Incorporating Scaffolding and Tutor Context into Bayesian Knowledge Tracing to Predict Inquiry Skill Acquisition

Michael A. Sao Pedro
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609 USA
(508) 831-5000
mikesp@wpi.edu

Ryan S.J.d. Baker
Teacher's College, Columbia
525 W. 120th St., Box 118
New York, NY 10027 USA
(212) 678-8329
ryan@educationaldatamining.org

Janice D. Gobert
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609 USA
(508) 831-5000
jgobert@wpi.edu

ABSTRACT
In this paper, we incorporate scaffolding and change of tutor context within the Bayesian Knowledge Tracing (BKT) framework to track students' developing inquiry skills. These skills are demonstrated as students experiment within interactive simulations for two science topics. Our aim is twofold. First, we desire to improve the models' predictive performance by adding these factors. Second, we aim to interpret these extended models to reveal if our scaffolding approach is effective, and if inquiry skills transfer across the topics. We found that incorporating scaffolding yielded better predictions of individual students' performance over the classic BKT model. By interpreting our models, we found that scaffolding appears to be effective at helping students acquire these skills, and that the skills transfer across topics.

Keywords
Science Microworlds, Science Simulations, Science Inquiry, Automated Inquiry Assessment, Educational Data Mining, Validation, Bayesian Knowledge Tracing, User Modeling

1. INTRODUCTION
Many extensions to the classic Bayesian Knowledge Tracing (BKT) model [1] have been developed to improve performance at predicting skill within intelligent tutoring systems, and to increase the interpretability of the model. For example, extensions have been made to account for individual student differences [2, 3], to incorporate item difficulty [4], to address learning activities requiring multiple skills [5], and even to incorporate the effects of automated support given by the system [6-8]. Extensions have also been added to increase model interpretability and to provide insight about tutor effectiveness. For example, [6] incorporated scaffolding into BKT to determine if automated support improved students' learning and performance. However, taking into account the differences in tutor contexts, the different facets of an activity or problem in which the same skills are applied, has only been studied in a limited fashion ([8] is one of the few examples). Context is important to consider because skills learned or practiced in one context may not transfer to new contexts [9], [10]. This, in turn, could reduce a model's predictive performance if it is to be used across contexts. Explicitly considering the context in which skills are applied within knowledge modeling may also increase model interpretability and potentially reveal whether some skills are more generalizable, and thus transferrable.

In this paper, we explore the impacts of incorporating two new elements to the BKT framework to track data collection inquiry skills [cf. 11] within the Inq-ITS inquiry learning environment [12]. These elements are scaffolding and change of tutor context. Like [6-8], we incorporate scaffolding by adding an observable and model parameters to account for its potential impacts on learning. We also add parameters and observables to account for change in the tutor context. In this work, we focus on one kind of tutor context, the specific science topic in which students practice and demonstrate inquiry skills. Predictive performance and interpretability of these extensions is addressed using data gathered from students who engaged in inquiry learning within scaffolded Inq-ITS activities on two Physical Science topics [12].

These proposed extensions are motivated by our prior work [13], [14] in constructing BKT models to track skills within an *unscaffolded* activities on a single science topic. Though these models could predict students' performance, we noticed they had very low learning rate parameters. Since then, we added scaffolding to these activities that automatically provides feedback to students when they engage in unproductive data collection. By incorporating scaffolding into our BKT models, we aim to improve prediction and to determine the degree to which scaffolding impacts skill acquisition. In other words, to paraphrase Beck et al. [6], we want to know "Does our help help?" Explicitly modeling this improvement may enhance the learning environment's ability to predict performance. In particular, if the scaffolding we provided is effective, we expect that learning rate should increase when students receive help by the system [cf. 6].

Similarly, the science topic (the context) in which skills are enacted may also play a role in models' predictive capabilities. Specifically, it is possible that inquiry skills may be tied to the science topic in which they are learned [15]. In other words, students who practice and learn inquiry skills in one science topic may not be successful at transferring skill to other science topics [cf. 9, 10]. Thus, from the viewpoint of predicting student performance, changing topics may reduce the success of a standard BKT model at predicting future performance. By explicitly modeling this, we may be able to improve models' predictive capabilities. In addition, by explicitly incorporating the science topic into models, it may become possible to discern from the model parameters the degree to which inquiry skills transfer.

2. INQ-ITS LEARNING ENVIRONMENT
We developed our models to track students' scientific inquiry skills within the Inq-ITS learning environment (www.inq-its.org), formerly known as Science Assistments [12]. This environment aims to automatically assess and scaffold students as they experiment with interactive simulations across several science topic areas such as Physical, Life, and Earth Science. Each Inq-ITS activity is a performance assessment of a range of inquiry skills; the actions students take within the simulation and work products they create are the bases for assessment.

Inq-ITS inquiry activities all have a similar look-and-feel. Each activity provides students with a driving question, and requires

them to conduct an investigation with a simulation and inquiry support tools to address that question. These inquiry support tools include a hypothesizing widget, a data analysis widget, and graphs and tables for automatically displaying and summarizing data. The tools not only help students explore and keep track of their progress, but also enable assessment because they make students' thinking explicit [12].

