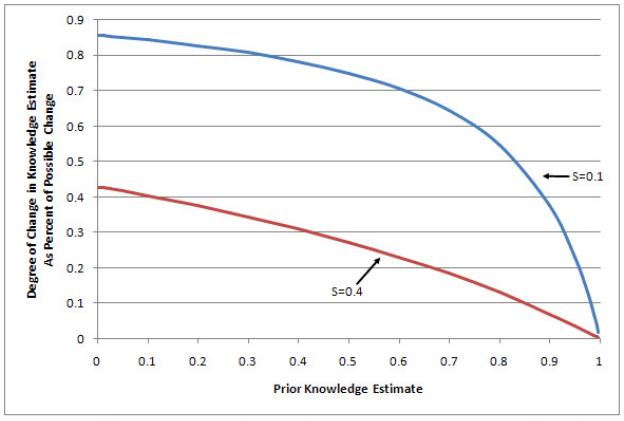

Figure 1. The degree to which an incorrect response can affect the knowledge estimate, for different levels of student prior knowledge and slip (S) parameters. G and T are held constant. The graph of how correct responses can affect the knowledge estimate is similar but reversed horizontally.

Pragmatically, it is more important for these estimations to be accurate when ($\mathbf{L_n}$) is distant from 0 and 1. As P($\mathbf{L_{n-1}}$) approaches 1, P(**S**) has less and less impact on P($\mathbf{L_n}$) – the base probability is too extreme. This can be seen in Figure 1. Similarly, as P($\mathbf{L_{n-1}}$) approaches 0, P(**G**) has less and less impact on P($\mathbf{L_n}$). Hence, it is more important for the model to be highly accurate in cases where P($\mathbf{L_n}$) is not very close to 0 or 1.

One way to accomplish this is to truncate the training set, so that actions where P($\mathbf{L_{n-1}}$) is too close to 0 or 1 are omitted. We choose the cutoffs 0.1 and 0.9, to err on the side of truncating too much rather than truncating too little. Hence, only cases where $0.1 < P(\mathbf{L_{n-1}}) < 0.9$ are included in the training set for the models of guessing and slipping. We can then follow the procedure given in the previous section to create the machine learned models of guessing and slipping, and then use these models in the model of student knowledge.

We call the resultant knowledge model Truncated Training Set Contextual Guess and Slip, or Contextual-Trunc for short. In the following sections, we will compare this model to a version of the Contextual model without any truncation of the training set, and to the Dirichlet Priors model. To avoid bias, all models are evaluated on non-truncated data.

5 Data

We evaluate the models of knowledge tracing discussed here within data sets drawn from three Cognitive Tutors, on Algebra, Geometry, and Middle School mathematics. Cognitive Tutors are a popular type of interactive learning environment now used by around half a million students a year in the USA. In Cognitive Tutors, students solve problems, with exercises chosen based on the student knowledge model [1], on-demand help, and instant feedback. Cognitive Tutors have been shown to significantly improve student performance on standardized exams and tests of problem-solving skill [13].

The Algebra and Geometry data sets were obtained from the Pittsburgh Science of Learning Center DataShop (https://learnlab.web.cmu.edu/datashop/). The DataShop is a public resource for the learning science community, giving free access to anonymized

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

	Actions	Problem Steps	Skills	Students
Middle School	581,785	171,987	253	232
Algebra	436,816	136,408	88	59
Geometry	244,398	32,997	144	88

data sets of student use of learning software. The Middle School data set was previously collected by the authors [cf. 3]. Each data set consisted of an entire year's use of an intelligent tutor in schools in the suburbs of a city in the Northeastern USA; we are not aware of any overlap in the student population between data sets. Within each data set, actions which were not labeled with skills (information needed to apply Bayesian Knowledge Tracing) were excluded. However, all other actions on all other skills (including actions eliminated from the Contextual and Contextual-Trunc training sets) are included. The magnitude of the data sets is shown in Table 1.

