# Real-time Assessment, Prediction, and Scaffolding of Middle School Students' Data Collection Skills within Physical Science Simulations

By,
Michael Sao Pedro

A Dissertation
Submitted to the PhD committee
in Partial Fulfillment of the Requirements for the
Dissertation Proposal of Doctor of Philosophy
in
Learning Sciences and Technologies

---

April 2013

APPROVED BY:

| | |
|---|---|
| ..... | ..... |
| Dr. Janice D. Gobert<br>WPI- Advisor | Dr. Neil T. Heffernan<br>WPI- Committee Member |
| ..... | ..... |
| Dr. Ryan S.J.d. Baker<br>WPI - Committee Member | Dr. Wouter van Joolingen<br>University of Twente –<br>External Committee Member |

# Abstract

Despite widespread recognition by science educators, researchers and K-12 frameworks that scientific inquiry should be an essential part of science education, typical classrooms and assessments still emphasize rote vocabulary, facts, and formulas. One of several reasons for this is that the rigorous assessment of complex inquiry skills is still in its infancy. Though progress has been made, there are still many challenges that hinder inquiry from being assessed in a meaningful, scalable, reliable and timely manner. To address some of these challenges and to realize the possibility of formative assessment of inquiry, we describe a novel approach for evaluating, tracking, and scaffolding inquiry process skills. These skills are demonstrated as students experiment with computer-based simulations. In this work, we focus on two skills related to data collection, designing controlled experiments and testing stated hypotheses.

Central to this approach is the use and extension of techniques developed in the Intelligent Tutoring Systems and Educational Data Mining communities to handle the variety of ways in which students can demonstrate skills. To evaluate students' skills, we iteratively developed data-mined models (detectors) that can discern when students test their articulated hypotheses and design controlled experiments. To aggregate and track students' developing latent skill across activities, we use and extend the Bayesian Knowledge-Tracing framework (Corbett & Anderson, 1995). As part of this work, we directly address the scalability and reliability of these models' predictions because we tested how well they predict for student data not used to build them. When doing so, we found that these models demonstrate the potential to scale because they can correctly evaluate and track students' inquiry skills.

The ability to evaluate students' inquiry also enables the system to provide automated, individualized feedback to students as they experiment. As part of this work, we also describe an approach to provide such scaffolding to students. We also tested the efficacy of these scaffolds by conducting a study to determine how scaffolding impacts acquisition and transfer of skill across science topics. When doing so, we found that students who received scaffolding versus students who did not were better able to acquire skills in the topic in which they practiced, and also transfer skills to a second topic when was scaffolding removed.

Our overall findings suggest that computer-based simulations augmented with real-time feedback can be used to reliably measure the inquiry skills of interest and can help students learn how to demonstrate these skills. As such, our assessment approach and system as a whole shows promise as a way to formatively assess students' inquiry.



**Keywords**: Science Microworlds, Science Simulations, Science Inquiry, Science Assessment, Performance Assessment, Inquiry Assessment, Computer-Based Assessment, Science Education, Inquiry Learning Environment, Open-Ended Learning Environment, Exploratory Learning Environment, Formative Assessment, Educational Data Mining, Validation, User Modeling, Skill Prediction, Behavior Detection, Designing and Conducting Experiments, Construct Validity, Generalizability, Text Replay Tagging, J48 Decision Trees, Bayesian Knowledge Tracing

*In memory of my beloved grandmother,*

*Amelia "Mitzi" Calo (née DeAngelis)*
*1919 - 1998*

# Acknowledgements

The research in this dissertation was 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.

# Thanks

They say it takes a village to raise a child. Well, I was a little more high maintenance than that, because I needed eleven villages' worth of people (and more) to get me where I am today! Coming back to grad school was one of the most difficult decisions I ever made in my life. I chose to give up a prosperous career in search of something for which I had a true passion. Looking back, I am glad to say it was the best decision I could ever have made during what was a tumultuous time in my life. Before grad school, there were many people who were pillars of strength and support as I made the leap, and many people during this experience who helped take me to the next level, giving me courage, opportunity, kindness, and a heck of a lot of laughs.

First and foremost, I would like to extend my deepest gratitude to my inspirational advisor, Janice Gobert. You were willing to take a chance on an unknown computer scientist and mold me into the learning scientist I have become today, while keeping me true to my roots. You have provided such great guidance and advice, and have been instrumental in giving me all the opportunities to succeed that you could. All of this has enabled me to grow and experience more than I could ever have hoped for during my graduate program. And of course, I commend you for all your patience in working with me. I know I can be a tough cookie! There is just one thing more to ask – are you gonna eat those chips?

Next, I would like to sincerely thank Ryan Baker who also has had a defining influence on my research. You too have taught me so much and given so much to help me succeed even though I was not your own student. I am blessed that our paths crossed, not just from what you taught me, but because you have become such a great friend and ally.

There are two other key people who made this transition possible: Stanley Selkow and Neil Heffernan. Stan has been a good friend and mentor for over 15 years now and he convinced

me to come back to grad school – the best decision I ever made in my life! Neil got me in the door and, like Janice, was willing to take a chance on someone he had not met. He also enabled me to experience firsthand what it was like to be at the forefront of taking new technology to classrooms, wonderful experiences I will never forget.

I also would like to thank Wouter van Joolingen for taking the time to be on the dissertation committee. Your wonderful comments and feedback helped shape the data analysis for this work, and also more broadly shaped my thinking about inquiry.

Next, I would like to graciously thank my mom Denise, my dad Carlos and my sister Nikki for their unconditional love and unyielding support during graduate school, and throughout my life. You all saw where I came from and have journeyed with me through my highest highs and darkest hours. I am blessed to have such a wonderful family that stands by me through thick and thin, and encourages me to task risks in order to achieve more than I could have imagined. I certainly could not have done this without you.

My love Chris, like my family, has been there through all the ups and downs of transitioning out of industry and through a doctorate degree. Chris, thank you for being there to listen, to comfort, to support, and to enjoy with me all this experience has given. You managed to sit on that front seat with me as I rode this rollercoaster without getting too nauseous. Thank you for your patience during those long nights and weekends, and patience for my erratic work habits. I love you dearly with all my heart.

Chris' family, Dan, Barb, Jill, Shaun and Madison, also welcomed me with open arms and showed me much love and support. As a "starving" Ph.D. student, they let me raid the fridge to my heart's content. They also provided many laughs during Whist games, crossword puzzle conundrums, and unforgettable holiday events. There was also plenty of encouragement and

great advice given along the way. Thank you all for making me a true member of your close-knit, wonderful, caring family.