The system also delivers scaffolds to students in a text-based format via a pedagogical agent named Rex, a cartoon dinosaur, shown in Figure 1. Primarily, Rex provides real-time feedback to students as they engage in inquiry. In other words, the system can "jump in" and support students as they work. Determination of who receives scaffolding is performed using both EDM-based detectors and knowledge engineered rules [12]. We will elaborate on these approaches to evaluate the data collection skills relevant to this paper in Section 4.

In this paper, we focus on tracking skills across two Physical Science Topics, Phase Change and Free Fall. We now present an overview of the inquiry activities pertaining to these topics.

2.1 Phase Change and Free Fall Activities

The Phase Change activities [12] (Figure 1) foster understanding about the melting and boiling properties of ice. In these activities, students are given an explicit goal to determine if one of three factors (size of a container, amount of ice to melt, and amount of heat applied to the ice) affects various measurable outcomes (e.g. melting or boiling point). Students then formulate hypotheses, collect and interpret data, and warrant their claims to address the goal. The inquiry process begins by having students articulate a hypothesis to test using a hypothesis widget [12]. The widget is set of pulldown menus that provide a template of a hypothesis. For example, a student may state: "If I change the container size so that it decreases, the time to melt increases."

After stating a hypothesis, students then "experiment" by collecting data to test their hypotheses (see Figure 1). Here, students are shown the Phase Change simulation and graphs that track changes of the ice's temperature over time. Students change the simulation's variables, and then run, pause and reset it to collect their data (trials). A data table tool is also present that shows all the data collected thus far.

Once students decide they collected enough data, they move to the final task, "analyze data". Similar to hypothesizing, students use pull-down menus to construct an argument whether their hypotheses were supported based on the data they collected [12].

The second set of activities we developed, the Free Fall activities [11], are similar to the Phase Change activities. These activities aim to foster understanding about factors that influence the kinetic, potential and mechanical energy of a ball when it is dropped. In these activities, students again try to address a driving question related to Free Fall by conducting an investigation.

As students collect data, they can receive real-time feedback from Rex (if feedback is turned on), as soon as the system detects they are not engaging in productive data collection (Figure 1). For example, if the system detects that a student is designing controlled experiments but is not collecting data to test their hypothesis (two skills associated with good data collection [12]), Rex will tell them "It looks like you did great at designing a controlled experiment, but let me remind you to collect data to help your test your hypotheses." If the student continues struggling, "bottom-out" feedback is given [cf. 1]: "Let me help some more. Just change the [IV] and run another trial. Don't change the other variables. Doing this lets you tell for sure if changing the [IV] causes changes to the [DV]" ([IV] and [DV] are replaced with the student's exact hypothesis) Thus, Rex's scaffolds provide multi-level help, with each level providing more specific, targeted feedback when the same error is made repeatedly, similar to Cognitive Tutors [e.g. 1]. A goal of this paper is to gain insight about the efficacy of this scaffolding approach.

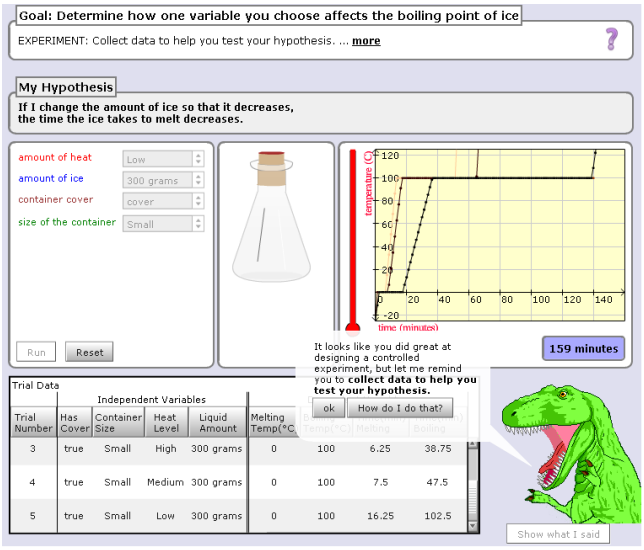

Figure 1. Pedagogical Agent Rex automatically provides support to students as they experiment with a Phase Change simulation in the Inq-ITS learning environment.

3. PARTICIPANTS AND PROCEDURE

We collected data from 299 eighth grade students as they engaged in inquiry within Inq-ITS. These students attended three different schools in suburban Central Massachusetts. Students at each school had the same teacher, and were separated into class sections. Some had prior experience conducting inquiry in Inq-ITS, and for others, this was their first experience.

These data were collected as part of a study to determine the impacts of automated scaffolding on acquisition and transfer of data collection skills across science topics. In this study, students were assigned 5 Phase Change inquiry activities, and two weeks later, 5 Free Fall activities. Students were allotted approximately two class periods per science topic to complete the activities. Due to time constraints, some students did not finish all the activities in each science topic.

Recall that in each activity, students formulated hypotheses, collected data and analyzed data. In the first 4 Phase Change activities, all students had scaffolding available as they formulated hypotheses. However, some students were randomly chosen to have data collection scaffolds available, whereas others did not. In the scaffolding condition, Rex (Figure 1) provided feedback to the students when they were evaluated as not demonstrating good data collection behavior. Students who were in the no-scaffolding condition received no feedback on their data collection. In the "analyze data" inquiry task, no students received scaffolding.

Both groups then completed a fifth Phase Change activity with no scaffolding. This enabled us to measure immediate impacts of scaffolding on skill acquisition within the same science topic. Approximately two weeks later, all students engaged in inquiry

within the Free Fall activities. Students did not receive any feedback on their data collection within these activities. These activities were used to determine the impacts of scaffolding on transfer of skill across science topics.

4. EVALUATING THE DEMONSTRATION OF DATA COLLECTION SKILL

Within this work, we used automated methods for evaluating data collection skills [13, 16, 17]. This evaluation was used both to trigger scaffolding, and to provide observables of student performance for building Bayesian Knowledge Tracing models. Specifically, we aim to assess two process skills associated with productive data collection, designing controlled experiments and testing stated hypotheses [12]. These are demonstrated as students collect data using the simulation in the "experiment" stage of inquiry. Briefly, students design controlled experiments when they generate data that make it possible to determine what the effects of independent variables (factors) are on outcomes. They test stated hypotheses when they generate data that can support or refute an explicitly stated hypothesis. These skills are separable; students may test their hypotheses with confounded designs, or may design controlled experiments for a hypothesis not explicitly stated. Since these are process skills, students are assessed based on the actions they take while collecting data.

We evaluate whether students demonstrate these skills by combining predictions made by data-mined detectors [13], with knowledge-engineered rules to handle specific edge cases. This process works briefly as follows. Detectors were constructed by applying machine learning to predict labels of student skill. These labels were generated using text replay tagging on log files [18] from students' interactions within the Phase Change activities. In this process, a human coder labels whether or not students are demonstrating the inquiry skills by viewing a "chunk" of student actions (the text replay) that has been formatted to highlight relevant information to make that coding easier. These labels can be used as "ground truth" for whether or not students demonstrate a skill, and subsequently for building and validating detectors that replicate human judgment.

To validate our detectors, we tested their predictive performance against *held-out test sets* of student data, data not used to construct the detectors. It is important to note that the students considered in this paper were not used to build the detectors. Performance was measured using A' [19] and Kappa (κ). Briefly, A' is the probability that when given two examples of students' data collection, one labeled as demonstrating skill and one not, a detector will correctly label the two. A' is identical to the Wilcoxon statistic, and approximates the area under the ROC curve [19]. A' of 0.5 indicates chance-level performance, 1.0 indicates perfect performance. Cohen's Kappa (κ) determines the degree to which the detector matches raw human judgment, with $\kappa = 0.0$ indicating chance-level performance and $\kappa = 1.0$ indicating perfect performance.

Using this validation process, we demonstrated that our detectors of can be used to evaluate students' inquiry in Phase Change when they complete their experimentation [17]. More specifically, the designing controlled experiments detectors work well when students have run the simulation at least three times (thus collecting three pieces of data) in their experimentation. For data collections of this type, the detectors can distinguish a student who has designed controlled experiments when they have completed their data collection from a student who has not A' = 94% of the time. They also could identify the correct class extremely well, $\kappa = .75$. The testing stated hypotheses detector also predicted quite well, without the limitation on the number of trials collected by the student, A' = .91, $\kappa = .70$.

We also found that these detectors could also be used as-is to drive scaffolding in Phase Change [17], *before* students finished collecting their data. The designing controlled experiments detector could successfully be applied by the student's third data collection with the simulation, and the testing stated hypotheses detector could be applied in as few as two simulation runs.

Finally, these detectors have been shown to generalize to evaluate skill within the Free Fall activities [11], a different science topic from which they were built (Phase Change), and an entirely different cohort of students. Under student-level stratification, the designing controlled experiments detector could distinguish a student who designed controlled experiments from one who did not A' = 90% of the time, and highly agreed with a human coder's ratings, $\kappa = .65$. Performance for the testing stated hypotheses detector was also high, A' = .91, $\kappa = .62$.

As mentioned, though performance of these detectors is quite good for evaluation of data collection skill and for driving scaffolding, there are edge cases where the detectors did not perform as well. In particular, the designing controlled experiments detector cannot be applied when students collect only 1 or 2 pieces of data with the simulation. The testing stated hypothesis detector cannot be applied when the student collects only a single trial. In these cases, which are well-defined, we authored simple knowledge engineered rules to evaluate students' data collection for a single trial [20] and two trials [21, 22].

Thus overall, combining data mining and knowledge engineering enabled successful evaluation of students' data collection process skills. In the next section, we describe the data distilled from students' usage of the Phase Change and Free Fall activities. These data are used to develop and test the BKT extensions.

5. DATASET FOR BKT MODELS

Students' skill demonstration was evaluated by the detectors and knowledge engineered rules outlined in Section 4. A full profile of student performances was generated for each skill and each activity. These evaluations are the observations used to build BKT models of latent skill.

Certain students and evaluations were removed. First, we only consider students' first opportunity to demonstrate skill prior to receiving scaffolding. More specifically, students can continue to collect data after they receive scaffolding, and be re-evaluated. These additional evaluations are not included in the data set. We do this to control for the possibility that specific scaffolds in our multi-level scaffolding approach may differentially impact learning. Thus, we look for the overall effects of scaffolding. Second, we removed 12 students who did not complete both the Phase Change and Free Fall activities due to absence. The final dataset contained 5878 unique evaluations of 287 students' inquiry, 2939 evaluations for each data collection skill.