6 Results

Bayesian Knowledge-Tracing models make predictions about student knowledge (i.e. the probability a student knows a skill at a given time). These predictions can be validated by comparing them to future performance in two ways. The first is to compare actions at time ***N*** to the models' predictions of the probability that actions at time ***N*** will be correct – $P(\mathbf{L_n})*P(\mathbf{\sim S})+ P(\mathbf{\sim L_n})*P(\mathbf{G})$. This method accurately represents exactly what each model predicts; however, this method biases in favor of the Contextual Guess and Slip models, since those models use information associated with the answer being predicted to estimate the probability of guessing and slipping. Therefore, we instead 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 under-estimates goodness of fit 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) as the measure of goodness-of-fit. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. To assess the statistical significance of the differences between models, we compute A' for each student in each model, compute the standard error of the A' estimates [12], use a Z test to find the difference between models within each student [11], use Stouffer's Z [15] to aggregate across students, and finally compute the (two-tailed) statistical significance of the Z score obtained. This method does not collapse across any data (i.e. it is not overly conservative) but accounts for the non-independence of actions within a single student.

Within the Middle School data set, the Dirichlet Priors approach achieves an average A', across students, of 0.641. The Contextual approach achieves an average A' of 0.749. The Contextual-Trunc approach achieves an average A' of 0.758. The Dirichlet Priors approach is statistically significantly poorer than the other two approaches, Z=59.56,

Table 2. The A' of each model within each tutor, across students. The Contextual-Trunc model is in boldface where it is statistically significantly better than the Dirichlet Priors model, and in italics where it is statistically significantly better than the Contextual model.

	Dirichlet Priors	Contextual	Contextual-Trunc
Middle School	0.641	0.749	***0.758***
Algebra	0.694	0.632	***0.707***
Geometry	0.638	0.666	**0.669**

p<0.0001, Z=64.17, p<0.0001. The Contextual-Trunc approach is statistically significantly better than the Contextual approach, Z=4.59, p<0.0001.

Within the Algebra data set, the Dirichlet Priors approach achieves an average A' of 0.694. The Contextual approach achieves an average A' of 0.632. The Contextual-Trunc approach achieves an average A' of 0.707. The Contextual-Trunc approach is statistically significantly better than the Dirichlet Priors approach, Z=2.89, p<0.01. However, the Contextual approach is statistically significantly worse than the Dirichlet Priors approach, Z= -27.76, p<0.0001. The Contextual-Trunc approach is statistically significantly better than the Contextual Approach, Z= 30.65, p<0.0001.

Within the Geometry data set, the Dirichlet Priors approach achieves an average A' of 0.638. The Contextual approach achieves an average A' of 0.666. The Contextual-Trunc approach achieves an average A' of 0.669. The Contextual-Trunc approach is statistically significantly better than the Dirichlet Priors approach, Z=2.52, p=0.01; the difference between the Dirichlet Priors approach and the Contextual approach is (at best) marginally significant, Z=1.60, p=0.11. The difference between the Contextual and Contextual-Trunc approaches is not significant, Z=0.92, p=0.35.

The full pattern of results is shown in Table 2. As can be seen, the Contextual-Trunc model consistently performed better than the Dirichlet Priors model. The Contextual model, by contrast, performed almost as well as the Contextual-Trunc model in two cases, but was far worse than the other models in the Algebra data set. The primary difference appears to have been that the Algebra Contextual model predicted massively more slips than the other two models did. Whereas the average value of P(**S**) (across skills) in the Algebra Dirichlet Priors model was 0.19, and the average value of P(**S**) (across actions) in the Algebra Contextual-Trunc model was 0.38, the average value of P(**S**) (across actions) in the Algebra Contextual model was 0.67. Values of the slip parameter above 0.5 are degenerate, as discussed earlier; these values cause the model to very quickly infer that a student has mastered a skill, even when the student displays poor performance. By truncating the data set used to train the contextual model of slipping, the Contextual-Trunc model avoids this degenerate performance and is significantly more successful at predicting student performance.

7 Conclusions

In this paper, we have presented an improvement to the Contextual Guess and Slip model proposed in [2]. Earlier models of student knowledge [cf. 7,9] estimated a single probability of guessing and slipping for each skill, and used that estimate for all actions. By contrast, the model presented here (and the model in [2]) contextually estimate the probability that a student obtained a correct answer by guessing, or an incorrect answer by slipping. The Contextual models also use fewer parameters to estimate student knowledge than previous models.