During this time, I lost two people near and dear to me who played huge roles in my life. First, my grandfather, Benjamin A. Calo (1918-2011), passed away after a long struggle with dementia. He was a tough, hard-headed, stalwart, self-made man who devoted his life to his family. He was well-known on our street as the sentinel; he had the uncanny ability to know whatever was going on in a quarter-mile radius on Prospect Ave. at any time of the day or night from his perch on "number 3", the third floor of our home. He and I always butt heads over the years, which in the end, led to a greater respect for each other. I know he wanted nothing more than for me to succeed in whatever endeavor I chose. I can only hope I will have the strength and perseverance to cut down trees and shovel the long driveway when I'm 85 years old. Second, Stephen E. Graham (1969-2010), passed away unexpectedly from SUDEP. Stephen and I met when he was thrust into my office at our old job by my dear friend Amy. From that point, we became very good friends and shared many hobbies together. For example, I owe my advancement of Scrabble skills, enhanced fine dining palate, and "snarkiness" to him. Stephen was a quiet yet sarcastic individual who had a knack for getting people out of their comfort zones to experience all the wonders of life. He also excelled at bringing people closer together. He always knew how to put a smile on your face and how to make you feel special. I have too many great memories to count and I thank you for all the love, support, and wisdom you instilled.

Finally, there are so many friends and family members who have been sounding boards, idea engines, motivators, foundations, inspirations, and kindred spirits. I am truly thankful for having you all in my life and have been blessed by your presence. None of this was possible without your kindness. I list below several "villages" of such great people in alphabetical order.

- *My loving extended family*: The Calos – Uncle Bobby, Auntie Karen, Cousins Robert, Andrea, Kathryn, and Chrystina; The Coughlins – Cousins Caroline, Anthony and Michael; my Godfather Tony Ciampa; my sister Lara Sao Pedro (and son Alex); and my wonderful brother-in-law Niles Welch (and his mom, Sharon)
- *Chris' extended family*: June Donahue, Scott and Judy Donahue (and kids), Cindy, John, Andrew and Brian Farley, Kathy and Don MacNeill, and Dotty Moore
- *Amazing friends from the labs, past and present, and their families*: Samantha Allen, Shane and Grey Almeida (Lily, too), Matt Bachmann, Cameron Betts, Sujith Gowda, Arnon Hershkovitz (and family), dearest Adriana Joazerio (Maria and Chico, too), Kimberly Jung, Nathan Krach, Andy Montalvo, Jaclyn Ocumpaugh (Ken and Will, too), Zach Pardos, Jozsef "Don't worry, you will feel it" Patvarczki, Dovan Rai, Cody Rank, Juelaila "Muh-muh" and Razi Raziuddin (Raashed and Rayyan, too), Leena Razzaq, Lisa Rossi, Cindy Rowan, Sweet San Pedro (of no relation to me ☺), Ermal Toto (and family), Shubendu Trivedi, Manasi Vartak, and Mike Wixon
- *Amazing friends who stood by me when I made this crazy decision*: Kate Baxendale, Billy "Fleeken" Brown, Jeremy Earl Denham (and Nana), Victoria Glazomitsky, Charlie Genaro, Rob Goldstein, Jade Huynh, Bonnie "Ryuu-chan" Leach, Andy and Ken Mallon (and puppies), Noelle and Jim Nichols (and family), and Brad Snow
- *Amazing new friends I made during this journey*: Elaine Alden, "Body By" Krista Baker, Hannah Bary, Hamid Ghadyani, Matt Hinds-Aldrich, Venky Raghavan, Monica Sanchez, Sho Suda, and Martin Werner
- *The original BAE entourage and their families who showed me the ropes*: Chris Barratt, Dan Bostwick, John Everett, Hillary Holloway, Rob and Gretchen Hyland, Steven W. Jilcott III

(and Bethany McCloy), Tanya Khovanova, Basil Krikeles, B. Keith Law, “Kim Possible” Mangino, Conor O’Brien, Brad and Alex Pielech, Alex Proschitsky, Brian Roberts (and Irma Servatius), Brian Shimkin, Greg Sullivan, Jorge Tierno, Michelle Torrelli, Mike Vosseller, Dave Walend, Chuckie Wilcox, Willie Wilson (and Ruth Duncan), and Bijaya and Tim Zenchenko.

- *Professors who guided me along the way*: Joe Beck, Dhiman Bhadra, George Heineman, Kathi Fisler, Chuck Rich, Carolina Ruiz, Suzanne Weekes, Craig Wills, and Marianne Wiser
- *The unsung heroes*: Tammi Chandler, Jennie Clark, Linda Cunningham, Martha McGuane, Marlyn Myers, Ann Marie Serighelli, Kim Thienpont, Kathy Valera, Mike Voorhis, Will Zuidema, the CCC Helpdesk, and all the teachers and students I have worked with over the years.
- *The video game crew*: Martin “The Cat” Bedard, Michelle Bergamo, Greg “Bondo” Bond, Anna and Jason Cram (Ella, too), Tracy and Shawn Cram, Christian Cram and family, PJ “Torch Run” DiCesare, Mr. Kelly and Mrs. Nicole Flewin, Cheryl & Donald “The Robot” Hayes (Mikayla and Matthew, too), JD Lowe, Chris McClard, Nik Meeks, Robert Mruczek, Jenn “Mama Hen” Moore, Dave Nelson, Rusty and Shelley Nunnelee, Nick-O Ortakales, Tom “Rick Rollin’ Bunch of Bananas” Votava, and Stephen Wagner
- *The pinballers*: Sharon and Jeff (Maw and Paw) Blair, Mike and Kathy McGourty, Bill Morrison, Maureen and David “Dr. Dave” O’Neil (Skimo, too), Eric “My Evil Twin” Stone, Bruce “The Cape Crusader” Willard, and “Top Score” Tom
- *The bridge crew*: John Arnold, Megan Cann, Fred Hutson, John Pearsall, Creighton Peet, Mike Pelessari, Rudy Pinkham, John Rodgers, Nathan Sarapas, and many others.

Sincerely, I thank you all. None of this would have been possible without you.

