Assessing the Learning and Transfer of Data Collection Inquiry Skills Using Educational Data Mining on Students' Log Files

Michael A. Sao Pedro, Janice D. Gobert, and Ryan S.J.d. Baker
{mikesp, jgobert, rsbaker}@wpi.edu
Department of Learning Sciences and Technologies
Worcester Polytechnic Institute
100 Institute Rd. Worcester, MA 01609

Abstract: In this paper we explored whether engaging in two inquiry skills associated with data collection, designing controlled experiments and testing stated hypotheses, within microworlds for one physical science domain (density) impacted the acquisition of inquiry skills in another domain (phase change). To do so, we leveraged educational data mining techniques to both assess and estimate students' inquiry skills across domains. Analyses revealed that honing these skills in density activities provided benefits in terms of transfer and skill acquisition. More specifically, students who practiced in density activities first were more likely to show mastery of the designing controlled experiments skill than those who had no prior practice. These same students were also more likely to test their stated hypotheses during their first data collection in phase change. Thus, practice in one domain can positively impact acquisition and transfer of skill in a second domain, suggesting that inquiry skills also have a degree of domain generality.

Introduction

Science educators agree that cultivating inquiry skills is critical for students to become scientifically literate (National Research Council, 1996, 2000, 2011; Kuhn, 2005a). However, typical standardized science tests do not adequately reflect or assess complex inquiry process skills (Quellmalz, Timms & Schneider, 2009). Performance assessments of inquiry, instead, have been argued to be better-suited for this purpose (cf. Black, 1999; Pellegrino, 2001). Devising scalable and reliable performance assessments, though, is difficult for two reasons. First, it is difficult to separate inquiry skills from content understanding (Mislevy, Steinberg & Almond, 2002; Mislevy, et al., 2003). Second, inquiry processes are mutli-faceted, and there is no one single "right or wrong" way to engage in science inquiry (Shute, Glaser & Raghavan, 1989; Glaser, Schauble, Raghavan, & Zeitz, 1992). Given inquiry's importance, proper techniques for measuring inquiry are needed.

It is also important to better understand inquiry learning so that we can foster transfer of such skills to novel tasks (Kirschner, Sweller & Clark, 2006; Hmelo-Silver, Duncan & Chinn, 2007). Regarding transfer, it has been suggested that inquiry skills are tightly tied to the domain in which they are learned (van Joolingen, de Jong & Dimitrakopoulout, 2007), but some evidence exists that long-term, repeated practice of inquiry (Kuhn, Schauble & Garcia-Mila, 1992; Dean Jr. & Kuhn, 2006; Kuhn & Pease, 2008), and scaffolding or teaching these skills explicitly (Klahr & Nigam, 2004) can lead to successful acquisition and transfer to novel tasks.

In the present paper, we address two goals. First, we describe our approach for developing reliable, scalable performance measures of inquiry. Second we leverage those assessment techniques to examine how inquiry skills transfer between two physical science domains. We focus on on two inquiry skills, designing controlled experiments and testing stated hypotheses. Designing controlled experiments entails selecting experiments to yield

data that supports determining the effects of manipulable variables on outcomes. Testing stated hypotheses refers to generating data with the intention to support or refute a specific hypothesis. These skills are measured as students conduct inquiry within microworlds for two domains, phase change and density, developed within the Science Assistments system (Gobert et al., 2007; Gobert et al., 2009).

In our approach, we leverage techniques from Educational Data Mining (cf. Baker & Yacef, 2009; Romero & Ventura, 2010) to assess and track inquiry skills across several activities within each domain. To assess these skills, we use validated detectors (models) of students' inquiry behaviors that were constructed based on student log files (Sao Pedro et al., 2010, in press). We then produce estimates of student proficiency for each skill by aggregating all assessments into a Bayseian Knowledge Tracing model (Corbett & Anderson, 1995). This approach is rigorous because an EDM affords the ability to estimate how well the models assess and track skill. Furthermore, these can be done in real time. Thus, we argue, the approach is scalable, and could provide a possible model for inquiry assessment.

These techniques were also leveraged to measure whether the two skills of interest, designing controlled experiments and testing stated hypotheses, transfer across two physical science domains. These two domains are Density and Phase Change. More specifically, we analyzed whether students who practiced in density activities first had a greater likelihood of demonstrating the skills or reaching mastery than students with no prior practice. With our assessment and skill tracking models, we can assess this transfer in a finer-grained way than other prior studies of inquiry (e.g. Kuhn & Pease, 2008).

The remainder of this paper is organized as follows. First, we describe the two skills of interest in more detail, and present related work on assessing data collection skills. We then present a high-level view of our approach for assessing and estimating proficiency at these data collection skills using our educational data mining techniques. Next, we present our results on the transfer of these skills between domains which leveraged our assessment and estimation techniques. Finally, we present a discussion and conclusions of our paper.

Background

Data Collection Skills of Interest

Skills related to designing and conducting experiments (cf. National Research Council, 1996) are important to inquiry learning for two reasons. First, they have been argued to support the development of other scientific inquiry skills such as correctly interpreting data, and warranting claims (Klahr & Dunbar, 1988; Kuhn, Schauble & Garcia-Mila, 1992; Schauble, Glaser, Duschl, Schulze & John, 1995; Kuhn, 2005a; de Jong et al., 2005). For example, if controlled experiments are not designed, then valid conclusions cannot be drawn about the effects of variables on outcomes. Second, students typically have difficulty with these skills (de Jong & van Joolingen, 1998; Kuhn, 2005a,b) and as a result, engage in unfruitful exploration. For example, students may not collect data that support or falsify their articulated hypotheses (van Joolingen & de Jong, 1991, 1993; Kuhn, Schauble, Garcia-Mila, 1992; Schauble, Klopfer, Raghavan, 1991). They may only run a single trial when trying to confirm a hypothesis, thereby not generating enough data to make inferences (Kuhn, Schauble, Garcia-Mila, 1992). They may also run the same trial repeatedly (Kuhn, Schauble & Garcia-Mila, 1992; Buckley, Gobert & Horwitz, 2006). Finally, they may change too many variables between experimental trials, preventing proper inferences from being made (Glaser et al., 1992; Reimann, 1991; Tsirgi, 1980; Shute & Glaser, 1990; Kuhn, 2005a; Schunn & Anderson, 1998, 1999; Harrison & Schunn, 2004; McElhaney & Linn, 2008, 2010).

In this paper, we focus on the acquisition and transfer of two such data collection skills, designing controlled experiments and testing stated hypotheses. Skill at designing controlled experiments is demonstrated when a student designs experiments that yield data to support determining the effects of manipulable (independent) variables on outcomes (dependent variables). This skill is related to understanding and successful use of the Control of Variables Strategy (CVS; cf., Chen & Klahr, 1999). CVS entails the procedural and conceptual understanding of how, when, and why a controlled experiment should be conducted so that one can make valid inferences about the effects of one independent variable on a dependent variable (Chen & Klahr, 1999; Kuhn, 2005b). We differentiate designing controlled experiments from CVS as follows. CVS is a skill which emphasizes creating a single, contrastive and controlled experiment (a single pair of trials) to determine the effects of a variable (e.g. Chen & Klahr, 1999; Klahr & Nigam, 2004). Designing controlled experiments, on the other hand, applies to the collection of an entire dataset during open-ended inquiry and could involve multiple trials and variables.

A second, related skill we track is whether students understand how to test their stated hypotheses. Testing stated hypotheses refers to generating data with the intention to support or refute a previously stated hypothesis about the relationship between an independent variable and a dependent variable. We track this in addition to designing controlled experiments for two reasons. First, this skill can be demonstrated separately as students collect data. Students may attempt to test their hypotheses with confounded designs, or may design controlled experiments for a hypothesis not explicitly stated. Second, skill at testing hypotheses may be indicative of students' successful planning and monitoring of their inquiry (de Jong, 2006).

Prior Work on the Transfer of Data Collection Skills Across Domains

It is an open question whether or not inquiry skills are tied to the domain in which they are learned (van Joolingen, de Jong, & Dimitrakopoulout, 2007). However, several researchers have provided evidence that this is not the case. For example, Glaser, Schauble, Raghavan & Zeitz (1991) inferred that college students' inquiry skills had a degree of domain generality from improvements in content gains across three different simulation domains. Harrison and Schunn (2004) found that two groups of experts, those with domain expertise and those without, showed comparatively skilled inquiry behavior. Though both studies provide evidence of the domain generality of inquiry skills in a broad sense, they did not track how development and transfer of specific skills occurred across domains.