6. EXTENSIONS TO BKT

We amalgamated students' performances across activities within a Bayesian Knowledge-Tracing framework [1]. BKT is a two-state Hidden Markov Model that estimates the probability a student possesses latent skill (L_n) after n observable practice opportunities ($Prac_n$). In our work, latent skill is knowing how to perform the data collection skills, and a practice opportunity is an evaluation of whether skill was demonstrated during data collection in an inquiry activity. A practice opportunity begins when students

enter the "experiment" task in an inquiry activity. An opportunity ends when a student switches from the "experiment" task to the 'analyze data" task (see Section 2.1). As mentioned, the detectors / knowledge engineered rules evaluate students' actions, and these evaluations act as the observables. A student is evaluated as not having demonstrated skill ($Prac_n = 0$) if one of two cases occurs. The first is if they are evaluated as not demonstrating a skill when they signal completion of data collection (e.g. attempt to switch to the "analyze data" task). The second is if, while collecting data, the system believes the student does not know either skill and provides scaffolding. This approach to address scaffolding's impact on student correctness is similar to others [e.g. 4].

The classic BKT model [1] is characterized by four parameters, *G*, *S*, L_0, and *T*. The Guess parameter (*G*) is the probability the student will demonstrate the skill despite not knowing it. Conversely, the Slip parameter (*S*) is the probability the student will not demonstrate the skill even though they know it. L_0 is the initial probability of knowing the skill before any practice. Finally, *T* is the probability of learning the skill between practice attempts. From these values, the likelihood of knowing a skill P(L_n) is computed as follows:

$$P(L_n) = P(L_{n-1}|Prac_n) + \big(1 - P(L_{n-1}|Prac_n)\big) * T, \text{ where}$$

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

$$P(L_{n-1}|Prac_n = 0) = \frac{P(L_{n-1}) * S}{P(L_{n-1}) * S + \big(1 - P(L_{n-1})\big) * (1 - G)}$$

This classic BKT model [1] carries a few assumptions. First, the model assumes that a students' latent knowledge of a skill is binary; either the student knows the skill or does not. The model also assumes one set of parameters per skill and that the parameters are the same for all students. Finally, the classic model assumes that students do not forget a skill once they know it.

Relevant to this work, the classic BKT model does not take into account whether students received any scaffolding from the learning environment [6] and does not account for the topic in which skills are demonstrated [8]. The same skill in different topics would either be treated as two separate skills (assuming no transfer), or as having no differences between topics (assuming complete transfer). Both of these assumptions are thought to be questionable [10, 23]. Below, we describe our approach to incorporate both of these factors.

6.1 Taking Scaffolding into Account

We introduce scaffolding into BKT as an observable, $Scaffolded_n$ = {True, False}, because it can directly be seen if our pedagogical agent provided help to students as they collected data. A similar approach was taken by [6] to develop the Bayesian Evaluation and Assessment model. In their domain, reading, this scaffolding observable was true if a student received help just before reading a word (each word was treated as a skill). The observable was linked to all four BKT model parameters, meaning that scaffolding could have an impact on initial knowledge (L_0), guess (*G*), slip (*S*) and whether or not students learn between practice opportunities (*T*). As a result, their BKT model contained 8 parameters to account for scaffolding.

Unlike [6], we instead chose to condition *only* the learning rate (*T*), for three reasons. First, the increase in the number of parameters could result in overfitting, especially since the classic BKT model is already known to be overparametrized [24]. Second, though the additional parameters may facilitate model interpretation, it is unclear whether conditioning all the classic BKT parameters on scaffolding improves predictive performance. In particular, [6] found no increase in predictive performance when accounting for scaffolding. Finally, the immediate effects of scaffolding on performance may not be relevant because we only look at first practice opportunities (thus looking at overall effects of scaffolding), and because there is a time delay between data collection performance attempts. In particular, students attend to a different inquiry task, analyzing data, after their data collection (see Section 2.1 for more details).

In our extension, conditioning learning on whether students receive scaffolding yields two learning rate parameters, *T_scaffolded* and *T_unscaffolded*. Thus, this model tries to account for the differential impacts scaffolding may have on whether or not students learn a skill (e.g. the latent variable knowledge transitions from "doesn't know" to "know" after practicing). Mathematically, the original equation for computing P(L_n) is conditionalized to account for the observable as follows:

$$P(L_n|Scaffolded_n = True) = P(L_{n-1}|Prac_n) + \big(1 - P(L_{n-1}|Prac_n)\big) * P\big(T_{scaff}\big)$$

$$P(L_n|Scaffolded_n = False) = P(L_{n-1}|Prac_n) + \big(1 - P(L_{n-1}|Prac_n)\big) * P\big(T_{unscaff}\big)$$

6.2 Taking Science Topic (Context) into Account

We also developed BKT extensions to the take into account the science topic in which students demonstrate their inquiry skills. Recall that students first practiced inquiry in Phase Change activities (possibly scaffolded or unscaffolded) and then practiced inquiry in unscaffolded Free Fall activities, a different science topic. As mentioned, modeling the change in science topic is of important since the degree to which inquiry skills transfer across topics is unclear [15]. We hypothesize that incorporating the change of science topic into our BKT framework may improve models' predictive performance.