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

Educators and researchers view scientific inquiry as essential to science instruction, and necessary for developing scientific literacy. It has been recognized as critical to science education reform efforts (NRC, 1996, 2000, 2011; Minstrell & van Zee, 2000), and more generally argued to be an essential component toward developing students' capability to seek knowledge, solve problems, and make informed decisions (Kuhn, 2005a, p.5). Such skills are further being recognized as necessary for competing in the 21st century, knowledge-based economy (e.g. Clarke-Midura et al., 2011). Thus, there is a strong push for a paradigm shift from science pedagogy that emphasizes rote learning of vocabulary, facts, and formulas in service of high-stakes tests to pedagogies that promote active, integrative learning of content, skills, and critical reasoning via scientific inquiry.

Despite inquiry's importance, hands-on inquiry activities are seldom used in today's schools. One of several reasons is that assessment of authentic inquiry skills requires the use of performance-based assessments to accurately capture skill (Black, 1999; Pellegrino, 2001; Quellmalz et al., 2007; Clarke-Midura et al., 2011). Though computer-based approaches are now beginning to tackle assessment of inquiry (e.g. Clarke-Midura et al., 2011; Quellmalz et al., 2012), assessment is still difficult to execute in a reliable, valid, and scalable manner. The difficulty is rooted in the complexity and multi-faceted nature of the inquiry skills themselves and how they could be demonstrated in open-ended tasks (Williamson et al., 2006). That there is no single "right" or "wrong" way to conduct inquiry makes its measurement challenging. Furthermore, if inquiry is assessed, it is typically done so in a summative manner, as an activity separate from learning. Results from assessments, performance-based or otherwise, can take too long to generate (NRC, 2006). Therefore, teachers are not provided timely feedback they can

utilize to help them formatively assess their students, information that can enable them to pinpoint students' difficulties.

To combat these challenge and promote learning, a web-based learning environment, Inq-ITS (Gobert et al., 2012) is being developed with the goal of automatically assessing and honing middle school students' inquiry skills. In this environment, middle school students conduct inquiry within simulations designed to address specific content standards in the Massachusetts Curriculuar Framework (Massachusetts Department of Education, 2006). The simulations span a variety of science domains including Physical Science, Life Science, and Earth Science. The environment aims to track the development of inquiry skills defined in the National Science Education Standards (NRC, 1996) across science domains; and, in the long term, aims to scaffold those students who need support in real-time as they experiment.

This dissertation has two objectives towards realizing the vision of the Inq-ITS learning environment (Gobert et al. 2012). First, we develop scalable and reliable measurement and tracking of skills related to designing and conducting of experiments (NRC, 1996), skills demonstrated when collecting data. Second, we analyze the complex relationship between inquiry tutoring, and skill acquisition and transfer of these skills across science topics. In this work, we focus specifically on two data collection inquiry skills, designing controlled experiments and testing stated hypotheses. Briefly, students design controlled experiments when they generate data that make it possible to determine what the effects of independent variables (factors) are on outcomes. They test their stated hypotheses when they generate data that can support or refute an explicitly stated hypothesis. This work focuses specifically on assessing these skills within Inq-ITS Physical Science simulations.

Achieving the first objective of measuring and tracking the two data collection skills requires proper *evaluation* of students' interactions as they collect data. In other words, as students experiment with a simulation, how can the learning environment tell if they are designing controlled experiments, testing their hypotheses or off-track? To evaluate students' experimentation, we developed and validated data-mined detectors (models). The detectors leverage student logfiles, and human judgment of whether their logfiles demonstrate the skills (Baker et al., 2006) to construct data-mined models that can replicate that human judgment. We directly address reliability and scalability by testing the predictive capabilities of these models using student data not used to construct them (cf. Efron & Gong, 1983; Witten & Frank, 2005). In this work, the detectors also have additional purposes besides automatically scoring whether or not students design controlled experiments and/or test their hypotheses.

One other use for the detectors is to provide a foundation for measuring students' latent skill over a series of activities. In other words, when students use the simulations multiple times, how do we aggregate the evaluations to estimate if they how to design controlled experiments or test their hypotheses? The detectors enable evaluation of a single data collection activity, but do not provide a means to amalgamate several attempts at data collection into overall measures of proficiency. To estimate latent student proficiency, we constructed and validated aggregate models of latent skill using Bayesian Knowledge-Tracing (Corbett & Anderson, 1995), an approach that has been successful for intelligent tutoring systems (e.g. Koedinger & Corbett, 2006; Feng et al., 2009).

The mechanisms to evaluate, track, and scaffold students' inquiry in real-time enable us to achieve our second objective of exploring the relationship between scaffolding, and skill acquisition and transfer. An automated scaffolding system first requires that students' actions can

be evaluated. The detectors can be used to trigger scaffolding since they can evaluate students' actions as they work. A scaffolding system also requires the authoring of specific feedback aimed at helping students when they are off track. As part of this dissertation, we developed a number of scaffolds to provide assistance on these data collection skills. These supports range from more broad, abstract support to more targeted support (cf. Veermans, 2003), aimed at honing those skills of interest.

We also aim to determine if our scaffolding approach is effective in helping students learn how to design controlled experiments and test their stated hypotheses. To that end, we address two research questions regarding the acquisition and transfer of skill:

- Does scaffolding data collection within a domain improve students' data collection skill compared to a condition with no data collection scaffolding, controlling for time and initial inquiry skill?
- Will scaffolding of data collection skill in one domain give students a "leg up" in authentic data collection skill in a second domain when the scaffolds are removed?

These questions are of importance for three reasons. First, many studies have shown that students have difficulty with inquiry (de Jong & van Joolingen, 1998), so finding ways to improve acquisition and transfer of inquiry skill is of great interest to researchers and educators. Second, it is an empirical question whether inquiry skills can transfer across domains, or if they are tied to the domain in which they were taught/learned (van Joolingen et al., 2007). This question has important ramifications for assessing inquiry using unfamiliar contexts and for teaching inquiry. Finally, it is an empirical question as to how best to foster the acquisition and transfer of inquiry skills (cf. Kuhn, 2005b; Kirschner, Sweller, & Clark, 2006; Hmelo-Silver, Duncan, & Chinn, 2007). Evidence exists that repeated practice can lead to increased skill within a domain (Kuhn,

Schauble & Garcia-Mila, 1992; Dean Jr. & Kuhn, 2006), and transfer between domains (Kuhn, Schauble & Garcia-Mila, 1992). Scaffolding or teaching inquiry skills can also lead to successful acquisition (e.g. Klahr & Nigam, 2004; Kuhn & Pease, 2008; Sao Pedro et al., 2009; Sao Pedro et al., 2010b; Siler, et al., 2010) and transfer to novel tasks (Klahr & Nigam, 2004). In this work, we determine if scaffolding the designing controlled experiments and testing stated hypotheses skills leads to improved acquisition and transfer of these data collection skills across two physical science topics, Phase Change and Kinematics.

The remainder of this thesis is organized as follows:

- Chapter 2 presents background and related research for this proposal. It begins with the role of scientific inquiry in science education and challenges to assessing inquiry using performance-based activities. Then, computer-based systems similar to Inq-ITS that leverage simulations and microworlds as a means for students to practice inquiry are described. Next descriptions of the We then describe the two data collection skills of interest, designing controlled experiments and testing stated hypotheses, and prior work on measuring them. Since the aim of this dissertation is to leverage data mining techniques to evaluate, track and scaffold these inquiry skills, we complete the literature review with relevant prior work on assessing ill-defined skills, and scaffolding students in computer-based environments.
- Chapter 3 presents the Inq-ITS system and describes the physical science activities utilized in this report to elicit students' inquiry skills.
- Chapter 4 discusses our data mining-based approach to developing detectors of the two data collection behaviors associated with the skills of interest. In particular, we present the details of how these detectors were constructed, starting from students' low-level interactions with the system. We also describe how we validated them using data not used in their construct

them, revised them to increase their predictive capability by increasing their construct validity, and tested their generalizability. The detectors are used in the subsequent chapters to evaluate students' skills, drive scaffolding, and address the efficacy of the scaffolding approach.

- Chapter 5 presents an initial attempt at using Bayesian Knowledge Tracing to aggregate in a principled way students' performances across inquiry activities. We address the degree to which these models can predict student performance on their next inquiry activity as well as performance on other transfer tests of inquiry.
- Chapter 6 describes our scaffolding approach and presents a study for addressing if that approach improves students' acquisition and transfer of inquiry skills. The analyses make use of the detectors' evaluations of whether students demonstrate inquiry skills (Chapter 4) and estimates of students' latent skill over time (Chapter 5).
- Chapter 7 presents an alternate approach for discerning the impacts of scaffolding by extending the Bayesian Knowledge Tracing approach from Chapter 5. In particular, the framework is extended by incorporating scaffolding and change of context (e.g. switching inquiry topics). We also explore in this chapter whether these extensions also improve the models' predictive capabilities.
- Chapter 8 presents a final summary of our findings, broad implications of these results, and future work.

# 2 Background

## *2.1 Role of Scientific Inquiry in Science Education*

Contemporary views of science education have grown beyond defining scientific competency as the skill at answering factual questions (National Research Council, 1996, 2000, 2011; Settlage & Southerland, 2007; Clarke-Midura et al., 2011). Today, such literacy from the view of educational research involves deep understanding and application of knowledge (Kuhn, 2005a,b). According to these views, students should become proficient in applying ideas to diverse contexts and should understand multiple representations such as equations, diagrams, or textual descriptions of a concept. Furthermore, students should learn how to challenge claims, how to become proficient in researching solutions to their own questions, and how develop an awareness of how their ideas change over time. Importance is also placed on summarizing all the previously mentioned information together in order to communicate ideas to others. The backbone of developing all these capabilities is solid scientific inquiry skills (Duschl, Schweingruber, & Shouse, 2007).

Generally speaking, scientific inquiry embodies the skills underlying fluency at using scientific reasoning to develop understanding within a science discipline (NRC, 1996, 2000). It is seen as necessary for developing general scientific literacy and reasoning skills (Kuhn, 2005a), and nowadays, critical for 21$^{st}$ century jobs in a knowledge-based economy (Clarke-Midura et al., 2011). Aside from being central to the field of science, inquiry can also act as a support for learning new science material (e.g. White, 1993; White & Frederiksen, 1998; Buckley et al., 2006).

Despite the importance and benefits, schools often do not focus on inquiry[1]. There are several possible reasons for this. First, inquiry activities are costly and usually require physical materials that many schools cannot afford. Second, the textbooks teachers use emphasize content knowledge, the "end results" of inquiry, as opposed to the processes of how that knowledge came to be (Eltinge & Roberts, 1993). Third, teacher preparation programs do not provide adequate examples of how to teach inquiry, nor do they provide teachers adequate opportunity to implement inquiry as part of their field experience (Barrow, 2006). Compounding this is the lack of consensus on how best to foster these skills, specifically if practices should be discovered by students on their own or directly taught to them (Klahr & Nigam, 2004; Kuhn, 2005b; Dean Jr. & Kuhn, 2006; Kirschner, Sweller, & Clark, 2006; Hmelo-Silver, Duncan, & Chinn, 2007). Finally, inquiry is generally not emphasized within the classroom because rote knowledge is and has been prioritized on high-stakes tests (Settlage & Southerland, 2007, p.280-282) like Massachusetts' MCAS exam (Massachusetts Department of Elementary and Secondary Education, 2008).

Going forward, inquiry will likely become more of a focus because influential organizations like NAEP, PISA, NETP, NRC, and PCAST are touting and recognizing its importance. A push is therefore beginning towards developing better materials for practicing, and automatically assessing inquiry. Since our work addresses inquiry assessment, we discuss some of these approaches in the next section.

[1] Barrow (2006) provides a nice overview of issues related to implementing inquiry within classrooms.

## 2.2 General Approaches and Challenges to Assessing Scientific Inquiry

Given the call to incorporate inquiry into science classrooms, there is also growing interest in how to properly and reliably measure students' inquiry skills in order to give feedback to interested stakeholders like parents and teachers in a timely manner. Typically, inquiry is assessed using paper-based tests like short answer or multiple choice exams (cf. Alonzo & Aschbacher, 2004). Such tests are now widely criticized because they cannot measure students' inquiry process skills (Black, 1999; Pellegrino, 2001; Quellmalz et al., 2007; Clarke-Midura et al., 2011; NRC, 2011, pp.262-263; Gobert et al., 2012, under review) for several reasons. These reasons include:

1) They are better suited to capture conceptual knowledge, not information about students' strategies and reasoning (NRC, 2006; Gobert et al., 2012);
2) They do not align well with current standards (Quellmalz et al., 2007);
3) They do not align well with how inquiry process should be taught (Clarke-Midura et al., 2011); and
4) They do not provide scenarios or situations that elicit students' inquiry reasoning and processes (Clarke-Midura et al., 2011), even though attempts have been made to do so (e.g. White & Frederiksen, 1999).

Therefore, performance assessments that require students to "do" inquiry are required in order to tap their skills (Baxter and Shavelson 1994; Ruiz-Primo and Shavelson, 1996; Clarke-Midura et al., 2011).

One kind of performance assessment used in state assessments is hands-on performance assessments (Linn, 1994). In such activities, students engage in authentic practices using

physical materials. Such assessments, though more authentic than multiple choice, have drawbacks regarding their scalability, reliability, and validity in terms of what they can measure. In terms of scaling, they are cost-prohibitive (Stecher & Klein, 1997) due to the cost of materials, and because scoring is laborious (NRC, 2006). Due to these limitations, it is typical for such performance assessments to be given only once (Shavelson et al., 1999). However, doing so calls into question the degree to which they then measure skill since they require more data to measure reliably (Shavelson et al., 1999). Finally, in terms of validity, these performance assessments are typically highly structured and thus limit the investigation strategies that can be measured (Quellmalz et al., 2007). Furthermore, since they are hands-on, it is very difficult to capture students' inquiry processes (cf. Rupp, 2010), the steps and actions they take as they experiment, since it is very difficult to "play back" what students did during their inquiry (unless one uses a video tape). Even then, their strategies are not necessarily exposed.

As an alternative, national and international organizations that focus on educational frameworks and assessment (e.g. NAEP, PISA, NETP, NRC) have now acknowledged the potential of computer-based environments as a scalable approach to overcome some of these limits and challenges. Deploying virtual activities on computer-based environments addresses scalability, because fewer materials are needed (just a computer and the software). Furthermore, since activities are computer-based, it is easier to ensure all students receive the same experience when engaging with the materials. Also, students' low-level actions within these activities can be captured in logfiles and can be used to provide insight on students' inquiry processes (e.g. Buckley, et al. 2006; Buckley, et al., 2010; Gobert et al., 2007). However, these environments are not yet used for performance assessment to a substantial degree because there are still several challenges pertaining to measurement and reliability (Quellmalz et al., 2009).

The crux of these issues has to do with the complexity of the log files generated within these environments as Williamson et al. (2006) identify. Complex tasks like those found in performance-based inquiry activities require students to take many steps to solve as compared to multiple choice items requiring a single answer. Because tasks are more complex, students may demonstrate skills in multiple ways, and thus there is wider variability in students' answers. Using inquiry as an example, though effective strategies have been identified (cf. Schunn & Anderson, 1998), there is no one exact "right or wrong" way to demonstrate inquiry skill. In addition, performance (observations) of multiple tasks in an activity may need to be aggregated to determine if students demonstrate and/or know skills. Finally, tasks and skills may not be independent from each other, and thus, assumptions of conditional independence (as made in Classical Test Theory, for example) do not hold (Mislevy et al, 2012). Thus, the complexity of performance-based tasks, like inquiry, makes measuring skills using conventional methods difficult.

In terms of feedback to stakeholders, inquiry assessments are typically summative, irrespective of being multiple-choice, hands-on, or computer-based. These assessments only permit feedback typically weeks or months after they have been completed, too late for effective and timely intervention by the teacher. It is now acknowledged that the same computer-based systems used to assess inquiry may also enable formative assessment and individualized support as students work (cf. Clark-Midura et al., 2011). However, little has been done with the goal of developing an intelligent tutoring system for scientific inquiry, except for a proof of concept paper that outlined a possible system (Koedinger et al., 1999).

In our work, we aim to develop a computer-based inquiry assessment environment capable of providing students individualized feedback as they work. We focus specifically on two skills

related to "designing and conducting experiments / data collection" subskills (NRC, 1996). Skills are measured as students collect data within simulations/microworlds. To ground this work, we first present similar systems that utilize simulations and microworlds to teach, assess and support inquiry. Then, we identify the inquiry subskills of interest and discuss prior work on measuring them in simulation-based environments.

## 2.3 Systems that Assess and Scaffold Scientific Inquiry within Simulations

There are a plethora of software systems that support students' scientific inquiry. Since the focus of our work entails scaffolding and assessment of scientific inquiry skills as students engage in inquiry within physical science simulations/microworlds[2], we review here similar systems. Briefly, microworlds (cf. Papert 1980; 1993) are " ...subset[s] of reality or a constructed reality whose structure matches that of a given cognitive mechanism so as to provide an environment where the latter can operate effectively. The concept leads to the project of inventing microworlds so structured as to allow a human learner to exercise particular powerful ideas or intellectual skills" (Papert, 1980, p. 204). In regards to science education, they enable students to study scientific phenomena in a dynamic, interactive way. With microworlds, students can observe concepts at various levels of abstraction, highlighting aspects that may normally be unobservable without expensive equipment. Microworlds enable students to perform virtual

[2] We note that we are not making an explicit distinction between microworlds and simulations. Typically, microworlds and simulations differ in the amount of control a user has within the learning environment. Microworlds more resemble programming activities and engage students more in "model building", whereas simulations more reflect the simulation designer's goals and engage students more in "model using" (Penner, 2000/2001; Rieber, 2005). The environments in which I will assess inquiry more closely resemble simulations along the microworld-simulation continuum, though they will be referred to interchangeably as "microworlds" and "simulations".

experiments that may normally require too many expensive physical apparatuses, but still afford authenticity because they share many features with real apparatuses (Gobert, 2005). Additionally, they can enable time-stepping and playbacks to see how an object's or system's properties changes over time.

In the following sub-section, we review others' approaches to assessing and scaffolding inquiry, highlighting important findings on inquiry skill acquisition. The systems reviewed include: Thinkertools (White, 1993; White & Frederiksen, 1998), Modeling Across the Curriculum (MAC) (e.g. Gobert, 2005), WISE Airbags (McElhaney & Linn, 2008; 2010), SimQuest (van Joolingen & de Jong, 2003), SCY: Science Created By You (de Jong, et al., 2010), and SimScientists (Quellmalz et al., 2009, 2012).

### 2.3.1 Thinkertools

The Thinkertools simulation environment (White, 1993) aims to improve students' understanding of Newtonian forces and motion by having them discover relevant concepts through exploration within a series of activities involving microworlds. Briefly, each microworld includes a game-like "task" or "goal" a student must achieve. These goals are meant to promote understanding of concepts, alternative representations of motion, and discovery of the causal relationships that govern the microworld. Each activity progressively incorporates more sophisticated models of how forces affect the motion of objects. Thus, students' prior knowledge is naturally built upon to develop a deeper, more complex understanding of the domain.

Though White (1993) developed some inquiry supports to align with Thinkertools, the inquiry process and supports for it play a more prominent role in later work (White and Frederiksen, 1998). Here, students partake in inquiry cycles and reflection over several weeks,

conducting experiments within Thinkertools and the real world. The inquiry cycles are broken into several steps to scaffold students' understanding: questioning, predicting (hypothesizing), experimenting, modeling (mathematizing), and applying models. Each cycle becomes less scaffolded over time, starting from more rigid sequences of steps students must follow, ending to open-ended research projects. Scaffolding and instruction comes from paper materials that students follow as they execute inquiry cycle steps, and from classroom-wide activities monitored by the teacher. Thus, the Thinkertools software itself does not automatically assess skills and does not provide automated support during students' exploration. Also, all students are scaffolded in the same way, with the same materials.

These two studies have several important findings with regard to content knowledge and acquisition of inquiry skills. I highlight a few key results as follows. In terms of content acquisition, White (1993) found that middle school students in the Thinkertools curriculum significantly outperformed high school-level students who studied Newtonian mechanics in a normal classroom setting on a conceptual transfer test. In terms of inquiry skill acquisition, White and Frederiksen (1998) found that students, overall, had a 0.8 sigma gain on their pen-and-pencil inquiry posttest. This test measured skill at hypothesizing, experimenting, analyzing data, forming conclusions, and overall consistency of their argument in linking together the various steps of their inquiry. Finally, it was found through correlational analyses that inquiry skills can play a significant role in the development of students' learning and understanding of physics concepts. Putting it all together, these studies show that robust understanding of inquiry and physics content can be produced by combining microworld activities with real-world and virtual experimentation, and scaffolded inquiry cycles.

### 2.3.2 Modeling Across the Curriculum (MAC)

The MAC project (Gobert, 2005; Buckley, et al. 2006; Buckley, et al., 2010; Gobert et al., 2007; Levy & Wilensky, 2006) developed a number of computer-based science activities to help students learn content in several science domains through exploration. These content areas include biology (BioLogica), physics (Dynamica), and chemistry (Connected Chemistry; Levy & Wilensky, 2006). Typical activities give students the freedom to explore a microworld in order to answer a given target question. Similar to Thinkertools (White 1993; White & Frederiksen, 1998), activities are sequenced to foster progressive model-based learning (White & Frederiksen, 1998; Gobert & Buckley, 2000) students are initially presented simple, static models and eventually engage in activities that include more complex, dynamic models. Unlike Thinkertools, however, all inquiry activities and occur within the computerized system. This enabled tracking of students' authentic inquiry skills via log files.

Using data from the log files, cognitive task analyses were performed to produce knowledge engineered rules encapsulating differing levels of systematic experimentation skill (Buckley et al., 2006, 2010; Gobert et al., 2007), and specific inquiry behaviors (Buckley et al., 2006, 2010). They leveraged these models to then study the relationship between students' systematic inquiry behavior and content acquisition. For example, Buckley et al., (2006, 2010) showed that systematic inquiry demonstrated within microworld-based activities in the domain of genetics positively students' acquisition of content knowledge, as measured by pre- and post-test gains. Specifically, they found that systematic performance on certain inquiry tasks within BioLogica, one of their microworld activities, predicted about 10% of the variance in students' post-test gain scores, irrespective of whether they actually succeeded at the inquiry task during

learning. In Dynamica, Gobert et al. (2007) identified strategic approaches to inquiry tasks that had significant positive correlations with post-test conceptual gains. In Connected Chemistry, Levy and Wilensky (2006) found that model exploration during inquiry led to greater conceptual gains.

### 2.3.3 WISE and Airbags: Too Fast, Too Furious?

WISE, the Web-based Inquiry Science Environment (Linn, Davis & Bell, 2004; Slotta & Linn, 2009), is a platform that aims to make scaffolded inquiry projects accessible, modifiable, and shareable to teachers and students all over the world. Typical WISE projects are sub-divided into steps in a sequence where each step involves various activities like answering questions, discussing, reading webpages, drawing graphs, developing concept maps, creating presentations, and exploring within simulations. Of relevance to our work, McElhaney and Linn (2008, 2010) designed a one-week curriculum for high school physics within WISE in which students explored which factors could impact airbag safety.

Similar to the MAC project (Buckley, et al. 2006; Buckley, et al., 2010; Gobert et al., 2007; Levy & Wilensky, 2006), McElhaney and Linn (2008) devised knowledge engineered rules to code students' experimental behavior. They then used these models to study the relationship between experimentation behaviors, skill at understanding graphs, and content acquisition. Specifically, they analyzed the effects of three exploration behaviors. The first was total number of trials. The second was trial variability, a measure to determine what range of values for each variable was tested). The third was experimentation validity, a **hand-scored** measure of the extent to which students controlled for variables (cf. Chen & Klahr, 1999) and were consistent with a chosen investigation question. They found that students who conducted

valid experiments (those in which valid inferences could be made from data) tended to conduct fewer trials, but learned more content knowledge from the Airbags unit, as measured by pre-post comparisons. No differences were found in understanding graphs in pre- versus post- test measures. The authors note that high scores on experimentation validity reflects several skills: controlling for variables, successfully mapping questions onto experimentation variables, correctly interpreting outcomes, and planning investigations in advance.

### 2.3.4 SimQuest

The SimQuest learning environment (van Joolingen & de Jong, 2003) is a discovery-based learning environment in which students construct their own conceptual knowledge in science domains by engaging in inquiry with simulations. In a typical activity, students are presented with one of several goals: set up a simulation in a certain way, investigate the relationship between variables in the simulation, predict values of variables of the simulation, or answer target questions using the simulation as a supporting tool. This learning environment also attempts to strike a balance between student and system control of learning activities. For example, the system can be scripted to let students pick which activities to do in any order, or the system could force students through a particular order.

Research involving SimQuest has been conducted by Veermans (2003) on augmenting SimQuest to provide feedback on several inquiry processes (Njoo & de Jong, 1993). The author focused specifically on testing predefined hypotheses and drawing conclusions about these hypotheses. Depending on students' actions, Veermans' system provided verbal feedback on several possible exploration heuristics which may aid in conducting inquiry. Example heuristics include: generating a small amount of data to glean an idea for a hypothesis, controlling for

variables, testing the effects of continuous variables at equal increments, testing extreme values of variables, and generating enough trials to fully test hypotheses. These scaffolds were intended as advice, meaning that the system did not force students to follow this feedback.

Veermans (2003) also conducted two studies to test the efficacy of this scaffolding on students' acquisition of knowledge in the physics topic of collisions. The first study compared students who received feedback on their experiments to a control condition who did not. Overall, no differences between the conditions were found on their knowledge posttests. In a second study, Veermans changed the feedback to include not only the guidelines for how to follow the heuristics, but also the rationale and name of the heuristic. Two conditions were again compared in which students received implicit scaffolding (just the guidelines) and explicit scaffolding (guidelines, rationale, and heuristic name). Again, however, no differences between groups were found. Veermans notes, though, that this may have been due to a ceiling effect in posttest scores.

### 2.3.5 Science Created By You (SCY)

SCY is a sequel project to SimQuest that enables students to conduct inquiry in larger research projects (de Jong, et al., 2010). Unlike SimQuest, there is a particular focus on inquiry processes, products resulting from those processes (called Emerging Learning Objects, or ELOs for short), and collaboration. The system aims to provide feedback to students' processes and work products by analyzing student actions and providing suggestions. Relevant to our work, SCY has one set of activities that utilize simulations. In these activities, students change simulations and run trials to produce a data collection ELO.

As mentioned, a goal of SCY is to provide feedback to students about their inquiry. Although results on the efficacy of providing this feedback have not yet been published,

Weinbrenner et al. (2010) developed pedagogical agents based on knowledge engineered rules for two data collection behaviors. The first agent determined the degree to which a sequence of actions demonstrates CVS (varying one thing at a time). The second agent determined if students collected data for a continuous-valued simulation variable at equal intervals.

### 2.3.6 SimScientists

SimScientists (Quellmalz et al., 2009, 2012) is an assessment environment that, like Inq-ITS, aims to formatively assess middle school-level inquiry skills and science concepts using simulations and other performance-based activities. Currently, simulations are developed for ecosystems (Life Science) and forces and motion (Physical Science). In regards to content measurement, the assessments aim to measure understanding of cross-cutting concepts like "components and their roles", "interactions among components", and "emergent behaviors of systems" (Timms et al., 2012, p.14). Evidence for these constructs is obtained by traditional questions and more interactive questions. These interactive questions require students to, for example, build food webs and collect data with a simulation (Timms et al., 2012, p.15).

Quellmalz and colleagues (2009, 2012) have employed a variety of methods to tease apart content knowledge from inquiry knowledge. For example, Quellmalz et al. (2009) used Bayesian Networks to link actions taken in the ecosystems simulation with latent measures of inquiry skills and content knowledge. In this model, each latent construct has four levels, "advanced", "proficient", "basic" and "below basic", and probabilities associated with them to capture student progress on that measure. Quellmalz et al. (2012) have also used confirmatory factor analysis and multidimensional IRT to measure inquiry skills.

## 2.4 Data Collection Skills of Interest

Skills related to designing and conducting experiments (cf. NRC, 1996) are of particular importance because they have been argued to support the development of other scientific inquiry skills like 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). Furthermore, it has been shown that students typically have difficulty with these skills (de Jong & van Joolingen, 1998). For example, when experimenting, students may not collect data that test their articulated hypotheses (van Joolingen & de Jong, 1991, 1993; (Kuhn et al., 1992; Schauble, Klopfer & Raghavan, 1991). They may only run one trial (Kuhn et al. 1992), run the same trial repeatedly (Kuhn et al., 1992; Buckley et al., 2006), or change too many variables (Glaser et al., 1991; Reimann, 1991; Tsirgi, 1980; Shute & Glaser, 1990; Kuhn, 2005b; Schunn & Anderson, 1998, 1999; Harrison & Schunn, 2004; McElhaney & Linn, 2008, 2010). They may also run experiments that try to achieve an outcome (i.e., make something burn as quickly as possible) or design experiments that are enjoyable to execute or watch (White, 1993), as opposed to attempting to understand the effects of variables (Schauble et al., 1991, 1995; Njoo & de Jong, 1993). In this work, we focus specifically on two 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 (e.g. with a simulation) 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).

Others have analyzed the acquisition of skills related experimentation within computer-based learning environment, particularly for designing controlled experiments and learning the CVS. For example, several have researched the effectiveness of environments meant to teach CVS in isolation from other inquiry skills (e.g. Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Zohar & David, 2008; Sao Pedro et al., 2009; Sao Pedro et al., 2010b; Siler et al., 2010). Additionally, some have analyzed performance at designing controlled experiments during open-ended inquiry tasks involving simulations, as I am proposing to do within Inq-ITS (e.g. Shute & Glaser, 1990; Schunn & Anderson, 1998; Koedinger et al., 1999; Harrison & Schunn, 2004; Kuhn & Pease, 2008; McElhaney and Linn, 2008, 2010; Gobert &

Koedinger, 2011). Many of these approaches developed methods to score if students were designing controlled experiments using *knowledge-engineered rules*.

The knowledge-engineered rules entail different ways of measuring how many times the student ran the simulation (thus collecting data) and counting controlled trials in that dataset. For example, McElhaney and Linn's (2008, 2010) approach determines if students design controlled experiments by examining the number of successive pairwise trials that are controlled (e.g. does trial 2 vs. trial 3 only have one variable change between them?). Gobert & Koedinger (2011)'s approach is more lenient; they count if *any* two trials are controlled (e.g. trial 1 vs. trial 4, trial 2 vs. trial 3, etc.). Harrison and Schunn (2004) computed average numbers of trials that were controlled.

By contrast, we propose constructing data-mined detectors that labels students' overall behavior as demonstrating the skill (or not), and considers several more attributes than successive pairwise controlled experiments. We will discuss our reasoning why and the process for doing so in Chapter 4. Part of our rationale is that these skills there are several valid ways they can be demonstrated as students collect data, and as such, are ill-defined. In the next section, we describe prior work on evaluating ill-defined domains within Intelligent Tutoring Systems.

## 2.5 Prior Work on Evaluating and Assessing Ill-Defined Skills in Intelligent Tutoring Systems

A large amount of work has gone into developing learning environments and intelligent tutoring systems (ITS) for well-defined domains. Examples include problem solving in mathematics (Koedinger & Corbett, 2006; Razzaq et al., 2005; Feng et al., 2009), physics (Gertner & VanLehn, 2000), genetics (Corbett et al., 2010), computer programming tasks (Corbett &

Anderson, 1995; Kasurinen & Nikula, 2009), and reading (Beck & Chang, 2007). These environments are designed to elicit student responses in which evaluation of student work is relatively straightforward. More specifically, the solutions to problems in these environments are well-defined (e.g. the solutions to the equation $x^2 + 9 = 0$, a word is correctly read, …), or the paths or actions a student follows in the environment to solve a problem are finite and explicit (e.g. model tracing, Anderson et al., 1990; Anderson, 1993). Since evaluation of student work or solution steps can be made clear, researchers instead turn to broader assessment issues, developing models that more accurately predict learning and student understanding, students' affective states, and students' metacognitive states (for a review, see Desmarais & Baker, 2012). For more ill-defined domains, however, assessment is more difficult because evaluation of students' solutions is ambiguous and open to interpretation (Voss, 2006; Lynch et al., 2009).

Much progress, though, has been made towards assessment for such domains. One set of examples comes from evaluating students' free text responses to open-ended questions. For example, Graesser et al. (2005) assessed conceptual physics knowledge by matching student responses against required components for a correct answer as well as known misconceptions. Text mining has also been used to evaluate students' explanations to provide support to students working on collaborative tasks (Kumar & Rose, 2011). In design-based tasks in which students' work products are evaluated, constraint-based modeling (Mitrovic & Ohlsson, 1999; Mitrovic et al., 2003) has successfully been used. For example, SQL-Tutor (Mitrovic, 2003), an ITS that teaches students how to write database queries, uses constraint-based assessment to evaluate the correctness of those queries. Rather than forcing students to describe exactly one query to solve a problem, the constraint-based approach lists out criteria a solution must have (or violations of criteria), enabling many queries to be considered correct, even though they are not explicitly

listed as solutions. Similarly, Roll, Aleven and Koedinger (2010) used constraint-based modeling to evaluate students' understanding of data variation and used student responses to generate new problems to address students' misconceptions automatically.

Rather than authoring components or aspects of solutions, models to evaluate students' skills, learning strategies, or behaviors can also be *learned* using data mining techniques (cf. Romero & Ventura, 2007; Baker & Yacef, 2009) from student interaction data. As mentioned, in our work we will employ this strategy for generating evaluation models for the two inquiry skills of interest. Next, we describe others' approaches for employing data mining techniques for such purposes, particularly for evaluating inquiry, to situate our work.

## 2.6 Prior Work on using Data Mining to Assess Skills and Complex Behaviors within Computer-Based Learning Environments

In our approach, we aim to develop machine-learned assessment and proficiency estimate models of two data collection subskills, designing controlled experiments and testing stated hypotheses. In a similar vein, previous research has successfully used machine learning techniques to distinguish students' problem solving strategies within exploratory learning environments. For example, Bernardini and Conati (2010) used clustering and Class Association Rules to capture learner models of effective and ineffective learning strategies within an environment for learning about a constraint satisfaction algorithm. Ghazarian and Noorhosseini (2010) constructed task-dependent and task-independent machine-learned models to predict skill proficiency in computer desktop applications.

Research has also been conducted on using machine learning techniques to model competency and knowledge within inquiry environments. Stevens et al. (2004) used self-

organizing artificial neural networks to build models of novice and expert performance using transition logs within the HAZMAT high school chemistry learning environment. They then leveraged those models to construct a Hidden Markov Model for identifying learner trajectories through a series of activities. Rowe and Lester (2010) developed Dynamic Bayesian Network models of middle school students' narrative, strategic and curricular knowledge as students they explored within a 3D immersive environment on microbiology, Crystal Island. Finally, Shores, Rowe and Lester (2011) compared machine learning algorithms' efficacy at predicting whether students would utilize a particular inquiry support tool shown to improve learning within that same environment.

The work presented here differs from this earlier work in one key fashion. Whereas previous work has looked for general indicators of problem solving skill in inquiry environments (Stevens, et al., 2004; Rowe & Lester, 2010), or predictors of whether students will use cognitive support tools (Shores et al., 2011), this work develops models of specific inquiry subskills (cf. NRC, 1996) and tracks them over a series of activities.

It is worth noting that several others have successfully utilized machine learning techniques to model and detect other complex learner behaviors within learning environments. Beck (2005), for example, developed an IRT-based model incorporating response times and correctness to predict disengagement in an approach called "engagement tracing". Cocea and Weibelzahl (2009) labeled raw log files, distilled features, and then built and compared several models of disengagement yielded by different machine learning algorithms. Cetintas et al. (2009) used a combination of timing features, mouse movements unique to each student to build off-task behavior detectors. Walonoski and Heffernan (2006) and Baker et al. (2008a) successfully built and validated gaming the system detectors by triangulating qualitative field observations

with features gleaned from log files. And finally, Baker and de Carvalho (2008) and Baker, Mitrovic and Mathews (2010) labeled the gaming the system behavior using text replays, which also led to successful detectors under cross-validation (cf. Efron & Gong, 1983).

Our work on developing models to assess students' data collection inquiry skills is similar to these projects in that we follow a similar paradigm to construct detectors of inquiry skills. Specifically, we leverage the success of Baker and de Carvalho's (2008) and Baker et al.'s (2010a) use of human labeling of log files as our method of classifying training instances. We will discuss this approach in more detail in Chapter 4.

## 2.7 Automated Learning Support in Computer-based Learning Environments

A wide variety of techniques have been employed to provide students' support as they learn computer-based learning environments. Many systems attempt to approximate the actions of an expert human tutor to simulate a one-on-one learning experience for the student, because this type of interaction has been shown to be highly fruitful for learning (Bloom, 1984). Such systems accomplish this feat in several ways (see VanLehn (2006) for a review of techniques related to Intelligent Tutoring Systems). For example, learning environments may provide implicit support in which the learning task is structured in such a way as to support students' learning (e.g. Quintana et al., 2004; Reiser, 2004). In contrast, they may provide explicit feedback based on the solutions students generate in the learning environment (e.g. the SQL Tutor, Mitrovic et al., 2003), or on the individual steps they take as they solve the problem (e.g. Cognitive Tutors, Anderson et al., 1995). Heffernan et al. (2006) designed a system, ASSISTments, that decomposes more complex problems into sub-problems students must solve

for mathematics. In terms of the types of support we aim to provide, our work most closely aligns with the Cognitive Tutor approach (Chapter 6), and as such, we describe this approach in more detail.

Cognitive Tutor-style learning environments, for example, provide hints and scaffolds by assessing students' problem solving steps (cf. