Identifying Transfer of Inquiry Skills across Physical Science Simulations using Educational Data Mining

Michael Sao Pedro, Worcester Polytechnic Institute, 100 Institute Rd. Worcester MA 01609, mikesp@wpi.edu
Yang Jiang, Luc Paquette, Ryan S. Baker, Teachers College, 525 W. 120th St. New York, NY 10027
Email: yang.jiang@tc.columbia.edu, luc.paquette@gmail.com, ryan@educationaldatamining.org
Janice Gobert, Worcester Polytechnic Institute, 100 Institute Rd. Worcester MA 01609, jgobert@wpi.edu

Abstract: Students conducted inquiry using simulations within a rich learning environment for 4 science topics. By applying educational data mining to students' log data, assessment metrics were generated for two key inqury skills, testing stated hypotheses and designing controlled experiments. Three models were then developed to analyze the transfer of these inquiry skills between science topics. Model one, Classic Bayesian Knowledge Tracing, assumes that either complete transfer of skill occurs or no transfer occurs; model two (BKT-PST), an extension of BKT, assumes partial transfer and tests that assumption; and model three, a variant of BKT-PST, assumes no transfer and tests this assumption. An analysis of models one and two suggest that transfer of these inquiry skills across topics did occur. This work makes contributions to methodological approaches for measuring fine-grained skills using log files, as well as to the literature on the domain-specificity vs. domain-generality of inquiry skills.

Introduction

Science educators and reseachers agree that inquiry skills are critical to science literacy (NRC, 2011; Kuhn, 2005). To cultivate skills, some researchers have developed interactive, computer-based activities like simulations and microworlds (e.g. Quellmalz et al., 2009). A benefit of these activities is that they yield rich log data which can be leveraged for fine-grained performance assessment (Pellegrino et al., 2001; Mislevy et al., 2012). Though promising, assessment is still challenging because inquiry is multi-faceted, and manifests itself over time in complex ways (Williamson et al., 2006). Some are addressing these challenges using Educational Data Mining (EDM) to automatically assess specific skills (e.g. Sao Pedro et al., 2013a; Baker & Clarke-Midura, 2013; Ketelhut et al., 2013), Such techniques have potential to not only provide teachers and students real-time feedback about skill progress, but also to contribute to the field's understanding of inquiry learning.

In this paper, we use existing EDM models for evaluating data collection inquiry skills (Sao Pedro et al., 2012, 2013a) to build new models that identify skill transfer across several science topics. We focus on these skills because they support the development of other sense-making skills such as interpreting data and warranting claims (e.g. Kuhn, 2005), and because students have difficulty with these (de Jong & van Joolingen, 1998). Inquiry skills will be particularly valuable if they can transfer (Thorndike & Woodworth, 1901; Singley & Anderson, 1989), but it has been suggested that skills are tightly tied to the domain in which they are learned (van Joolingen et al., 2007), and thus may not transfer to new topics. However, other researchers have found evidence that inquiry skills can transfer and have a domain-general component (Glaser et al., 1991; Harrison & Schunn, 2004), or that content knowledge and inquiry skills co-develop (Kuhn et al., 1992; Kuhn & Pease, 2008). Though impressive, these studies had relatively small sample sizes and conflated data analysis skills with experimental design skills, skills unpacked in the present study. Our approach builds on our prior research (Sao Pedro et al., 2013c) in which we extended Bayesian Knowledge Tracing (Corbett & Anderson, 1995) to evaluate transfer of two data collection inquiry skills across two science topics. In particular, we address inquiry skill transfer at a larger scale with more students and across more science topics than seen in prior work.

Methodology

Participants

Participants were 299 eighth grade students from five middle schools in suburban Central Massachusetts who conducted inquiry across at least two science topics within Inq-ITS.

Materials: Inq-ITS Learning Environment Physical Science Inquiry Activities

Inq-ITS (**Inq**uiry **I**ntelligent **T**utoring **S**ystem, Gobert et al., 2012) is a web-based virtual science lab environment that automatically assesses students' inquiry skills (NRC, 2011). In this environment, students conduct inquiry with interactive simulations aligned to middle school Physical, Life, and Earth Science content described in the Massachusetts curricular frameworks, and inquiry support tools. In this paper, we focus on inquiry activities for four Physical Science topics: Phase Change (Figure 1), Free Fall Energy (Figure 2), Free Fall Speed, and Liquid Density. In a typical inquiry activity, students are first presented with a driving question.