In earlier work [2], contextual models of guess and slip were trained using every action involving non-degenerate skills. In this paper, we adjusted the training set, removing

actions where the probability that the student already knows the skill is below 0.1 or above 0.9. Truncating the training set in this fashion avoids training on cases where probabilities of guess or slip are close to 0 or 1 due to prior probabilities rather than the information contained in successive actions.

We show that using a truncated training set leads to models which are statistically significantly better at predicting future student performance than the Dirichlet Priors approach to parameter selection. A non-truncated training set is also better than Dirichlet Priors in two cases, but in a third case (the Algebra data set) performs significantly worse, due to assigning degenerate values for the slip parameter. This shows that it is valuable to test new student modeling methods on data sets from different learning software (increasingly available in publicly accessible databases such as the PSLC DataShop), since the non-truncated data set would have been perfectly adequate in the Geometry and Middle School data sets.

Further investigation of how to optimally truncate training sets is probably warranted. The choice of 0.1 and 0.9 as cut-offs in this data set is based on data but ultimately arbitrary, and while the solution is effective, a more principled method for selecting cut-offs may lead to better performance. Studying whether truncation of training sets is useful to other classification problems in educational data is another area for future work; input probabilities very close to 0 or 1 are likely to bias the output of any Bayesian method.

At this point, contextual estimation of guess and slip has proven to be better at predicting future performance than earlier methods for student knowledge modeling, for three different learning systems. In the long term, more sensitive and accurate estimation of student knowledge has the potential to improve the effectiveness of learning software. Additionally, as accurate knowledge modeling is a key component of models of complex student behavior used in data mining analyses [cf. 4, 10], better knowledge modeling is likely to be useful to the broader advancement of the field of educational data mining.

8 Acknowledgements

We would like to thank Project LISTEN and Joseph Beck for offering the BNT-SM toolkit used within our model creation process. 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.

9 References

[1] Anderson, J.R., Corbett, A.T., Koedinger, K.R., and Pelletier, R. Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 1995, 4 (2), 167-207.

[2] Baker, R.S.J.d., Corbett, A.T., Aleven, V. (to appear) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. To appear in *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. Online at http://www.cs.cmu.edu/~rsbaker/BCA2008V.pdf

[3] Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (to appear) Developing a Generalizable Detector of When Students Game the System. To appear in *User Modeling and User-Adapted Interaction*. Online at http://www.cs.cmu.edu/~rsbaker/USER475.pdf

[4] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., Koedinger, K. Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments. *Journal of Interactive Learning Research*, 2008, 19 (2), 185-224.

[5] Barnes, T. The Q-matrix Method: Mining Student Response Data For Knowledge. *Proceedings of the AAAI 2005 Educational Data Mining Workshop.*

[6] Beck, J. 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, 2007, 21-30.

[7] Beck, J.E., Chang, K.-m. Identifiability: A Fundamental Problem of Student Modeling. *Proceedings of the 11th International Conference on User Modeling,* 2007.

[8] Cen, H., Koedinger, K.R., Junker, B. Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems,* 2006, 164-175.

[9] Corbett, A.T., Anderson, J.R. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 1995, 4, 253-278.

[10] Feng, M., Heffernan, N.T., Koedinger, K.R. Looking for Sources of Error in Predicting Student's Knowledge. *Educational Data Mining: Papers from the 2005 AAAI Workshop,* 54-61.

[11] Fogarty, J., Baker, R., Hudson, S. Case Studies in the use of ROC Curve Analysis for Sensor-Based Estimates in Human Computer Interaction. *Proceedings of Graphics Interface (GI 2005)*, 129-136.

[12] Hanley, J.A., McNeil, B.J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology,* 1982, 143, 29-36.

[13] Koedinger, K. R., Corbett, A. T. Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences*, 2006, pp. 61-77. New York, NY: Cambridge University Press.

[14] Reye, J. Student Modeling based on Belief Networks. *International Journal of Artificial Intelligence in Education*, 2004, 14, 1-33.

[15] Rosenthal, R., Rosnow, R.L. *Essentials of Behavioral Research: Methods and Data Analysis*, 1991. Boston: McGraw-Hill.

[16] Witten, I.H., Frank, E. *Data Mining: Practical machine learning tools and techniques,* 2005. San Francisco: Morgan Kaufmann.