Others have researched the development of inquiry skills in grade school and middle school students at a more fine-grained level (Kuhn et al., 1992; Kuhn & Pease, 2008). In these studies, a recurring finding was that repeated practice over time is necessary for transfer. More specifically, Kuhn et al. (1992) and Kuhn and Pease (2008) showed that with repeated, long-term practice, inquiry skills can co-develop across domains. Though comprehensive in identifying how inquiry skills develop and transfer over time, both studies had some limitations. First, smaller sample sizes of at most 30 students were used. Second, the skills of data analysis and interpretation skills were conflated with experimental design skills, thereby, not providing data about how each develop separately. Finally, scaling using this approach is difficult because all performance data consisted of hand-scored open responses and/or reports.

In our approach, we aim to develop scalable assessments of inquiry which can be used, in part, to study how skills develop over time and transfer across domains. Our approach aims to assess students log files which provide rich performance metrics of students' inquiry skills. In order to do so, we require a rigorous way of assessing such skills; we discuss others' approaches for doing so below.

Prior Work on Assessing Data Collection Skills

Several researchers have assessed data collection skills and tracked their development to address a variety of research questions. Buckley, Gobert et al. (2006, 2010) defined a broader notion of inquiry skill with regard to data collection, "systematic" exploration, and measured it looking at students' log files within microworlds. They then studied the relationship between systematic inquiry and content knowledge gains, that systematic inquiry was beneficial at post-test, even if the students' inquiry lead them to the incorrect answer. Others looked specifically at the impact of designing controlled experiments on various outcomes. For example, Shute and Glaser (1990) analyzed whether certain exploration behaviors, one being the number of variables changed between experiments, impacted content gains. Schunn and Anderson (1999) compared the number of extraneous variables that were changed between successive trials between novices and experts in a domain during inquiry. Harrison and Schunn (2004) also explored differences between novices and experts' inquiry tendency to design controlled experiments, but did so by computing several ratios of the number of controlled trials compared to the total number of trials. Similarly, McElhaney and Linn (2008, 2010) computed a "CVS score" for open-ended inquiry by computing the number of successive pairwise CVS trials. They then compared the degree to which students designed controlled trials depending on the task goal. Finally, Kuhn and Pease (2008) tracked students' developing inquiry skills, in part, by computing the degree to which students make inferences, assessed on a 5-level scale. A common thread in all these approaches is that data collection, in particular the degree to which experiments were controlled, was measured using rules defined by the researchers. In other words, these rules were *knowledge-engineered*.

Such approaches, however, may fail to properly measure skill in environments where students may exhibit a variety of data collection strategies. For example, consider employing McElhaney and Linn's (2008, 2010) approach to measure skill at designing controlled experiments by computing successive pairwise CVS trials. This approach may fail to catch "corner cases" in which students exhibit additional behaviors. For example, a student may run repeated trials to observe the microworld, change one variable, run a few more repeated trials, change one variable, etc. As another example, a student may initially run pairwise experiments and then search for interaction effects. In both cases, students appear to understand how to design controlled experiments, but were engaging in other kinds of valid exploration behaviors. The successive pairwise controlled experiments rule, though, would yield a low estimate of skill. The averaged-based approaches of Harrison and Schunn (2004) also would yield lower estimates. As illustrated, since students may collect any data they like and exhibit a variety of strategies, engineering rules and identifying all potential "corner cases" can be quite difficult.

Rather than engineer rules, we, instead, developed validated, machine-learned detectors (models) to assess these skills using an Educational Data Mining (EDM) approach (cf. Baker & Yacef, 2009; Romero & Ventura, 2010). In this approach, student log files are used as a basis for discovering the "rules" describing what it means to design controlled experiments and test stated hypotheses. This differs from knowledge engineering in that rules are not prescribed a-priori. Instead, given student data, human-classified labels, and a feature set derived from student data, we use machine learning techniques to build models of various inquiry behaviors. Generally speaking, there are several advantages to a machine learning approach over knowledge engineering. First, the resulting models can capture relationships that humans cannot easily codify rationally, while leveraging the human ability to recognize demonstration of skill. Thus, this approach may be less subject to the "expert blind spot"

about what students will do. The models can also capture corner cases, and the fuzziness at the edges of these cases, more appropriately than knowledge engineering approaches. Finally, the accuracy and generalizability to new student populations or other domains of models are easier to verify than for knowledge engineering, since machine learning is amenable to cross-validation. Cross-validation is a standard method for predicting how well models will generalize to new data (cf. Efron & Gong, 1983). Thus, this approach facilitates concrete determination of model goodness.

In the following sections, we describe our approach for analyzing the degree of data collection skill transfer between two physical science domains. We also describe the EDM-based models which enabled us to conduct this research. In particular, we describe at a high level our EDM approach to automatically assess these skills, and aggregate assessments to yield estimates of student knowledge (Sao Pedro et al., 2010, in press).

Method

Participants

Participants were 148 eighth grade students (12-14 years old) from a public school in suburban Central Massachusetts. They had no previous experience using microworlds within Science Assistments.

Materials

We studied the acquisition and transfer of the two "designing and conducting experiments" skills using inquiry activities developed for the Science Assistments System (www.scienceassistments.org). This system is a web-based inquiry learning environment for Physics, Life Science, and Earth Science that automatically assesses (and in the future scaffolds) scientific inquiry skills in real-time within interactive microworld simulations. These simulations, designed for use at the middle school level, span several science domains including physical, life, and earth science (Gobert et al., 2007, 2009). Each microworld targets domain-specific concepts defined in the Massachusetts Curricular Frameworks content standards for Middle School Science (Massachusetts Department of Education, 2006). Within each microworld, inquiry skills identified in the National Science Education Standards for middle school (National Research Council, 1996, 2011) are assessed. These skills include: hypothesizing, designing and conducting experiments, interpreting data, warranting claims, and communicating findings.

In the present work, we examined transfer of skills between two physical science topics, phase change and density. We describe in more detail below the microworlds and associated activities in which behaviors were detected and skills were measured.

Phase Change Activities

Four activities built around one microworld (Figures 1 and 2) focused on the phase change of ice. They aimed to foster understanding about the invariant properties of a substance's melting and boiling point through experimentation. Each activity provided students with an explicit goal to determine if one of four variables (container size, heat level, substance amount, and container covered) affected properties of a substance's phase change (melting point, boiling point, time to melt, and time to boil). Students addressed the goal by hypothesizing, collecting and analyzing data, and communicating findings about how a variable affected the outcomes. Each of these tasks was structured into different phases that supported students' overall experimentation: "observe", "hypothesize" (Figure 1), "experiment" (Figure 2), and "analyze data". Within each phase, inquiry support tools were provided. For example, a hypothesizing widget supported writing of a well-structured,

testable hypothesis, and a "data table" tool (Figure 2) kept track of students' experimental designs and the results of running trials. Though the overall inquiry process was organized this way, students still had a moderate degree of control within the activities. Students had some freedom to navigate between inquiry phases and had flexibility within each phase to conduct many actions. Furthermore, students could choose to ignore the explicit goals and test whatever hypotheses they wished. Finally, though inquiry was structured into phases, explicit scaffolding on students' experimentation processes was not provided. For more information about these activities, see Sao Pedro et al. (in press).

Density Activities

Three activities utilized the Density Microworld (Figure 3). This microworld enabled students to inquire about the relationships between mass, volume, and density and is based on Archimedes' principle of buoyancy. Similar to Phase Change, a typical task in this domain provided students with an explicit goal to determine if a particular independent variable (orientation of object, type of liquid, volume of object and mass of object) affects density. However, unlike Phase Change, the activities were more open-ended in that the activities had fewer inquiry support tools. For example, support tools for hypothesizing and analyzing data are not provided. Students instead write hypotheses and analyses in open response boxes. Second, the manner in which students moved between phases of inquiry is slightly different. Unlike Phase Change, students engage in hypothesizing only once. Finally, students could not elect to "observe" before forming a hypothesis. These differences existed because the inquiry support tools and navigation components had not been implemented for Density at the time we collected our data.

Procedure

First, students took a paper-style pretest to assess initial data collection skills. Next, they received an introduction on relevant vocabulary needed for the activities. Then, the Science Assistments System randomly assigned students to a domain activity order, phase change first or density first. After completing both activity sets, another paper-style inquiry test was administered. Since the pre and post-tests are not a focus of this paper, we do not discuss their contents. This procedure took two class periods, about 1.5 hours in total.

We explored whether practicing data collection skills in density activities, more open-ended tasks, improved skill acquisition in the phase change activities, slightly more structured tasks. We note that though the experimental design enabled testing if phase change practice impacted density activity performance, these analyses were not conducted here because the density log data has not yet been distilled. Next, we describe how we developed automated mechanisms for measuring acquisition of the two inquiry skills within the phase change activities. These mechanisms will be leveraged to analyze the degree of transfer between the two domains.

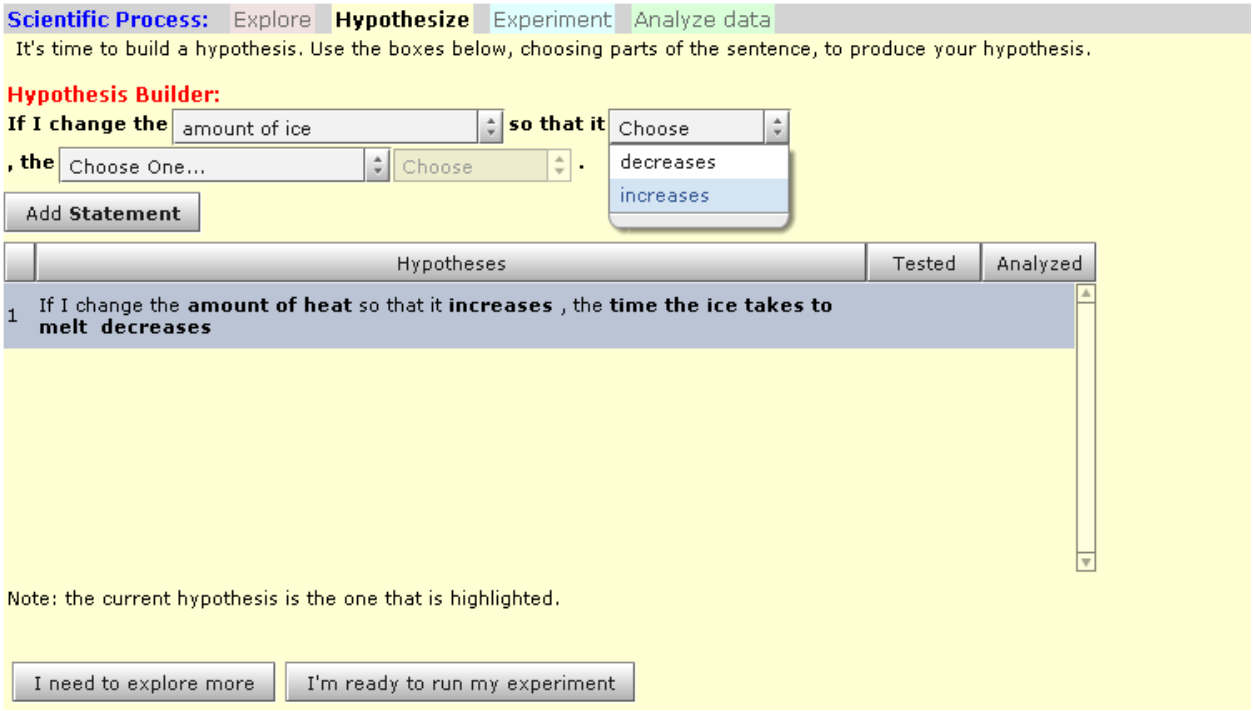

Figure 1. Hypothesizing tool for the Phase Change microworld.

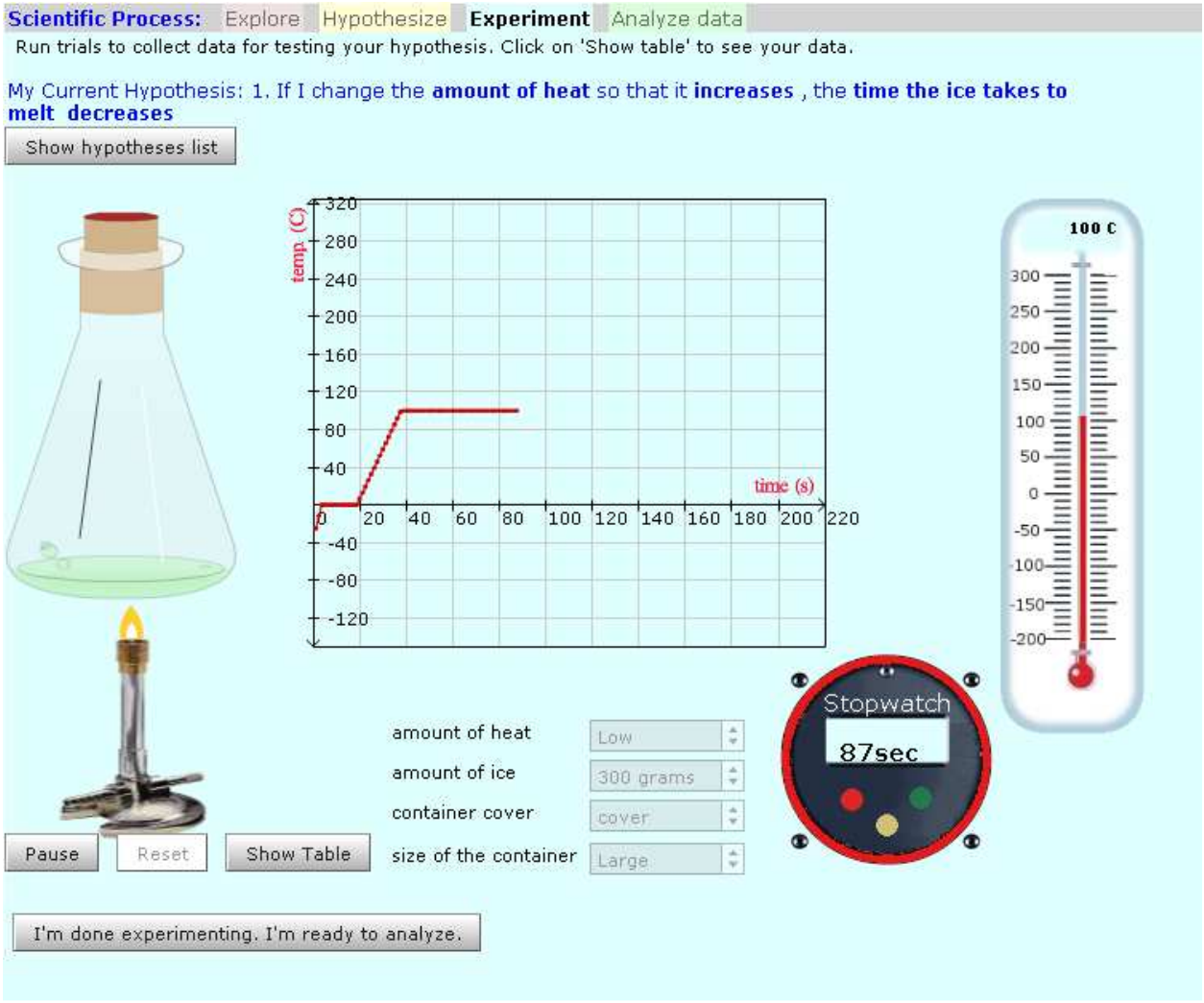

Figure 2. Experiment Phase of the Phase Change microworld.

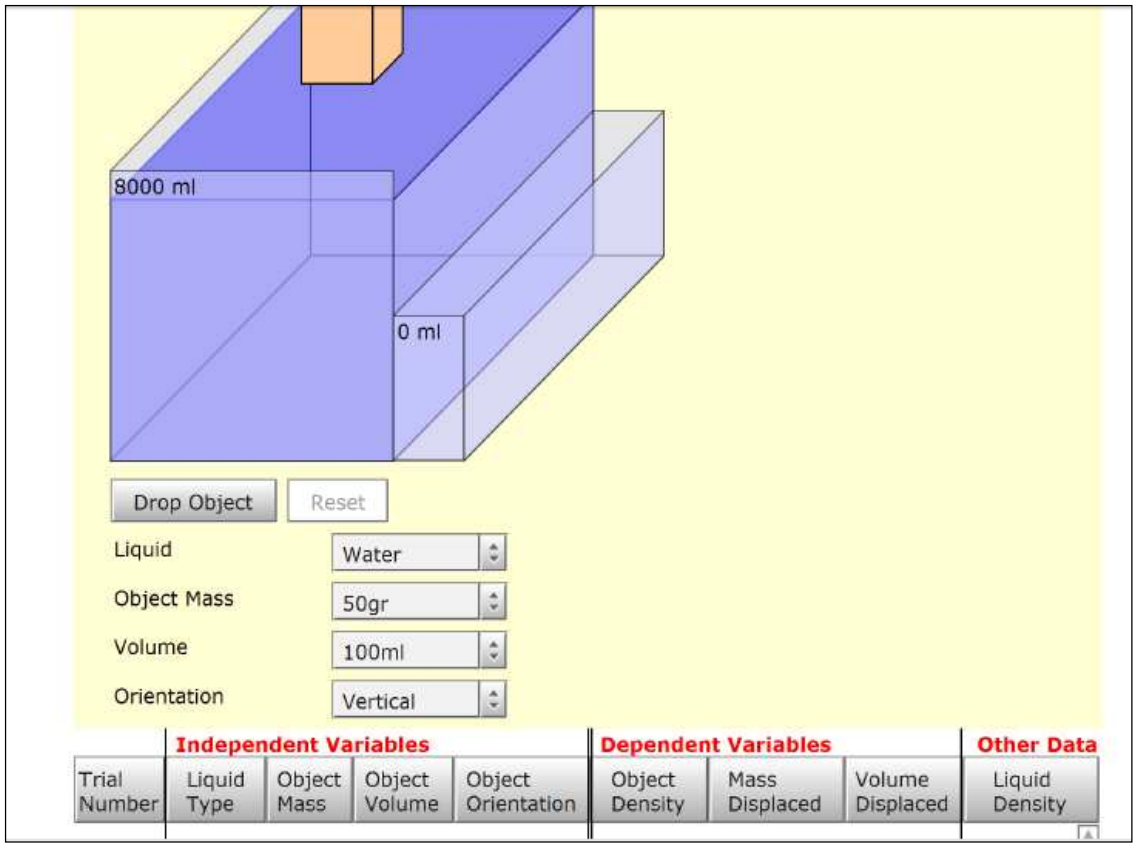

Figure 3. Density microworld on Archimedes' Principle

Leveraging Educational Data Mining Techniques to Measure and Track Data Collection Skills

Measuring and tracking the skills within the phase change activities involved two steps. First, we used an automated method for assessing whether students designed controlled experiments or tested their stated hypotheses in a microworld activity. Second, we used another automated approach to aggregate single assessments over all activities, and produce final estimates of latent skill. These methods are briefly described below; a full description of both approaches appears in Sao Pedro, et al. (in press).

Assessing Data Collection Skills with Machine-Learned Behavior Detectors

Automatic assessment was performed using machine-learned behavior detectors (models) of behaviors associated with each skill. At a high level, this approach leverages machine-learning to "discover" what it means to design controlled experiments and test stated hypotheses in our learning environment. Thus, unlike knowledge engineering in which rules to describe behaviors are authored by a human (cf., Koedinger & MacLaren, 2002), our machine-learning approach attempts to derive rules based, in part, on student data. More specifically, we employed "text replay tagging" of log files (Sao Pedro, et al., 2010; Montalvo et al., 2010; Sao Pedro et al., in press), an extension to the text replay approach developed in Baker, Corbett and Wagner (2006) to build and validate behavior detectors. Text replay tagging, a form of protocol analysis (Ericsson & Simon, 1980, 1984), leveraged human judgment to identify whether students' log files demonstrated inquiry skill.

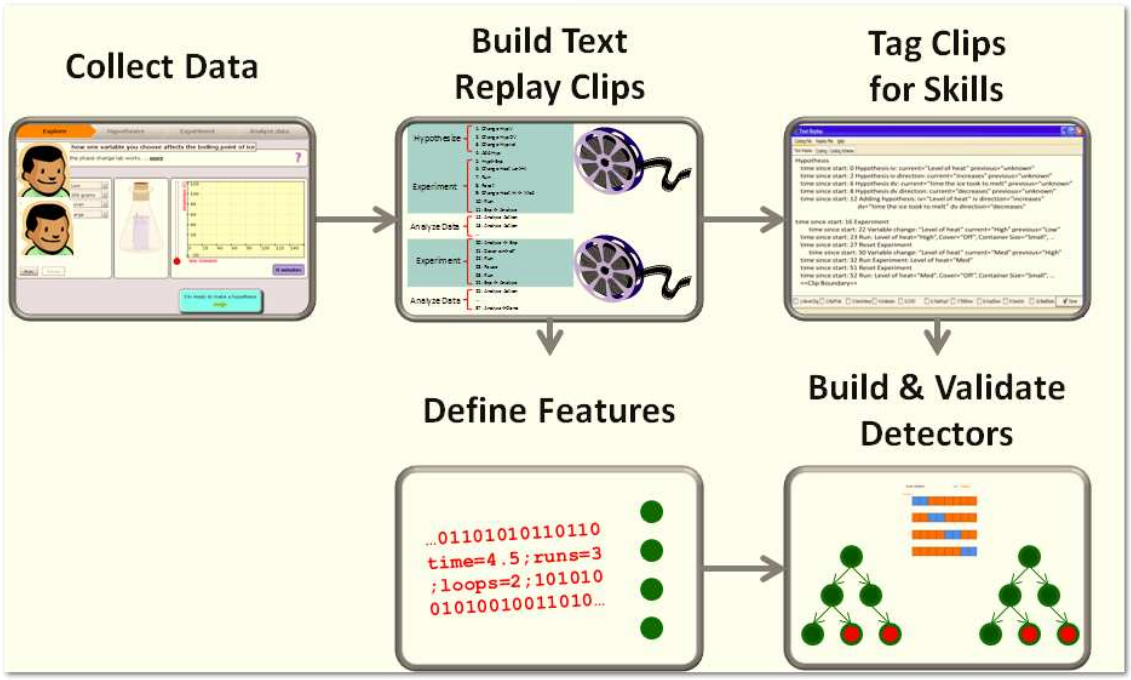

Figure 4. Overview of the text replay tagging process that enabled the construction of validated, machine-learned behavior detectors. These detectors are used to assess whether students design controlled experiments or test their stated hypotheses during their experimentation.

As shown in Figure 5, there are several steps involved our text replay tagging approach. The process begins by having students engage in inquiry within the Phase Change activities and collecting their log files. From there, the log files are segmented into meaningful sets of student actions called *clips*. Human coders then tag a subset of these clips with the behaviors of interest, designing controlled experiments and testing stated hypotheses. These tags are combined with a set of *features* (attributes) which summarize the clips. The clips, represented as a combination of behavior tags and features, provide the backbone for discovering behavior models and testing how well the model performs. The clips (features and tags) are given to a machine-learning algorithm to "discover" models relating the features to demonstration of each behavior. Finally, the models are validated by measuring how well they predict behavior in clips not used to build the models. The output of this process is two validated behavior detectors which can be leveraged to assess whether or not a student demonstrates behavior during a clip, a segment of their experimentation in an activity.

Two key processes in the text replay tagging approach are having human coders label behavior within clips and building and validating the detectors. We describe each in more detail below to provide a more concrete sense of how text replay tagging was conducted.

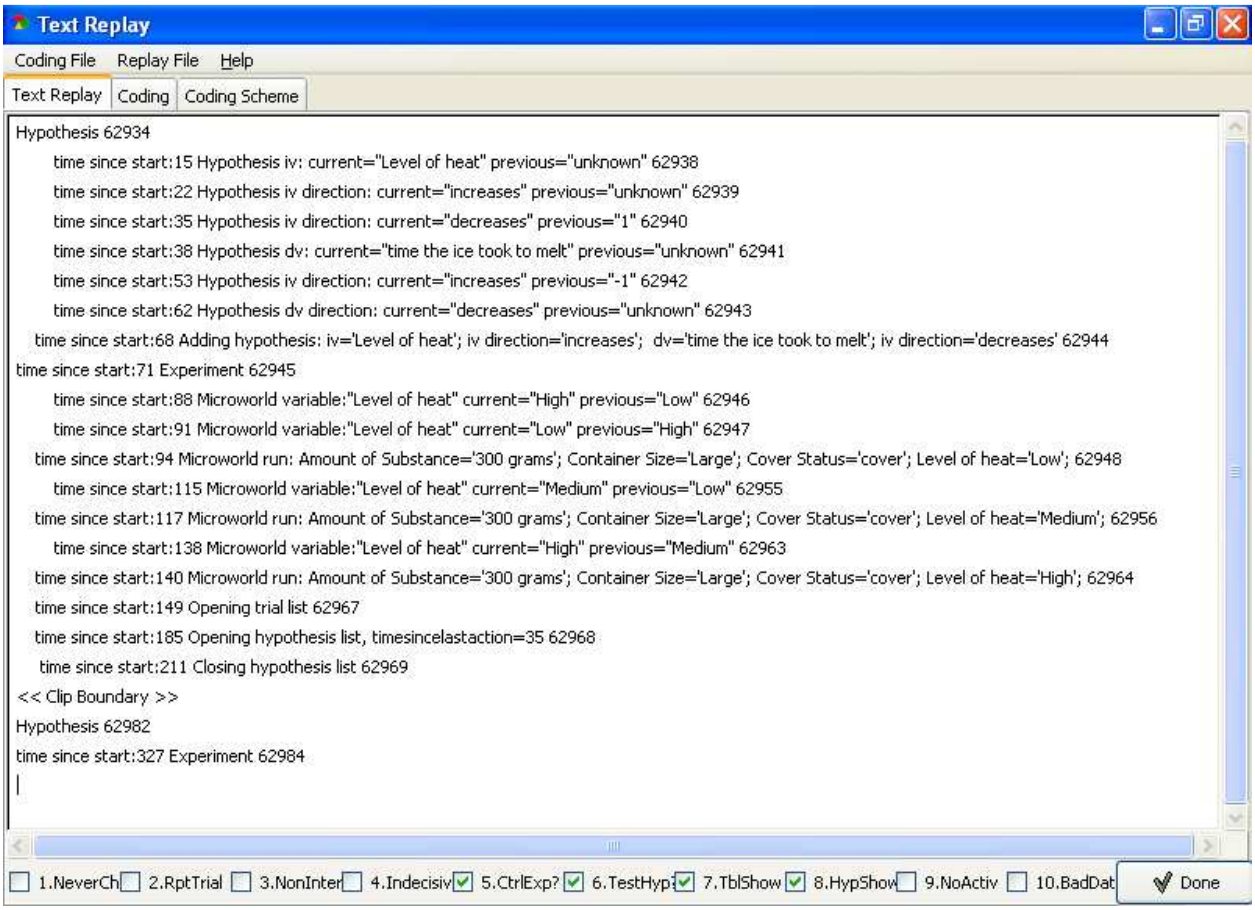

Figure 5. An example clip labeled by a human coder. This clip was tagged as involving designing controlled experiments and testing stated hypotheses, in addition to other behaviors.

Tagging Behavior in Clips for Phase Change Activities

A key part of this process is to have human coders apply behavior tags to clips. These behavior tags act as the "gold standard" from which detectors are built and validated. To help human coders identify behavior, a clip is cleaned and summarized into a text replay. A text replay, shown in Figure 6, summarizes clip actions and highlights important aspects of students' inquiry processes. Choosing which actions should be included in the replay is of particular importance since a human coder needs sufficient information to identify whether a student is demonstrating the behaviors. In the phase change microworld, such behavior is exhibited in the hypothesizing and (primarily) the experimenting phases of inquiry. Thus, to properly identify inquiry behavior, a clip contains all student actions relevant to hypothesizing and experimenting. This enabled coders to obtain a more comprehensive view of students' inquiry processes, necessary for labeling processes (such as these) that unfold over time.

Given a text replay representing a clip, that clip could be tagged with either the "designing controlled experiments" or "testing stated hypotheses" behaviors. It could also be tagged with both behaviors, or neither behavior. When a tag is not applied, it means the student did not demonstrate its corresponding behavior in the clip. To give a better sense of how behavior is labeled by a human coder, consider the text replay shown in Figure 6, which was tagged, in part, as demonstrating the "designing controlled experiments" and "testing stated hypotheses" behaviors. To tag these behaviors, the human coder focuses primarily on

the trials run by the student. In the experimentation cycle, he specified one hypothesis relating the level of heat applied to the ice to the time it takes to melt. He then ran a total of three trials as indicated by the "microworld run" statements at time 94s, 117s, and 140s. For each trial, he changed only the "level of heat" variable in a successive manner, comparing a 'low' level to 'medium', and then 'medium' to 'high'. He spent 78 seconds doing so. Due to the consistency in manipulating only one variable at a time between trials, this clip was tagged as demonstrating the "designing controlled experiments" behavior. In his experiments, he focused specifically on the independent variable stated in his hypothesis, the level of heat. Because of this, the clip was also tagged as "testing stated hypotheses."

In general, students' experimentation varied greatly within the phase change activities. Some students took very few actions and engaged in few experimentation cycles within an activity. Others had more complex experimentation patterns. To ensure there was consistency in identifying behaviors, we establish inter-rater reliability by having two human coders tag 50 clips. Prior to this testing, the two coders discussed the coding scheme and coded several clips together. Inter-rater reliability for each behavior was high. The Cohen's κ for designing controlled experiments was 0.69, and 1.00 for testing stated hypotheses (Sao Pedro et al., 2010, in press). This level of agreement is on par with prior text replay-based behavior detectors (Baker & de Carvalho, 2008; Baker, Mitrovic & Mathews, 2010; Lee et al., 2011).

Building and Validating Behavior Detectors

Following the text replay tagging methodology, behavior detectors for "designing controlled experiments" and "testing stated hypotheses" were constructed and validated within the phase change microworld. We present here only high-level details and a summary of the results from Sao Pedro et al. (in press). Detectors were constructed using all 148 students' interactions within four phase change activities. Then, after segmenting student actions into clips, clips were tagged by two human coders who had achieved good inter-rater reliability. One clip per student, per activity was randomly selected to be tagged. This ensured there was a representative sample of all students and all activities. In all, the human coders tagged 570 clips. In addition, a set of 73 features (Sao Pedro et al., in press) was distilled to summarize clips. Example features computed per clip include: number of trials run, number of hypotheses stated, count of pairwise controlled trials, time spent running experiments, and number of simulation pauses. The corpus of clips, represented as a combination of summary features and behavior tags, was used to train and validate detectors of each behavior.

We built and validated the detectors by following a six-fold student-level cross-validation approach. In this approach, students are randomly selected to be in one of 6 groups. Five of the six group's data are used to train (build) a detector. The remaining group is used to test how well the detector can predict behavior. This process is repeated, using each group as a test group once. This approach enabled us to estimate how well the detectors will work for new groups of students in the phase change environment. Within each training and testing loop, a detector was built as follows. First, all correlated features above 0.6 were removed. Then, J48 decision trees with automated pruning to control for over-fitting were used to derive models (Quinlan, 1993). These decision trees relate feature values to behavior predictions. Note that separate decision trees were constructed for each behavior.

As part of the cross-validation process, we can estimate how well the behavior detectors work by observing how well the detectors' predictions match the human coder's labels for all clips. We quantified the degree of agreement between the two by computing two metrics, A' (Hanley & McNeil, 1982) and Cohen's Kappa (κ). A' is the probability that if the detector is comparing two clips, one involving the category of interest (designing controlled experiments, for instance) and one not involving that category, it will correctly identify which

clip is which. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. Cohen's Kappa assesses whether the detector is better than chance at identifying the correct action sequences as involving the category of interest. A κ of 0 indicates that the detector performs at chance, and a κ of 1 indicates that the detector performs perfectly.

In Sao Pedro et al. (2010, in press), we reported that the detectors worked very well overall. The detector for designing controlled experiments could distinguish a clip which demonstrated the behavior from a clip which did not 85% of the time (A' = .85). The detector's $\kappa = .47$ indicated that its predictions agreed with coders' tags better than chance. The testing stated hypotheses detector also worked well. It could distinguish clips 85% of the time (A' = .85) and also agreed with coders' tags better than chance, $\kappa = .40$. This level of performance is comparable to other behavior detectors' which have been refined over several years (e.g., Baker & de Carvalho, 2008; Baker, Mitrovic & Mathews, 2010). Thus, the detectors can be used to automatically assess students' data collection. In the next section, we describe how we leveraged the detectors to classify all student clips, and aggregated them to estimate each student's proficiency at each skill.

Estimating Proficiency at Data Collection Skills Across Practice Attempts

To amalgamate students' performances across activities and produce proficiency skill estimates, we used Bayesian Knowledge-Tracing (BKT, Corbett & Anderson, 1995). This is a classic approach that has been successfully used to model learning within Intelligent Tutoring Systems for mathematics problem solving (e.g. Koedinger & Corbett, 2006; Feng, Heffernan & Koedinger, 2009), genetics problem solving (Corbett, Kaufmann, MacLaren, Wagner, & Jones, 2010), computer programming (Corbett & Anderson, 1995; Kasurinen & Nikula, 2009), and reading (Beck & Chang, 2007). A BKT model (shown in Figure 7) is a two-state Hidden Markov Model that estimates the probability a student possesses latent skill (L_n) after *n* observable practice opportunities. This model assumes that knowledge of a skill is binary; either the student knows the skill or does not. Given student performance data, it estimates the likelihood that a student knows the skill. To concretize this for our domain, the observable student performance is whether or not a student demonstrates one of the data collection behaviors. This is determined using the behavior detectors. Latent skill (L_n) is the estimate of whether or not a student knows how to design controlled experiments or test stated hypotheses after her *n*th time collecting data.

BKT models are characterized by four parameters, *G*, *S*, L_0, and *T*, used in part to compute latent skill (L_n). 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. Within the BKT framework, these four parameters are assumed to be the same for all students.

In this approach, a BKT model for each skill is fit from student data in order to make predictions about current students, and future students. Thus, given student data, values for the four parameters are found that minimize the error in predicting whether or not they will demonstrate behavior during data collection. In Sao Pedro, et al. (in press), we used the behavior detectors to label all students' inquiry within each activity. Then, we used a brute force search to find the best fitting parameters over the data. This method previously has been found to produce comparable or better model parameters than other methods (Baker, Pardos, Gowda, Nooraei, & Heffernan, 2011). Employing this process led to BKT models of each skill estimated students' skills at each practice opportunity reasonably well (Sao Pedro et al., in press). This was determined by measuring how well the model could predict whether a

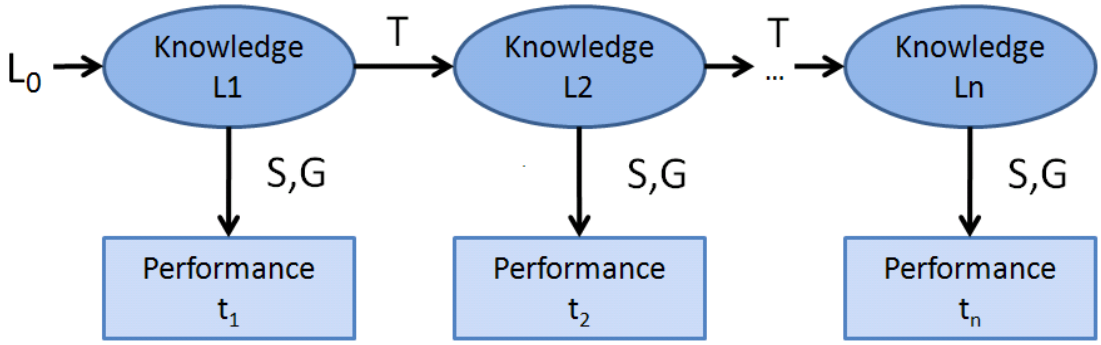

Figure 6. Classic Bayesian Knowledge Tracing model (Corbett & Anderson, 1995) for a skill, e.g., knowing how to design controlled experiments. The model estimates the likelihood the student knows a skill (L_n) after n observable practice opportunities. It does so using four parameters: L_0 is the initial knowledge, S is the likelihood of slipping, G is the likelihood of guessing and T is the learning rate of the skill.

student would demonstrate the skill at time n, based in part on the estimate of knowing the skill up to that point (L_{n-1}). BKT models for each skill could predict better than chance, A' = .74 for designing controlled experiments and A' = .79 for testing stated hypotheses.

At this point we have described a process for assessing two data collection skills during a single data collection activity, and aggregating those assessments over several activities to produce estimates of latent skill. We leverage the BKT models, in particular, to explore our original research question: does practicing in density activities first before phase change activities lead to better acquisition of skill? In other words, do inquiry skills related to data collection learned in density transfer to the domain of phase change? We address this question of transfer in the next section.

Results

Our main goal is to analyze whether practicing two data collection skills, designing controlled experiments and testing stated hypotheses, in one domain (density), will lead to improved performance on those skills in a different domain (phase change). We anticipate students who practiced in density first would possess more skill in phase change than those who did not. In other words, we hypothesize the skills will transfer to the second domain. Two design choices facilitate determining whether transfer occurred. First, the domain practice order was counterbalanced; students were randomly assigned to phase change activities first, or density activities first. This allows us to compare the two groups, which differ solely in whether they use density prior to phase change Second, there were no significant between-group differences on a pretest (before both environments) requiring designing controlled experiments skills ($t(147) = 1.23$, $p = .222$) and knowledge of hypotheses ($t(147) = 0.80$, $p = .428$). This finding implies that initial prior knowledge between groups is likely not a source of any differences that may be found between the groups.

How should transfer between the two domains be measured? We chose to detect the "additional skill" in phase change in two ways. First, students who had prior practice in density activities may be more likely to demonstrate skill on their first attempt at data collection. In other words, these students may have more initial skill when starting the

activities, and thus show immediate transfer. A second way this transfer can be measured is to examine whether more students in one condition showed proficiency by the end of the activities. Thus, practice in density first may impact the degree to which inquiry skill is acquired in phase change. We address each of these below as possible indicators of skill acquisition and transfer across domains.

Comparing Initial Performance in the Phase Change Activities

If practicing in density activities impacted acquisition and transfer of these skills, we would expect students in that condition to demonstrate skill during their first data collection opportunity within the phase change activities. To test this, we used the behavior detectors to assess whether each student designed controlled experiments or tested stated hypotheses during their first data collection within the phase change activities. As shown in Table 1, 23 out of 147 students (15.7%) designed controlled experiments during their first data collection. Almost twice as many students in the density-first condition (15 students) did so as compared to the phase-change first condition (8 students). Recall that at this point, this was the first time these students in the phase-change first condition engaged in our inquiry activities. This difference approached significance, $\chi^2(1)=3.66$, $p=.056$. For testing stated hypotheses, 35 out of 147 students (23.8%) did so during their first data collection. Of those students, more than twice as many had practiced in density activities first (24 students), as compared to those who did not (11 students). This difference was significant, $\chi^2(1)=8.63$, $p=.003$. In summary, data collection practice in the density environment appeared to be associated with greater acquisition and immediate transfer, particularly for testing stated hypotheses, to the phase change tasks.

Comparing Mastery Levels upon Completing the Phase Change Activities

It is also relevant to analyze whether initial practice in density improves students' acqusition of the two skills over multiple practice attempts in phase change. If the initial, additional practice in density provided such a benefit, we would expect students in the density-first condition to have higher final estimates of latent skill (L_{final}). This estimate of skill at the end of the phase change activities is given by the BKT model.

Descriptive analyses revealed the L_{final} sample distributions for each skill were bimodal, with the two modes centered near 0 and 1, meaning that students either "knew" the skills or did not when they had completed the phase change activities. We therefore compared which condition contained a greater proportion of students who had mastered or approached mastery of each skill by the end of the phase change activities, indicated by their L_{final} being above 0.8. As shown in Table 2, 30 out of 147 students (20.4%) in total demonstrated mastery at designing controlled experiments. However, twice as many students in the density-first condition (20 students) mastered this skill than the phase-change first condition (10 students). This difference was significant, $\chi^2(1)=5.89$, $p=.015$. Thus, practicing inquiry in the density microworld improved acquisition of the designing controlled experiments skill in phase change. For testing stated hypotheses, 50 out of 147 students (34%) demonstrated mastery. Again, more students who practiced density first (28 students) achieved mastery than students who did not (22 students), but this difference was not quite significant, $\chi^2(1)=2.50$, $p=.114$. In summary, initial practice in density appears to improve acquisition of the desiging controlled experiments skill in the phase change environment, but not for the testing stated hypotheses skill. We discuss possible reasons for this in the following section.

Table 1. Crosstabulations of domain order practice condition, and display of behavior in first phase change data collection activity

	Designed Controlled Experiments?		Tested Stated Hypotheses?	
	No	Yes	No	Yes
Density -> Phase Change	54	15	45	24
Phase Change -> Density	70	8	67	11
	$\chi^2(1) = 3.66$		$\chi^2(1) = 8.63$**	

**p < .01

Table 2. Crosstabulations of domain order practice condition, and mastery of each data collection inquiry skill.

	Designing Controlled Experiments		Testing Stated Hypotheses	
	Not Mastered	Mastered	Not Mastered	Mastered
Density -> Phase Change	49	20	41	28
Phase Change -> Density	68	10	56	22
	$\chi^2(1) = 5.89$*		$\chi^2(1) = 2.50$	

*p < .05

Discussion and Conclusions

In this paper, we presented an approach for developing models to assess and track students' inquiry via Educational Data Mining (EDM) techniques (cf. Baker & Yacef, 2009; Romero & Ventura, 2010). Two inquiry skills related to designing and conducting experiments (National Research Council, 1996, 2011), namely, designing controlled experiments and testing stated hypotheses were assessed and tracked as students conducted inquiry with computerized microworlds. We leveraged our models to explore whether practicing inquiry with a microworld in one physical science domain (density) impacted skill acquisition in another domain (phase change). To do so, we counterbalanced the order in which students practiced inquiry within the two domains, and analyzed students' performance in the phase change activities. Thus, some students had prior practice in density whereas others had no prior practice. We evaluated transfer performance in two ways. First, we compared the groups on whether they demonstrated the skill in their first data collection task. This metric provided a benchmark for determining whether students immediately recognize to use these skills when collecting data in the second domain. Second, we compared groups on whether they achieved mastery by the end of the phase change activities. This enabled us to estimate whether prior practice in density improved students' ability to acquire these skills over time.

We found that more students who were in the density plus phase change group were able to demonstrate the skill designing controlled experiments in their first data collection task when compared to the phase-change only group. This difference approached statistical significance. However, significantly more students in the density plus phase change group achieved mastery on this skill than did students in the phase-change only group. This provides evidence that the skill of designing controlled experiments may have a domain-general component to it. In addition, we interpret the group differences on our significant findings to have two possible meanings. First, designing controlled experiments may be a more difficult skill to learn than the testing hypotheses skill (as evidenced by fewer students mastering this skill across both conditions than the testing hypotheses skill). Thus, the learning trajectory for this skill may be longer and require more practice than the phase-change first condition received. Second, it may be that prior practice in density better prepared students for the skill when they reached the phase change activities. In other words,

the prior experience may have prepared them for future learning (Bransford & Schwartz, 1999). To test the latter, we will analyze students' performance in the density activities and separate out those who mastered the skill in density from those who did not. This hypothesis would be supported if more density-first students who were not "masters" at the end of density activities became "masters" by the end of the phase change activities than students with no prior practice.

For testing stated hypotheses, significantly more density-first students demonstrated this skill in their first data collection than the phase-change first condition. However, there was no significant difference on levels of mastery. We interpret these findings to mean that this skill, too, has a domain general component, given the group difference on this skill in the first activity. We also believe that since this skill may be easier to acquire than the designing controlled experiments skills, as evidenced by the finding that it took fewer practice opportunities to acquire it. Thus, overall our findings support earlier studies that data collection skills have a domain-general component to them, and that once learned/mastered, they can be transferred (e.g. Klahr & Nigam, 2004; Harrison & Schunn, 2004; Dean Jr. & Kuhn, 2006).

It is also worth noting that even with additional practice in density, only 29% of the students showed mastery of designing controlled experiments and only 41% showed mastery at testing stated hypotheses after completing the phase change activities. This may be for two reasons. First, students were not given explicit feedback on their experimentation procedures. Such feedback may help students acquire these skills (Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Sao Pedro et al., 2009, 2010). In addition, students did not engage in long-term, repeated practice, which has been shown to promote acquisition and transfer of these skills (Dean Jr. & Kuhn, 2006). In future work, we will address if providing both real-time feedback on students' experimentation strategies and/or practicing across several domains will improve learning and transfer of these skills.

One possible limitation of this study is that we did not analyze whether transfer was bidirectional, meaning whether practice in phase change first impacted performance in density activities. We did not do so because our detectors had not yet been validated to work across physical science domains. As mentioned earlier, these analyses could be used to help determine if practicing inquiry in separate domains can function as preparation for future learning. This transfer is of interest to us also since the density activities were slightly more open-ended than the phase change activities since the inquiry support tools (e.g. hypothesizing widget) were not present in the density activity. If transfer from state change to density were borne out, this would provide further evidence of domain generality of inquiry skills. Additionally, since phase change included widgets that supported students' inquiry and density did not include these widgets, transfer from phase change to density would demonstrate mastery of inquiry processes. Such a finding would illustrate Vygotsky's (1978) notion of scaffolding.

As previously mentioned, central to our approach is the use of EDM for the development of our models, one for assessing these skills during a data collection activitiy, and one for aggregating assessments to yield an estimate of skill after completing the activity. This approach, which requires as a first step text replay tagging (Montalvo et al., 2010; Sao Pedro et al., 2010, in press) and educational data mining, is novel in its application to the systematic study of inquiry learning. Text replay tagging, a form of protocol analysis (Ericsson & Simon, 1980, 1984), leveraged human judgment to identify whether students' log files demonstrated inquiry skill. The data mining portion enabled us to leverage human's codes to build and validate automated "detectors" of each skill that can replicate human judgment. Our skill proficiency estimation (aggregation) model was built using a Bayesian Knowledge-Tracing framework (Corbett & Anderson, 1995). This approach, was chosen for

two reasons: 1) it had demonstated prior success in estimating skill in several domains (e.g. Koedinger & Corbett, 2006; Beck & Chang, 2007), and 2) enabled us to measure the validity of these skill estimates (Sao Pedro et al., in press).

We believe our approach has three primary benefits over previous approaches. First, analyzing log data in this way enables a rigorous and scalable way to assess students' inquiry processes (Rupp et al., 2010). Second, with regard to the two skills of interest, our approach is advantageous over knowledge engineered approaches (e.g. Schunn & Anderson, 1999; McElhaney & Linn, 2008, 2010) in that the validity of our assessments can be more easily determined (Sao Pedro et al., in press). Finally, our approach can identify skill in situations where students also employ a variety of other valid inquiry strategies (cf. Schunn & Anderson, 1998; Veermans, 2003), whereas other approaches cannot because they exclusively code sequential pairwise trials (e.g. Dean Jr. & Kuhn, 2006; Kuhn & Pease, 2008; McElhaney & Linn, 2008, 2010).

In the future, we aim to address whether we can leverage data mining for other complex inquiry skills such as interpreting data and warranting claims with data (NSES, 1996; NRC, 2011). This will involve similar methods and techniques to those described in this paper, namely text replay tagging and educational data mining, to identify such skills in students' log files. We also aim to leverage the existing models to study transfer across more disparate science domains, namely biology and earth science. Such models can not only help to more quickly assess students inquiry in a more principaled way, but also can enable conducting broader-scale studies to empirically address questions such as the domain generality of inquiry skills.

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). Any opinions expressed are those of the authors and do not necessarily reflect those of the funding agencies.

References

Alonzo, A., & Aschbacher, P. (2004, April 15). Value Added? Long assessment of students' scientific inquiry skills. *Paper presented at the annual meeting of the American Educational Research Association.* San Diego, CA: Retrieved December 20, 2010, from the AERA Online Paper Repository.

Baker, R. S., Mitrovic, A., & Mathews, M. (2010). Detecting Gaming the System in Constraint-Based Tutors. In P. De Bra, P. Kobsa, & D. Chin (Ed.), *Proceedings of the 18th Annual Conference on User Modeling, Adaptation and Personalization, UMAP 2010. LNCS 6075*, pp. 267-278. Big Island of Hawaii, HI: Springer-Verlag.

Baker, R., & de Carvalho, A. (2008). Labeling Student Behavior Faster and More Precisely with Text Replays. In R. S. Baker, T. Barnes, & J. E. Beck (Ed.), *Proceedings of the 1st International Conference on Educational Data Mining, EDM 2008*, (pp. 38-47). Montreal, Quebec, Canada.

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

Baker, R., Corbett, A., & Wagner, A. (2006). Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop held at the 8th International Conference on Intelligent Tutoring Systems, ITS 2006*, (pp. 29-36). Jhongli, Taiwan.

Baker, R., Pardos, Z., Gowda, S., Nooraei, B., & Heffernan, N. (2011). Ensembling Predictions of Student Knowledge within Intelligent Tutoring Systems. In J. Konstan, R. Conejo, J. Marzo, & N. Oliver (Ed.), *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization, UMAP 2011. LNCS 6787*, pp. 13-24. Girona, Spain: Springer.

Baker, R., Pardos, Z., Gowda, S., Nooraei, B., & Heffernan, N. (2011). Ensembling Predictions of Student Knowledge within Intelligent Tutoring Systems. In J. Konstan, R. Conejo, J. Marzo, & N. Oliver (Ed.), *Proceedings of the 19th International Conference on User Modeling, Adaptation and Personalization, UMAP 2011. LNCS 6787*, pp. 13-24. Girona, Spain: Springer.

Baxter, G., & Shavelson, R. (1994). Science performance assessments: benchmarks and surrogates. *International Journal of Education Research, 21(3),* 279-298.

Beck, J., & Chang, K. (2007). Identifiability: A Fundamental Problem of Student Modeling. In C. Conati, K. McCoy, & G. Paliouras (Ed.), *Proceedings of the Eleventh International Conference on User Modeling, UM 2007. LNAI 4511*, pp. 137-146. Corfu, Greece: Springer-Verlag.

Beck, J., & Chang, K. (2007). Identifiability: A Fundamental Problem of Student Modeling. In C. Conati, K. Mccoy, & G. Paliouras (Ed.), *Proceedings of the 11th International Conference on User Modeling, UM 2007. LNCS 4511*, pp. 137-146. Corfu, Greece: Springer-Verlag.

Black, P. (1999). *Testing: Friend or Foe? Theory and Practice of Assessment and Testing.* New York, NY: Falmer Press.

Bransford, J., & Schwartz, D. L. (1999). Rethinking Transfer: A Simple Proposal with Multiple Implications. In A. Iran-Nejad, & P. Pearson, *Review of Research in*

Education, 24 (pp. 61-101). Washington, D.C.: American Educational Research Association.

Buckley, B., Gobert, J. D., & Horwitz, P. (2006). Using Log Files to Track Students' Model-Based Inquiry. *Proceedings of the 7th International Conference on Learning Sciences*, (pp. 57-63). Bloomington, IN.

Buckley, B., Gobert, J., Horwitz, P., & O'Dwyer, L. (2010). Looking Inside the Black Box: Assessments and Decision-making in BioLogica. *International Journal of Learning Technology, 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., & Anderson, J. (1995). Knowledge-Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

Corbett, A., Kaufmann, L., MacLaren, B., Wagner, A., & Jones, E. (2010). A Cognitive Tutor for Genetics Problem Solving: Learning Gains and Student Modeling. *Journal of Educational Computing Research, 42*, 219-239.

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.

de Jong, T., Beishuizenm, J., Hulshof, C., Prins, F., van Rijn, H., van Someren, M., et al. (2005). Determinants of Discovery Learning in a Complex Simulation Learning Environment. In P. Gardenfors, & P. Johansson, *Cognition, Education and Communication Technology* (pp. 257-283). Mahwah, NJ: Lawrence Erlbaum Associates.

Dean Jr., D., & Kuhn, D. (2006). Direct Instruction vs. Discovery: The Long View. *Science Education*, 384-397.

Efron, B., & Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician, 37*(1), 36-48.

Ericcson, K., & Simon, H. (1980). Verbal Reports as Data. *Psychological Review, 87*, 215-251.

Ericcson, K., & Simon, H. (1984). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: Bradford Books/MIT Press.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the Assessment Challenge in an Intelligent Tutoring System that Tutors as it Assesses. *User Modeling and User-Adapted Interaction, 19*, 243-266.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific Reasoning Across Different Domains. In E. DeCorte, M. Linn, H. Mandl, & L. Verschaffel, *Computer-based Learning Environments and Problem-Solving* (pp. 345-371). Heidelberg, Germany: Springer-Verlag.

Gobert, J.; Heffernan, N.; Koedinger, K.; Beck, J. (2009). ASSISTments Meets Science Learning (AMSL). Proposal (R305A090170) funded by the U.S. Dept. of Education.

Gobert, J.; Heffernan, N.; Ruiz, C.; Kim, R. (2007). AMI: ASSISTments Meets Inquiry. Proposal NSF-DRL# 0733286 funded by the National Science Foundation.

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

Harrison, A., & Schunn, C. (2004). The Transfer of Logically General Scientific Reasoning Skills. In K. Forbus, D. Gentner, & T. Regier (Ed.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society, CogSci 2004* (pp. 541-546). Chicago, IL: Erlbaum.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Krischner, Sweller, and Clark (2006). *Educational Psychologist, 42(2)*, 99-107.

Kasurinen, J., & Nikula, U. (2009). Estimating Programming Knowledge with Bayesian Knowledge Tracing. *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education, ITiCSE 2009* (pp. 313-317). New York, NY: ACM Press.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist, 41(2)*, 75-86.

Klahr, D., & Dunbar, K. (1988). Dual search space during scientific reasoning. *Cognitive Science, 12*(1), 1-48.

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.

Koedinger, K., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R. Sawyer, *The Cambridge Handbook of the Learning Sciences* (pp. 61-77). New York, NY: Cambridge University Press.

Koedinger, K., & MacLaren, B. (2002). *Developing a Pedagogical Domain Theory of Early Algebra Problem Solving.* Pittsburgh, PA: CMU-HCII Tech Report 02-100.

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

Kuhn, D. (2005b). What needs to be mastered in mastery of scientific method? *Psychological Science, 16(11)*, 873-874.

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., & M., G.-M. (1992). Cross-Domain Development of Scientific Reasoning. *Cognition and Instruction, 9*, 285-327.

Lee, D., Rodrigo, M., Baker, R., Sugay, J., & Coronel, A. (2011). Exploring the Relationships Between Novice Programmer Confusion and Achievement. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Ed.), *Proceedings of the 4th Bi-Annual International Conference on Affective Computing and Intelligent Interaction, ACII 2011 - Volume Part I. LNCS 6975*, pp. 175-184. Memphis, TN: Springer-Verlag.

Massachusetts Department of Education. (2006). *Massachusetts Science and Technology/Engineering Curriculum Framework.* Malden, MA: Massachusetts Department of Education.

McElhaney, K., & Linn, M. (2008). Impacts of Students' Experimentation Using a Dynamic Visualization on their Understanding of Motion. *Proceedings of the 8th International Conference of the Learning Sciences, ICLS 2008, Volume 2* (pp. 51-58). Ultrecht, The Netherlands: International Society of the Learning Sciences, Inc.

McElhaney, K., & Linn, M. (2010). Helping Students Make Controlled Experiments More Informative. In K. Gomez, L. Lyons, & J. Radinsky (Ed.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010) - Volume 1, Full Papers* (pp. 786-793). Chicago, IL: International Society of the Learning Sciences.

Mislevy, R., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., & Haertel, G. (2003). *Design Patterns for Assessing Science Inquiry.* Menlo Park, CA: SRI International.

Mislevy, R., Steinberg, L., & Almond, R. (2002). On the roles of task model variables in assessment design. In S. Irvine, & P. Kyllonen, *Item generation for test development* (pp. 97-128). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R., Steinberg, L., & Almond, R. (2002). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Montalvo, O., Baker, R. S., Sao Pedro, M. A., Nakama, A., & Gobert, J. D. (2010). Identifying Students' Inquiry Planning Using Machine Learning. In R. Baker, A. Merceron, & P. Pavlik (Ed.), *Proceedings of the 3rd International Conference on Educational Data Mining*, (pp. 141-150). Pittsburgh, PA.

National Research Council. (1996). *National Science Education Standards.* National Science Education Standards. Washington, D.C.: National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards: A Guide for Teaching and Learning.* Washington, D.C.: National Academy Press.

National Research Council. (2011). *A Framework for K-12 Science Education.* Washington, D.C.: National Academies Press.

Pellegrino, J. (2001). *Rethinking and redesigning educational assessment: Preschool through postsecondary.* Education Commission of the States, US Department of Education, Denver, CO.

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

Quinlan, J. (1993). *C4.5: Programs for Machine Learning.* San Francisco, CA: Morgan Kaufmann.

Reimann, P. (1991). Detecting Functional Relations in a Computerized Discovery Environment. *Learning and Instruction, 1*(1), 45-65.

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State-of-the-Art. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 40*(6), 601-618.

Rupp, A., Gushta, M., Mislevy, R., & Shaffer, D. (2010). Evidence-centered Design of Epistemic Games: Measurement Principles for Complex Learning Environments. *The Journal of Technology, Learning, and Assessment, 8*(4), 1-45.

Sao Pedro, M. A., Baker, R. S., Montalvo, O., Nakama, A., & Gobert, J. D. (2010). Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. In R. Baker, A. Merceron, & P. Pavlik (Ed.), *Proceedings of the 3rd International Conference on Educational Data Mining*, (pp. 181-190). Pittsburgh, PA.

Sao Pedro, M. A., Gobert, J. D., & Raziuddin, J. (2010). Comparing Pedagogical Approaches for the Acquisition and Long-Term Robustness of the Control of Variables Strategy. In K. Gomez, L. Lyons, & J. Radinsky (Ed.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences, ICLS 2010, Volume 1, Full Papers* (pp. 1024-1031). Chicago, IL: International Society of the Learning Sciences.

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

Sao Pedro, M., Gobert, J., Heffernan, N., & Beck, J. (2009). Comparing Pedagogical Approaches for Teaching the Control of Variables Strategy. *N.A. Taatgen & H. vanRijn (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1294-1299). Amsterdam, Netherlands: Cognitive Science Society.

Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (1995). Students' Understanding of the Objectives and Procedures of Experimentation in the Science Classroom. *The Journal of the Learning Sciences, 4*, 131-166.

Schauble, L., Klopfer, L., & Raghavan, K. (1991). Students' Transition from an Engineering Model to a Science Model of Experimentation. *Journal of Research in Science Teaching, 28*(9), 859-882.

Schunn, C. D., & Anderson, J. R. (1998). Scientific Discovery. In J. R. Anderson, *The Atomic Components of Thought* (pp. 385-428). Mahwah, NJ: Lawrence Erlbaum Associates.

Schunn, C., & Anderson, J. (1999). The Generality/Specificity of Experise in Scientific Reasoning. *Cognitive Science, 23*(3), 337-370.

Shute, V., & Glaser, R. (1990). A Large-Scale Evaluation of an Intelligent Discovery World: Smithtown. *Interactive Learning Environments, 1*, 51-77.

Shute, V., Glaser, R., & Raghavan, K. (1989). Inference and Discovery in an Exploratory Laboratory. In P. Ackerman, R. Sternberg, & R. Glaser, *Learning and Individual Differences: Advances in Theory and Research* (pp. 279-326). New York, NY: W.H. Freeman.

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills; Instructional effectiveness and path independence. *Cognitive Development, 23*(4), 488-511.

Tsirgi, J. (1980). Sensible Reasoning: A Hypothesis about Hypotheses. *Child Development, 51*, 1-10.

van Joolingen, W., & de Jong, T. (1991). Supporing Hypothesis Generation by Learners Exploring an Interactive Computer Simulation. *Instructional Science, 20*(5-6), 389-404.

van Joolingen, W., & de Jong, T. (1993). Exploring a Domain through a Computer Simulation: Traversing Variable and Relation Space with the Help of a Hypothesis Scratchpad. In D. Towne, T. de Jong, & H. Spada, *Simulation-based Experiential Learning* (pp. 191-206). Berlin: Springer-Verlag.

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.

Veermans, K. (2003). *Intelligent Support for Discovery Learning, Ph.D. Thesis.* Eindhoven, The Netherlands: Twente University Press.

Vygotsky, L. (1978). *Mind in Society: The Development of Higher Psychological Processes.* Cambridge, MA: Harvard University Press.