For example, in a typical activity for Phase Change, students are asked to determine if one factor (e.g. size of container or amount of ice) affects specific outcomes (e.g. boiling point of water). Then, they conduct a semi-structured scientific inquiry process to address the goal: First, they articulate a hypothesis to be tested using a hypothesis widget with pulldown menus. Next, students collect data to try and test their hypothesis with a simulation. Students are required to run at least one trial before continuing. Once they finish running trials, they analyze their data by forming an argument (similar to hypothesizing) and selecting trials as evidence. A key aspect of this system is that activities provide performance assessment metrics on students' inquiry skills. Assessment of inquiry is based on the processes a student follows while experimenting, and the work products s/he creates using the support widgets.

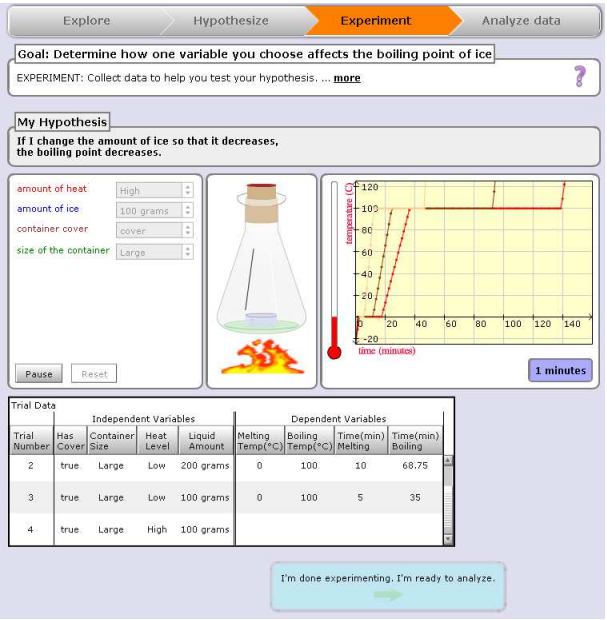

Figure 1. Phase Change simulation

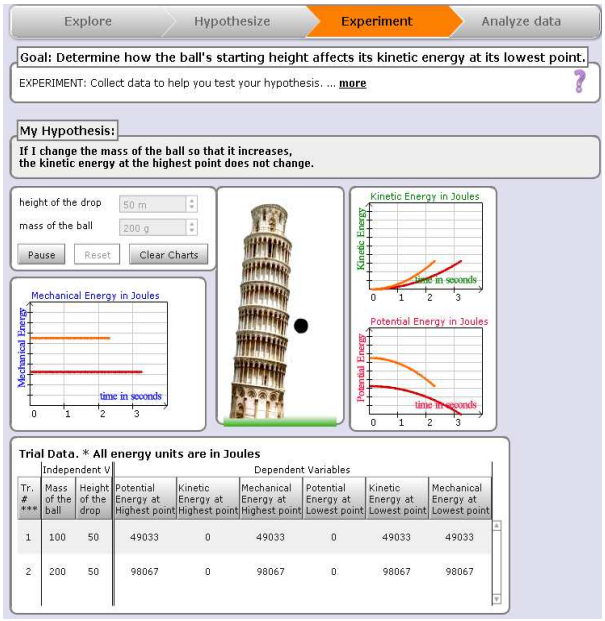

Figure 2. Free Fall Energy simulation

Procedure

Throughout the 2011-2012 school year, students at the five partner schools participated in inquiry within Inq-ITS. We coordinated with teachers regarding which activities would be used and when. Table 1 shows the specific activities chosen by each school and the order they were administered.

Each science topic included between 3 and 5 activities and were administered over two class periods of about 45 minutes each. Over the year, students completed between 2 and 4 sets of activities. The time delay between activity sets varied between schools, according to the respective teacher's pedagogical decisions (see Table 1). For example, at school 4, two science topics were done without any time gap (Free Fall Speed, then Free Fall Energy); at school 5, Free Fall Speed and Free Fall Energy were assigned 3.5 months apart. As students worked, Inq-ITS automatically logged all students' interactions, and automatically assessed their inquiry skills, as described in the next section. Unlike other Inq-ITS activities that provide personalized support (Sao Pedro et al., 2013c), students did not receive any explicit feedback on their inquiry processes or work products in the activities used in this study.

Evaluating Students' Data Collection Skills within Activities

Our work focuses on two data collection skills: designing controlled experiments, and collecting data to test hypotheses (Figures 1 and 2). Students design controlled experiments when they generate trials that make it possible to infer how changeable factors affect outcomes. This skill is related to the Control of Variables Strategy (CVS; cf., Chen & Klahr, 1999) that focuses on creating a single, contrastive and controlled experiment (a single pair of sequential trials). Unlike CVS, designing controlled experiments takes into consideration all a student's trials overall to determine whether a student demonstrates this skill (Sao Pedro et al. 2013a). The second skill, collecting data to test a hypotheses, is demonstrated when a student collects data that can support or refute an explicitly stated hypothesis. We track this in addition to designing controlled experiments because: 1) students may attempt to test their hypotheses with confounded designs, or may design controlled experiments for a hypothesis not explicitly stated; and 2), skill at testing hypotheses may be indicative of a student's successful planning and monitoring of their inquiry (de Jong, 2006).

Our process skills assessment is based on students' actions taken while collecting data with the simulation, and we evaluate whether students design controlled experiments and collect data to test their hypothesis using a combination of data-mined detectors and knowledge-engineered rules (Sao Pedro et al., 2013a,b). Our data mining approach accounts for "corner" cases when students do not conduct their inquiry in lock-step fashion, unlike other approaches that require sequential trials as demonstration of CVS (e.g. McElhaney & Linn, 2010). The goodness and generalizability of data mined detectors also can be determined by testing how well they can predict skill for students who were not used to build the detectors, e.g., we conducted extensive validation tests to show that these detectors agree with expert judgments of inquiry skill performance across our physical science activities (Sao Pedro et al., 2013b,c; Gobert et al., 2013), and new student populations (Sao Pedro et al., 2013c). The detectors are the backbone for generating models of skill transfer across topics, discussed in the next section.

Table 1: Topic order for each school, time delay between activities, and number of participants who conducted inquiry in each pair of topics.

Simulation Topic Pair	School	Delay Between Topics	Number of Participants	Error Rate (% students not demonstrating skill)		
				Last Attempt 1st topic	First Attempt 2nd Topic	Last Attempt 2nd Topic
PhCh → FF Energy	1,2,3	2-3 weeks	140	CtrlExp: 21.4% TestHyp: 21.4%	CtrlExp: 19.3% TestHyp: 14.3%	CtrlExp: 14.3% TestHyp: 15.7%
Density → FF Speed	4	5 weeks	33	CtrlExp: 69.7% TestHyp: 18.2%	CtrlExp: 24.2% TestHyp: 24.2%	CtrlExp: 12.1% TestHyp: 12.1%
FF Speed → FF Energy		no delay	31	CtrlExp: 12.9% TestHyp: 12.9%	CtrlExp: 16.1% TestHyp: 12.9%	CtrlExp: 6.5% TestHyp: 6.5%
FF Energy → PhCh		3 weeks	31	CtrlExp: 12.9% TestHyp: 12.9%	CtrlExp: 9.7% TestHyp: 3.2%	CtrlExp: 9.7% TestHyp: 9.7%
FF Energy → FF Speed	5	14 weeks	64	CtrlExp: 62.5% TestHyp: 57.8%	CtrlExp: 40.6% TestHyp: 39.1%	CtrlExp: 42.2% TestHyp: 42.2%

Developing Models of Transfer to Track Students' Performance across Topics

We model student knowledge and estimate the probability that students are transferring science inquiry skill between topics using Bayesian Knowledge Tracing (BKT, Corbett & Anderson, 1995). BKT is a two-state model (in technical terms, a Hidden Markov Model or simple Dynamic Bayesian Network) that estimates whether student knows a specific latent skill, based on the student's past history of observed performance on that skill. Here, we use BKT to estimate if students know how to design controlled experiments and how to collect data to test a hypothesis (cf. Sao Pedro et al., 2013a,c). The observable performance is whether a student actually demonstrated skill, determined by the detectors discussed previously. BKT has been widely and successfully used to model student knowledge in various intelligent tutoring systems, including the widely-used Cognitive Tutor (Corbett & Anderson, 1995) and ASSISTments systems (Pardos & Heffernan, 2010). BKT performs equivalently to or better than competing approaches (Gowda et al., 2011), and has been extended to support analysis of the nature of student learning (e.g. Beck et al., 2008; Sao Pedro et al., 2013c).

In the classic BKT framework, it is assumed that a skill is either known or not known, and that there is a certain probability of each. Students demonstrate an inquiry skill when (1) they already know the skill and they do not make a slip (a careless mistake); or when (2) they do not know the skill but guess how to do it correctly. The model is defined by a set of four parameters: $P(L_0)$, the probability that the skill is already known before the first opportunity to use it; *T,* the probability that the skill will be learned at each opportunity to use it (classical BKT does not include forgetting, though many extensions do); *G,* the probability that a student will guess and demonstrate the skill despite not knowing it; and *S*, the probability that the student will slip and make a mistake despite knowing the skill. In classical BKT, the four parameters are assumed to be the same for all students (many variants on BKT relax this constraint as well).

Using these parameters, the classic BKT model can incrementally calculate the likelihood $P(L_n)$ that a student knows a skill, such as how to design controlled experiments, after the student finishes their n^{th} attempt practicing the skill ($Prac_n$) in an inquiry activity. It can also estimate the likelihood that a student will demonstrate a skill *before* they begin their inquiry in the *n*th attempt, $P(Prac_n = True)$ using the prior estimate of knowledge, $P(L_{n-1})$. The equations for computing these two estimates are as follows:

$$PLn = PLn-1 Pracn + 1 - PLn-1 Pracn * T, \text{ where}$$

$$PLn-1 Pracn = False = PLn-1 * 1 - SPLn-1 * 1 - S + 1 - PLn-1 * G$$

$$PLn-1 Pracn = True = PLn-1 * SPLn-1 * S + 1 - PLn-1 * 1 - G$$

$$PPracn = True = PLn-1 * 1 - S + 1 - PLn-1 * G$$

One assumption of the Classic BKT model that is relevant to the present work is that it assumes either that complete transfer of skill occurs or no transfer occurs (cf. Sao Pedro et al., 2013c). That is, in Classic BKT, full transfer can be assumed by treating two skills as the same skill (e.g., designing controlled experiments is the same skill whether it is in Phase Change or Density); no transfer is assumed by treating the skill as a separate, independent skill within each topic (e.g. designing controlled experiments in Phase Change and designing controlled experiments in Density are different skills). Since we believe that the acquisition of inquiry skills is richer than this, rather than make either assumption, we developed an extension to BKT that aims to capture the possibility of partial transfer of skill (cf. Singley & Anderson, 1989) across science topics, the BKT-PST model. Capturing partial transfer enables us to determine empirically whether transfer occurred and the degree to which it occurred across pairs of science topics.

BKT-PST: Accounting for Partial Transfer of Skills

The proposed BKT-PST model builds upon our prior work (Sao Pedro et al., 2013c) in which we extended BKT to account for partial transfer. In this work, we added two components in the model to adjust the likelihood of knowing a data collection skill, $P(L_n)$, in a new science topic. The first was an observable $Topic\_Switch_n$ = {True, False} to indicate when the student begun a new set of inquiry activities for a different science topic. The second was a degradation parameter, $k \in (0.0, 1.0)$ that lowers the likelihood of knowing the skill by a constant factor k when switching science topics. The k parameter captures that students may not readily know to apply (transfer) the same data collection skills within different simulations (cf. Singley & Anderson, 1989). We believe, though, that the original approach may not accurately model transfer. Though $k = 1$ in this model accurately models full transfer (the estimate $P(L_n)$ does not get degraded when the topic switches), $k = 0$ would predict with certainty that the student would have no skill at all, degrading $P(L_n)$ to be 0. Thus, for low values of k the model may be too strict.

The BKT-PST model has the same Bayesian Network topology as our prior work (Sao Pedro et al., 2013c), but instead we change how the k parameter impacts the estimate of $P(L_n)$. In BKT-PST, the k parameter represents the percentage of learning accumulated within the first science topic that is transferred to the second topic. So, when $Topic\_Switch_n$ = True, the likelihood that students know the skill before the second science topic $P(L_n)$ is equal to the sum of the initial latent skill, $P(L_0)$, and the learning that is transferred, k*($P(L_{n-1} \mid Prac_n)$ - $P(L_0)$). The modified equations to compute $P(L_n)$ for BKT-PST become:

$PLnTopic\_Switch_n=True=PST+1-PST*T$, with $PST= PL0+k*PLn-1|Pracn-PL0$

$$PLnTopic\_Switch_n=False=PLn-1Pracn+1-PLn-1Pracn*T$$

If full transfer is assumed ($k = 1$), BKT-PST behaves the same way as the classic BKT model and indicates that a student's latent skill does not degrade for a new topic. When $k = 0$, $P(L_n)$ returns back to the original estimate of initial knowledge, adjusted for the possibility of learning from the practice attempt, a more realistic assumption. In other words, mathematically when $k = 0$, $P(L_n \mid Topic\_Switch_n = True) = P(L_0) + (1 - P(L_0))*T$.

Though BKT-PST may better represent transfer, it is worth noting that it has an important limitation for a somewhat uncommon special case. Take a student who fails to demonstrate the skill completely on all attempts (*n-1* attempts) in the first science topic. After observing all these failures, the likelihood of knowing the skill, $P(L_{n-1})$, will be less than $P(L_0)$. In this case, for sufficiently low values of *k,* the PST computation will be *larger* than $P(L_{n-1} \mid Prac_n)$, the updated estimate of $P(L_{n-1})$ after observing the performance $Prac_n$. In other words, the BKT modification could yield an increase when switching topics, the opposite of our assumption of degradation after switching topics. For unusual cases like this, BKT-PST may not be an ideal model. In our study, this special case occurred on 18.06% and 26.42% of topic switches for the skill of designing controlled and testing stated hypotheses, respectively.

Model Fitting

We employed a brute force grid search approach (Baker et al., 2010), a standard approach for fitting BKT models, to determine the value of each set of parameters for our three models. In order to find the best-fitting parameters, all potential parameter combinations in the search space were tried at a grain-size of 0.01 for each skill per simulation pair. The best set of parameters is the one that yields the lowest sum of squares residual (SSR) between the likelihood of demonstrating skill, $P(Prac_n = True)$, and the actual data. The values of Guess parameter G and Slip parameter S were bounded to be below 0.5 to avoid "model degeneracy" (Baker et al., 2008), where a model may estimate that the student has a lower probability of knowing $P(L_n)$ after observing the student demonstrate the skill. All other variables were allowed to have values from 0.01 to 0.99. For the previously found best parameter set, the same brute force search process was repeated around these parameters at a grain-size of 0.001 to find a tighter fit. For the "no transfer" BKT-PST model, we applied the classic BKT model on the data from activities in both science topics to calculate overall $P(L_0)$, before using brute force grid search strategy again to calculate the other three parameters. This was done in order to avoid the model from accounting for transfer of skills across science topics by increasing the initial learning probability $P(L_0)$.

Results

As previously mentioned, we applied three models that assume, respectively, full transfer, no transfer, or partial transfer to evaluate students' mastery and transfer of science inquiry skill between pairs of science topics (e.g. between Phase Change and Free Fall Energy, Density and Free Fall Speed, etc.). This is done in two ways. First, we fit and compare the parameters of three models: Classic BKT (Corbett & Anderson, 1995) that assumes full transfer, our new BKT-PST model that empirically estimates partial transfer (BKT-PST with $k >= 0$), and a model that assumes data collection skills do not transfer across science topics (BKT-PST with $k = 0$). Then, we conduct a more stringent test, comparing whether Classic BKT and BKT-PST predicts student performance better than the no transfer model. If the Classic BKT or BKT-PST models fit student performance data better than the no transfer model, it implies that transfer occurred.

To get a sense of student performance across the activities, we first conducted a descriptive analysis by computing error rates (% students who fail to demonstrate the skills during a practice opportunity) at three key points at each topic pair where transfer (or the lack thereof) can be seen: 1) at the last practice opportunity of the first science topic, 2) at the first practice opportunity for the second science topic, and 3) at the last practice opportunity of the second topic. In order for the calculated error rates to be indicative of transfer, we expect error rates to be constant or decreasing at each point. As shown in Table 1, almost all error rates fit this criteria. There were only two exceptions where error rates slightly increased – from 18.2% to 24.2% when transferring the skill of testing hypotheses from "Density" to "Free Fall Speed" activities, and from 12.9% to 16.1% when transferring the skill of designing controlled experiments from "Free Fall Speed" to "Free Fall Energy" activities. In addition, the error rates were always smaller for the last practice opportunity of the second topic than the last practice opportunity of the first topic. Though we report on these three specific points, we note that *all* of the students' activities were used to construct and test the models.

Interpreting Parameters of the Predictive Models

Partial transfer of science inquiry skills across topics was captured by the linear transfer factor k for "transfer" models built for five science topic pairs (see Table 2). The high value of k for almost all skills and topic pairs (between 0.839 and 0.990) suggests close to full skill transfer across science topics. The only exception was for the skill of testing hypotheses for students in the "Free Fall Energy" and "Free Fall Speed" topic pair for which transfer was poor. In this special case, both the learning rate ($T = 0.001$), and transfer of what they learned in the first science topic to the second topic ($k = 0.341$) were low. But overall, these finding suggests these inquiry skills transfer between science topics in Inq-ITS, replicating earlier findings (Sao Pedro et al., 2013c).

We can understand these models better by looking at the four remaining parameters of the BKT-PST models ($P(L_0)$, *S, G, T*), and comparing these parameters to those in the classic BKT models. Because the BKT-PST model behaves exactly the same as the classic BKT model when the linear transfer factor k is equal to 1, and the k values were high across the different science topic pairs, we would expect the four remaining parameters to be very similar between models. The highly similar parameters obtained when comparing classic BKT models and the "transfer" models (Table 2) meet our expectations and indicate that transfer occurred. As such, we can conclude that students were able to apply what they had learned about data collection skill from one science topic to another with very little degradation of the skill. For the Free Fall Energy to Phase Change pair, we noticed that the Guess parameter (G) for both the classic BKT model and the "skill degradation model" hit its 0.5 boundary for designing controlled experiments skill, indicating that students who did not know the skill were as likely to get the question correct by guessing than they were to get it incorrect. More research will be needed to determine why this occurred.

Comparing Models' Overall Predictive Capability

To test and compare the how well the three BKT models performed in accurately tracking the development of each inquiry skill, we conducted six-fold student-level cross validation for all science topic pairs to determine which models hold better predictive performance in predicting skill demonstration. Specifically, we stratified students randomly into six folds per skill per science topic pair and trained and tested the models' performance by comparing the estimated $P(Prac_n = True)$ with the actual student performance at time n. This cross-validation process ensures that the models can be generalized to other groups of students beyond those whose data were used to train the models originally. Model goodness was determined by computing A', which is the probability that the model will be able to distinguish practice opportunities in which the skill is and is not demonstrated. A' was used because it is an appropriate metric to use when using predictions with a confidence value to predict a binary variable (Fogarty et al., 2005). An A' value of 0.5 implies chance-level performance, and one of 1.0 indicates perfect performance.

Overall, the "transfer" models and the classic BKT models showed similar model goodness with A' values ranging from 0.512 to 0.870 for designing controlled experiments skill, and values ranging from 0.575 to 0.900 for testing stated hypotheses skill (see Table 2). All the A's per skill per simulation pair are above the 0.5 chance level. The "transfer" model for testing stated hypotheses in the "Free Fall Energy" to "Free Fall Speed"

pair performed slightly better than the classic BKT model and an unusually low A' (0.512) was observed for the classic BKT model of designing controlled experiments in the "Free Fall speed" to "Free Fall Energy" pair. Further investigation showed that one of the training folds for "Free Fall speed" to "Free Fall Energy" pair yielded a very high learning rate (T = 0.990), which causes the model to immediately update its P(L_n) estimate to be near 1.0. The low A' is consistent with this type of degenerate model. It is also worth noting that the A' obtained for the "no transfer" BKT models are lower than those of the corresponding BKT-PST models. The fact that the BKT-PST models performed better than the "no transfer" models at predicting the students' skills, combined with the high values obtained for the linear transfer factors k supports our hypothesis that science inquiry skill transfers between two science topics.

Table 2: Parameter values for four BKT models across all science topic pairs.

Topic Pair	Model	Skill	$P(L_0)$	G	S	T	K	A'
Phase Change → Free Fall Energy (n = 140) Schools 1,2,3	Classic BKT	CtrlExp	0.621	0.138	0.053	0.142		0.870
		TestHyp	0.645	0.145	0.036	0.130		0.895
	No Transfer	CtrlExp	0.621	0.168	0.035	0.173		0.836
		TestHyp	0.645	0.177	0.018	0.150		0.860
	Transfer	CtrlExp	0.617	0.142	0.052	0.144	0.990	0.867
		TestHyp	0.615	0.149	0.017	0.148	0.904	0.891
Free Fall Energy → Phase Change (n = 31) School 4	Classic BKT	CtrlExp	0.839	0.131	0.005	0.259		0.829
		TestHyp	0.879	0.161	0.002	0.119		0.897
	No Transfer	CtrlExp	0.839	0.336	0.001	0.259		0.800
		TestHyp	0.879	0.171	0.003	0.169		0.834
	Transfer	CtrlExp	0.839	0.131	0.007	0.259	0.990	0.832
		TestHyp	0.879	0.171	0.002	0.119	0.990	0.900
Density → Free Fall Speed (n = 33) School 4	Classic BKT	CtrlExp	0.147	0.158	0.174	0.323		0.731
		TestHyp	0.489	0.001	0.06	0.356		0.831
	No Transfer	CtrlExp	0.147	0.383	0.001	0.092		0.606
		TestHyp	0.489	0.181	0.003	0.372		0.739
	Transfer	CtrlExp	0.148	0.159	0.173	0.323	0.990	0.730
		TestHyp	0.489	0.001	0.060	0.357	0.990	0.829
Free Fall Speed → Free Fall Energy (n = 31) School 4	Classic BKT	CtrlExp	0.531	0.500	0.001	0.279		0.512
		TestHyp	0.661	0.371	0.001	0.229		0.757
	No Transfer	CtrlExp	0.531	0.500	0.001	0.368		0.611
		TestHyp	0.661	0.471	0.001	0.139		0.599
	Transfer	CtrlExp	0.531	0.500	0.001	0.279	0.969	0.672
		TestHyp	0.692	0.321	0.001	0.239	0.839	0.760
Free Fall Energy → Free Fall Speed (n = 64) School 5	Classic BKT	CtrlExp	0.505	0.001	0.354	0.297		0.642
		TestHyp	0.287	0.33	0.174	0.087		0.575
	No Transfer	CtrlExp	0.505	0.176	0.253	0.079		0.632
		TestHyp	0.287	0.378	0.072	0.001		0.573
	Transfer	CtrlExp	0.506	0.001	0.353	0.296	0.990	0.633
		TestHyp	0.429	0.318	0.164	0.001	0.341	0.593

Discussion and Conclusions

In this paper, we leveraged Educational Data-Mined models to investigate whether two data collection inquiry skills, designing controlled experiments and testing stated hypotheses, transferred across four physical science simulations in Inq-ITS (Gobert et al., 2012). To empirically test for transfer, we developed two different models based on Bayesian Knowledge Tracing (Corbett & Anderson, 1995). Each makes different presuppositions about the likelihood of transfer occurring. The first, Classic BKT (Corbett & Anderson, 1995), assumes either *complete transfer,* or *complete skill independence*. The second model, BKT-PST, captures partial transfer of skill. BKT-PST assumes that inquiry skills are more nuanced in their acquisition and transfer, and that they are likely to be honed more gradually. We determined whether transfer occurred between topics by comparing the BKT-PST transfer model to a BKT-PST model with the assumption of no transfer, which posits that skills are tied to the domain in which they are learned (cf. van Joolingen et al., 2007). Our results indicated that both skills

transferred across nearly all the pairs of the physical science topics tested. This was demonstrated by the BKT-PST transfer parameter having very high values for both inquiry skills. In addition, we found that the BKT-PST model better captured student performance than the BKT-PST model with no transfer assumed in 5 of the 6 topic pairs for both inquiry skills, increasing our confidence that transfer occurred.

This paper makes two main contributions towards understanding of inquiry learning and scalable, performance-based assessment of inquiry. Our findings contribute to the understanding of domain-specificity vs. domain-generality of inquiry skills (Kuhn et al., 1992; Klahr & Nigam, 2004; van Joolingen et al., 2007) since they suggest that skills have some domain-general aspects. For example, once one knows how to design experiments, they can do so in a new domain to better understand a phenomena under investigation (e.g. Gobert et al., 2012). However, we note that all of the physical science simulations studied here have a similar, linear causal structure, which may have facilitated transfer. Skills may manifest themselves differently for simulations with more complex causal systems (Jacobson & Wilensky, 2006). In addition, we note that transfer was determined for activities solely within the learning environment. In the future, it will be beneficial to determine if the models can also predict skill knowledge on other tests external to the system in order to better understand how general these inquiry skills are (Klahr & Nigam, 2004; Baker et al., 2011).

This work also contributes to the literature on scalable, performance-based formative assessment of inquiry skills across domains. Our models explicitly capture skill transfer, and can be used to estimate students' performance and drive scaffolding in real-time (Sao Pedro et al., 2013c). We note that it is important to use metrics for inquiry skills that do not require that students conduct sequential experimental trials in lock-step fashion (e.g. McElhaney & Linn, 2010). Skill at designing controlled experiments can manifest itself multiple ways (Sao Pedro et al., 2013a), and a distinction needs to be made between students designing controlled experiments in unusual ways and students engaging in haphazard inquiry (cf., Buckley et al., 2010). One limitation of our BKT-PST model is that it cannot cleanly identify *what* causes transfer. There were substantial gaps in time between assessments during which teachers may have provided supports that helped students to acquire and transfer skills.

In closing, we note that developing science inquiry skills is a *necessary but not sufficient condition* for deep science learning. We believe that it is the application of these skills to science phenomena in rich meaningful ways that has the potential to result in deep conceptual learning. As such, being able to identify and track how these skills develop and transfer is crucial towards promoting rich skill development.

References

Baker, R., Clarke-Midura, J. (2013) Predicting Successful Inquiry Learning in a Virtual Performance Assessment for Science. In *Proc. of the 21st Int'l Conf. on User Modeling, Adaptation, and Personalization* (pp. 203-214).

Baker, R., Corbett, A., Aleven, V. (2008). More accurate student model through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 406-415).

Baker, R., Corbett, A., Gowda, S., Wagner, A., MacLaren, B., Kauffman, L., et al. (2010). 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* (pp. 52-63).

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

Baker, R., Gowda, S., & Corbett, A. (2011) Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. In *Proc. of the 4th Int'l Conf. on Educational Data Mining* (pp. 179-188).

Buckley, B.C., Gobert, J., Horwitz, P. & O'Dwyer, L. (2010). Looking inside the black box: Assessments and decision-making in BioLogica. *International Journal of Learning Technologies*, 5(2), 166 - 190.

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

Corbett, A. T. & Anderson, J. R. (1995). Knowledge-Tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4,* 253-278.

de Jong, T. (2006). Computer Simulations - Technological advances in inquiry learning. *Science*, 312(5773), 532-533.

de Jong, T., & van Joolingen, W. (1998). Scientific Discovery Learning with Computer Simulations of Conceptual Domains. Review of Educational Research, 68, 179-201.

Fogarty, J., Baker, R., & Hudson, S. (2005). Case Studies in the Use of ROC Curve Analysis for Sensor-Based Estimates in Human Computer Interaction. *Proc. of Graphics Interface* (pp. 129-136).

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1991). Scientific Reasoning Across Different Domains. In *Computer-based Learning Environments and Problem-Solving* (pp. 345-371).

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

Gobert, J., Sao Pedro, M., Raziuddin, J., & Baker, R. (2013). From Log Files to Assessment Metrics for Science Inquiry using Educational Data Mining. *Journal of the Learning Sciences, 22*(4), 521-563.

Gowda, S., Baker, R., Pardos, Z., Heffernan, N. (2011). The sum is greater than the parts: Ensembling student knowledge models in ASSISTments. In *Proc. of the KDD Workshop on KDD in Educational Data*.

Harrison, A., & Schunn, C. (2004). The Transfer of Logically General Scientific Reasoning Skills. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, (pp. 541-546).

Jacobson, M. J., & Wilensky, U. (2006). Complex Systems in Education: Scientific and Educational Importance and Implications for the Learning Sciences. *Journal of the Learning Sciences*, 15 (1), 11-34.

Ketelhut, D., Nelson, B., Sil, A., & Yates, A. (2013). Discovering what students know through data mining their problem-solving actions within the immersive virtual environment, SAVE Science. Presented at the American Educational Research Association, April 27-May 1, 2013, San Francisco, CA.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: effects of direct instruction and discovery learning. *Psychological Science*, 15(10), 661-667.

Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.

Kuhn, D., & Pease, M. (2008). What Needs to Develop in the Development of Inquiry Skills? *Cognition and Instruction*, 26(4), 512-559.

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

McElhaney, K., & Linn, M. (2010). Helping Students Make Controlled Experiments More Informative. In *Proc. of the 9th International Conference of the Learning Sciences - Volume 1, Full Papers* (pp. 786-793).

Mislevy, R.J., Behrens, J.T., DiCerbo, K., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *J. of Educational Data Mining, 4*, 11-48.

National Research Council. (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC: The National Academies Press.

Pardos, Z. & Heffernan, N. (2010). Modeling individualization in a Bayesian Networks implementation of knowledge tracing. In *Proc. of the 18th Int'l Conf. on User Modeling, Adaptation and Personalization* (pp. 255-266).

Pellegrino, J., Chudowsky, N., and Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Quellmalz, E., Timms, M., & Schneider, S., (2009). *Assessment of Student Learning in Science Simulations and Games*. Washington, DC: National Research Council Report, Washington, DC.

Sao Pedro, M., Baker, R., & Gobert, J. (2012). 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* (pp. 249-260).

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

Sao Pedro, M., Baker, R., & Gobert, J. (2013b). What Different Kinds of Stratification Can Reveal about the Generalizability of Data-Mined Skill Assessment Models. *Proceedings of the 3rd Conference on Learning Analytics and Knowledge*, (pp. 190-194).

Sao Pedro, M., Baker, R., & Gobert, J. (2013c). Incorporating Scaffolding and Tutor Context into Bayesian Knowledge Tracing to Predict Inquiry Skill Acquisition. In *Proceedings of the 6th International Conference on Educational Data Mining*, (pp. 185-192).

Singley, M. & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard Univ. Press.

Thorndike, E.L. & Woodworth, R.S. (1901). The influence of improvement in one mental function upon the efficacy of other functions. *Psychological Reviews,* 8, 247-261.

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

Williamson, D., Mislevy, R., & Bejar, I. (2006). *Automated 0053coring of Complex Tasks in Computer-Based Testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Acknowledgements

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.