We incorporate changing of science topics in two ways. First, we hypothesized that there may be a differential effect in learning between topics. For example, practice in Phase Change may prepare students to learn (and subsequently demonstrate) skills in Free Fall, called "preparation for future learning" [23]. To model differential learning between topics, we again break out the learning rate (*T*), this time for each topic: *T_PhCh*, *T_FF*. A new observable is also added for the current science topic, $Topic_n$ = {PhaseChange, FreeFall}. The result is a "BKT learn rate topic" model with a modification to the P(L_n) equation similar to the "scaffolded BKT model" described previously.

Our second model for incorporating the change of science topics posits that students may not understand that the skills are applicable across topics. We model this notion by adding in a linear degradation factor, $k \in (0,1)$, to potentially offset the likelihood students know the skill P(L_n) when the science topic switches. If $k = 1$ this implies there is no effect on students' knowledge when the topic switches. When $k = 0$, students will be presumed to not know the skill when the topic switches. One benefit of this approach is that it relaxes the assumption of skill independence if we had chosen to fit separate classic BKT models per skill, per science topic. Instead, *k* captures the potential for partial transfer of skill between science topics [cf. 10]. We also add an observable $Topic\_Switch_n$ = {True, False} to address when

the science topic changes from Phase Change to Free Fall (just before the student's first opportunity to practice in Free Fall). The corresponding P(L_n) modification for the "BKT skill degradation model" is:

$$P(L_n|Topic\_Switch_n = True) = k * \left[P(L_{n-1}|Prac_n) + \left(1 - P(L_{n-1}|Prac_n)\right) * T\right]$$

$$P(L_n|Topic\_Switch_n = False) = P(L_{n-1}|Prac_n) + \left(1 - P(L_{n-1}|Prac_n)\right) * T$$

Note that the degradation parameter k is different than modeling "forgetting" in the BKT framework [cf. 1, 8] in two ways. First, we note that the factor is applied to both conditional expressions in the P(L_n) equation, not just P($L_{n-1}|Prac_n$) as done when modeling forgetting. Second, in these earlier approaches forgetting is modeled at each practice opportunity, whereas our factor is applied at a single point, when the science topic switches.

6.3 Combining Models

The above models introduce three new potential observables to the BKT framework relevant to our learning environment: *Scaffolded?* = {True, False}, *Topic* = {PhaseChange, FreeFall}, and *Topic_Switch?* = {True, False}. The models above individually incorporate the observables by conditioning the learning rate parameter, *T*, on them, or by adding a multiplicative reduction factor, *k*, to the computation of P(L_n). As part of this work, we also combined the extensions described above into larger models. The most complicated model incorporated all observables and contained seven parameters: (L_0, *G*, *S*, *T_Scaff_PhCh*, *T_Unscaff_PhCh*, *T_Unscaff_FF*, *k*). We next describe our process for fitting these models.

6.4 Model Fitting

As in [3], [13], we use brute force search to find the best fitting parameters. This method has been found to produce comparable or better model parameters than other methods [25]. In this approach, all potential parameter combinations in the search space are tried at a grain-size of 0.01. The best parameter set yields the lowest sum of squares residual (SSR) between the likelihood that the student would demonstrate skill, P($Show\_Skill_n$), and the actual data. This likelihood is computed as follows [1]:

$$P(Show\_skill_n) = P(L_{n-1}) * (1 - S) + (1 - P(L_{n-1})) * G$$

Once this set has been found, another brute force search around those parameters is run at a grain-size of 0.001 to find a tighter fit. We bound *G* to be less than 0.3 and *S* to be less than 0.1 [cf. 25]; all other parameters can be assigned values in (0.0, 1.0).

When fitting our models, we found the brute force search to be realistically tractable only up to fitting 5 parameter models. To fit the combined models with more parameters, we used a two-stage process. First, we fit a classic BKT model with four parameters (L_0, *G*, *S*, *T*). Then, we fit a combined model using fixed values for *G* and *S* from the classic model. These parameters were fixed because we believe the extended models described above will have the most impact on estimates of learning between practice opportunities and initial knowledge, not on guessing and slipping.

7. RESULTS

We determine if extending the classic BKT model to include scaffolding and changing of science topics will 1) improve predictions of future student performance in our learning environment, and 2) yield insights about the effectiveness of our scaffolding approach, and the transferability of the inquiry skills.

To address predictive performance, we determined if the new models' predictions of skill demonstration P($Show\_Skill_n$), aggregated from evidence over times {1…*n*-1}, can predict actual student performance at time *n* better than the classic BKT model. We train and test our models' performance by conducting six-fold student-level cross-validation, stratifying by both learning condition (having scaffolding available in Phase Change or not) and class section. Cross-validating in this way helps ensure that each fold equally represents learning conditions, and students from each class section/school. This increases assurance that models can be applied to new students.

As in [13], model goodness was determined using A' [19]. This is an appropriate metric to use when the predicted value is binary (either students demonstrated skill in $Prac_n$ or they did not), and the predictors for each model are real-valued, e.g. P($Show\_skill_n$). As a reminder, a model with A' of 0.5 predicts at chance level and a model with A' of 1.0 predicts perfectly.

Two variants on A' for student performance data are computed as follows. First, we compute overall A' values of each model collapsing over students as we did in [13]. Second, we compute the A' values of each model per student [3], and report the average per-student A'. These approaches have different strengths and weaknesses [cf. 3, 13, 25]. Collapsing over students is straightforward and enables comparison of models' broad consistency in predicting skill demonstration. In other words, this approach can show, in general, whether or not high likelihoods of demonstration of skill predicted by the model correspond with actual demonstration of skill. In addition, collapsing can be used when there is not enough within variance for each student to produce a meaningful per student A' [cf. 13]. Collapsing over students, however, provides weaker estimates of predicting an individual student's learning and performance than the A' per student metric [3, 25]. Collapsing may also yield estimates that are biased towards students who practiced more with the system since they contribute more data [25].

Only used students who had variation in their evaluations were used when computing A' per student. In other words, a student was not considered if they were evaluated correct on all practice opportunities or incorrect on all practice opportunities. This was necessary because A' is undefined unless there is at least one 'positive', and at least one 'negative' evaluation for a student [19]. As a result, 175 students remained for designing controlled experiments and 132 students for testing stated hypotheses.

We ascertain whether any BKT model variant outperforms the classic model by comparing A' values computed under the cross-validation scheme described. These results are described next.

7.1 Models' Overall Predictive Capability

As shown in Table 1, all of the models show strong consistency, meaning that high estimates of skill demonstration are associated with actual demonstration of skill. This is evidenced by collapsed A' values ranging from .817 to .837 for the designing controlled experiments skill, and collapsed A' values ranging from .840 to .853 for the testing stated hypotheses skill. Recall that these high collapsed A' values do not reflect the models' ability to predict individual student trajectories [25], because they factor out the student term. The model with the highest A' = .837 for predicting future performance of the designing controlled experiments skill 1) conditioned the learning rate on whether the student received scaffolding (T_Scaffolded extension), and 2) incorporated skill degradation when switching between science topics (kLn_TopicSwitch extension). This represents a small increase in

Table 1. BKT model variant performance predicting whether students will demonstrate skill in their next practice attempt in the learning environment. The A' values were computed under six-fold student-level cross-validation Overall, the best model for both skills is the one in which the learning rate is conditioned on whether or not the student received scaffolding during Phase Change (*T_Scaffolded*).

BKT Model Variant			Designing Controlled Experiments		Testing Stated Hypotheses	
T_Scaffolded	T_Topic	kLn_TopicSwitch	A' per student avg[a]	A' collapsed[b]	A' per student avg[a]	A' collapsed[c]
X			.685	.827	.656	.846
	X		.633	.818	.610	.840
		X	.641	.825	.612	.844
X	X		.678	.829	.648	.848
	X	X	.630	.826	.601	.845
X		X	.680	.837	.638	.852
X	X	X	.676	.836	.645	.853
		Classic BKT:	.635	.817	.613	.841

[a] $N = 287$ students; [b] $N = 175$ students; [c] $N = 132$ students

performance over the classic BKT model (A' = .817). The model with the highest A' = .853 for predicting future performance of testing stated hypotheses was the full model that incorporated all three extensions. This again was a small improvement over the classic BKT model (A' = .841).

In terms of predicting individual student performance, some of the models performed reasonably well. As a baseline, the Classic BKT model for designing controlled experiments had a per-student average A' = .635. For testing stated hypotheses, the Classic BKT model had a per-student average A' = .613. These values, though above chance A' (.5), are somewhat low.

When incorporating some of the BKT variants, the per-student average A' increased. In particular, BKT variants that leveraged conditioning on scaffolding (T_Scaffolded model) performed better than the Classic BKT model (Table 1). For example, the best BKT model variant for both skills incorporated only scaffolding. The per-student average A' of this model for designing controlled experiments was .685, a jump over the Classic BKT model. The per-student average A' for testing stated hypotheses was .656, and again, outperformed the Classic BKT model. These A' values are on par with the extended BKT models developed in [6] that incorporated scaffolding.

7.2 Model Interpretation

Like [6], we interpreted the models' parameters to understand what they reveal about the impacts of scaffolding and the learning and transfer of scientific inquiry skills between Physical Science topics. Since the full models with 7 parameters had A' performance on par with the other best performing models, we chose to interpret their parameters. The parameter averages and standard deviations for each skill model across all six folds are presented in Table 2. We focus on interpreting the new parameters we added to the model.

In Phase Change, the learning rate when students were scaffolded is much higher than the learning rate without scaffolding, *T_Scaff_PhCh* = .638 vs. *T_UnScaff_PhCh* = .190 for designing controlled experiments, and *T_Scaff_PhCh* = .823 vs. *T_UnScaff_PhCh* = .158 for testing stated hypotheses. These values indicate that scaffolding students' inquiry appears to have a positive effect on whether students learn the skills [6].

The learning rate for the Free Fall activities, which were unscaffolded and practiced after the Phase Change activities, was comparatively lower for each skill, *T_UnScaff_FF* = .094 for designing controlled experiments, and *T_UnScaff_FF* = .089 for testing stated hypotheses. The meaning of these values is more difficult to discern because all students had prior opportunity to practice in Phase Change before attempting the Free Fall tasks. It could be that the unscaffolded Free Fall activities, like the unscaffolded Phase Change activities, are less effective for helping students acquire these inquiry skills. However, it could also be that the lower learning rates reflect that many students already mastered the skills in Phase Change and thus these new activities afforded no additional learning opportunities. We believe the latter to be the case because 1) more than 85% of students demonstrated each skill in their first Free Fall practice opportunity (data not presented in this paper), and 2) the initial likelihood of knowing the skills (L_0) was high.

Finally, the skill degradation parameter *k*, which captures the degree of skill transfer between science topic (0 is no transfer, 1 is full transfer), was high for both skills. For designing controlled experiments, *k* = .973 and for testing stated hypotheses, *k* = .961. These high values suggest that skill transfers from Phase Change to Free Fall within our learning environment [cf. 15]. We elaborate on this finding in more detail in the next section.

8. DISCUSSION AND CONCLUSIONS

In the classic Bayesian Knowledge Tracing framework [1], scaffolding and the tutor context, the nature of the activities in which skills are applied, are not taken into account when predicting students' future performance. Similar to others' prior work [6-8] we explored here whether extending the BKT framework to incorporate these factors improves prediction of students' skill demonstration. This work was conducted to predict students' acquisition of two data collection inquiry skills, designing controlled experiments and testing stated hypotheses [cf. 12, 13], in performance-based inquiry tasks across two Physical Science topics, Phase Change and Free Fall. Specifically, we added three extensions to the BKT model: 1) conditioning the learning rate on whether or not students were scaffolded; 2) conditioning the learning rate depending on the topic in which students practiced inquiry (Phase Change or Free Fall); and 3) adding a degradation parameter to potentially lower the likelihood

Table 2. Means and standard deviations of the parameter values for full BKT model variant across all six folds.

	Full BKT Model Parameters						
Skill	L_0	G	S	$T\_UnScaff\_PhCh$	T_Scaff_PhCh	$T\_UnScaff\_FF$	k
Designing Controlled Experiments	.470 (0.014)	.196 (0.029)	.050 (0.006)	.190 (0.018)	.638 (0.035)	.094 (0.010)	.973 (0.006)
Testing Stated Hypotheses	.602 (0.026)	.198 (0.023)	.042 (0.007)	.158 (0.026)	.823 (0.057)	.089 (0.011)	.961 (0.009)

of a student knowing a skill when the science topic changed. Overall, we found that BKT can track development of both skills, in accordance with our prior work [13], and that our extensions led to improvements in prediction and model interpretability.

In comparing our BKT extension that incorporates scaffolding, our approach is closest to the one taken in [6]. Our model assumes that scaffolding will *only* impact learning, whereas [6] captures that scaffolding may differentially impact learning *and* immediate performance. Our modeling choice was motivated in part by parsimony given that BKT is already overparametrized [24], a possibility hypothesized in [6], and by the delay between performance attempts of the skills in our learning environment. Unlike [6], we found that taking scaffolding into account improved the ability to predict individual student learning and performance over the classic BKT model, possibly due to increased parsimony. We also teased apart the effects of scaffolding on our models' predictive abilities overall (collapsing over students) and on predicting individual student performance.

When interpreting the parameters of the extended model, we found that scaffolding appears to have a positive impact on learning, as in [6]. We do note, though, that we did not tease out the differential impacts of specific scaffolds in our multi-level scaffolding approach. It is possible that specific scaffolds trigger different degrees of learning. One possible way to incorporate this is to condition learning rate on the different kinds of scaffolds, not just whether or not students received scaffolding in general.

We also incorporated parameters to account for the possible effects of demonstrating inquiry skill within different science topics (Phase Change and Free Fall). This modeling was inspired by the empirical question of whether inquiry skills are tied to the science topic in which they are learned [15], or if they transfer across topics [9, 10]. Though incorporating these parameters did not increase the predictive performance of our models, they do provide possible insights to inquiry learning. In particular, the model parameters suggest that the data collection skills of interest transfer across science topics, which supports earlier findings [e.g. 20, 26, 14]. There are limits to how certain we can be about this interpretation, though. First, in our study design, we only randomized whether students received scaffolding in Phase Change, and then measured transfer to Free Fall. A stronger approach to increase parameter interpretability would be to also randomize the science topic order. Second, it is possible that the implied transfer of skill may be due to the structural similarities of the activities [9] across Physical Science tasks. In the future, it will be beneficial to conduct a similar study across different science topic areas, like Life and Earth Science [12], with different activity structures to tease apart these possible effects.

This paper offers three contributions. First, to our knowledge, this work is the first application of BKT to track the development of inquiry process skills across science topics. This work strengthens our earlier findings in using BKT for a single group of students and single topic [13], because we cross-validated our models with students from multiple schools who engaged in two science topics. Second, we extended BKT by incorporating scaffolding. Though this extension is similar to others' [e.g. 6, 8], it enabled a "discovery with models" analysis [cf. 27] that shed light on the potential relationships between performance in the environment, scaffolding, and transfer of inquiry skills [15]. Furthermore, conditioning the BKT learning rate on whether students received scaffolding improved prediction of individual students' trajectories over the classic model. Finally, we incorporated a form of tutor context (the science topic in which skills were demonstrated) directly in the BKT model, unlike [8], which addressed context by selecting subsets of training and testing data. By adding these additional parameters, we discerned that the data collection skills transferred across the two science topics by interpreting the extended BKT model.

In closing, we note that this work focuses primarily on validation and interpretation of skill *within our learning environment.* In our prior work [13], we also showed that BKT models not only had this internal reliability, but were also moderately predictive of other measures of inquiry. In the future, we will determine if our model extensions can also improve external validation, thus realizing the full potential of using our learning environment to estimate and track authentic inquiry skills.

9. ACKNOWLEDGMENTS

This research is funded by the National Science Foundation (NSF-DRL#0733286, NSF-DRL#1008649, and NSF-DGE#0742503) and the U.S. Department of Education (R305A090170 and R305A120778). Any opinions expressed are those of the authors and do not necessarily reflect those of the funding agencies. Special thanks are also given to Joseph Beck for advice and feedback on this work.

10. REFERENCES

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

2 Pardos, Z.A. and Heffernan, N.T. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In *Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization* (Big Island, HI 2010), 255-266.

3 Baker, R.S.J.d, Corbett, A.T., Gowda, S.M. et al. Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. In *Proceedings of the 18th Annual Conference on User Modeling, Adaptation and Personalization* (Big Island of Hawaii, HI 2010), 52-63.

4 Pardos, Z.A. and Heffernan, N.T. KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In *Proceedings of*

the 19th International Conference on User Modeling, Adaptation and Personalization (Girona, Spain 2011), 243-254.

5 Koedinger, K.R., Pavlik, P.I., Stamper, J., Nixon, T., and Ritter, S. Avoiding Problem Selection Thrashing with Conjunctive Knowledge Tracing. In *Proceedings of the 3rd International Conference on Educational Data Mining (EDM 2010)* (Pittsburgh, PA 2010), 91-100.

6 Beck, J.E., Chang, K., Mostow, J., and Corbett, A. Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (Montreal, QC 2008), 383-394.

7 Jonsson, A., Johns, J., Mehranian, H. et al. Evaluating the Feasbility of Learning Student Models from Data. In *AAAI05 Workshop on Educational Data Mining* (Pittsburgh, PA 2005), 1-6.

8 Yudelson, M., Medvedeva, O., and Crowley, R. A Multifactor Approach to Student Model Evaluation. *User Modeling and User-Adapted Interaction*, 18 (2008), 349-382.

9 Thorndike, E.L. and Woodworth, R.S. The Influence of Improvement in One Mental Function Upon the Efficacy of Other Functions. *Psychological Review*, 8 (1901), 247-261.

10 Singley, M. and Anderson, J.R. *The Transfer of Cognitive Skill*. Harvard University Press, Cambridge, MA, 1989.

11 NATIONAL RESEARCH COUNCIL. *National Science Education Standards*. National Science Education Standards, Washington, D.C., 1996.

12 Gobert, J., Sao Pedro, M., Baker, R., Toto, E., and Montalvo, O. Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, 4, 1 (2012), 111-143.

13 Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J.D., Montalvo, O., and Nakama, A. Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction*, 23 (2013), 1-39.

14 Sao Pedro, M., Gobert, J., and Baker, R. Assessing the Learning and Transfer of Data Collection Inquiry Skills Using Educational Data Mining on Students' Log Files. In *Paper presented at The Annual Meeting of the American Educational Research Association.* (Vancouver, BC, Canada 2012), Retrieved April 15, 2012, from the AERA Online Paper Repository.

15 van Joolingen, W.R., de Jong, T., and Dimitrakopoulout, A. Issues in Computer Supported Inquiry Learning in Science. *Journal of Computer Assisted Learning*, 23, 2 (2007), 111-119.

16 Sao Pedro, M.A., Baker, R.S.J.d., and Gobert, J.D.. In *Proceedings of the 3rd Conference on Learning Analytics and Knowledge* (Leuven, Belgium 2013), 190-194.

17 Sao Pedro, M., Baker, R., and Gobert, J. Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. In *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP 2012)* (Montreal, QC, Canada 2012), 249-260.

18 Baker, R.S.J.d., Corbett, A.T., and Wagner, A.Z. Human Classification of Low-Fidelity Replays of Student Actions. In *Proceedings of the Educational Data Mining Workshop held at the 8th International Conference on Intelligent Tutoring Systems* (Jhongli, Taiwan 2006), 29-36.

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

20 Kuhn, D., Schauble, L., and M., Garcia-Mila. Cross-Domain Development of Scientific Reasoning. *Cognition and Instruction*, 9 (1992), 285-327.

21 Chen, Z. and Klahr, D. All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development*, 70, 5 (1999), 1098-1120.

22 Koedinger, K., Suthers, D., and Forbus, K. Component-Based Construction of a Science Learning Space. *International Journal of Artificial Intelligence in Education (IJAIED)*, 10 (1998), 292-313.

23 Bransford, J.D. and Schwartz, D. L. Rethinking Transfer: A Simple Proposal with Multiple Implications. In *Review of Research in Education, 24*. American Educational Research Association, Washington, D.C., 1999.

24 Beck, J.E. and Chang, K. Identifiability: A Fundamental Problem of Student Modeling. In *Proceedings of the 11th International Conference on User Modeling* (Corfu, Greece 2007), 137-146.

25 Pardos, Z.A., Gowda, S.M., Baker, R.S.J.d., and Heffernan, N.T. The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software. *ACM SIGKDD Explorations*, 13, 2 (2012), 37-44.

26 Glaser, R., Schauble, L., Raghavan, K., and Zeitz, C. Scientific Reasoning Across Different Domains. In DeCorte, E. et al., eds., *Computer-based Learning Environments and Problem-Solving* (Heidelberg, Germany, 1992), 345-371.

27 Baker, R.S.J.d. and Yacef, K. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1, 1 (2009), 3-17.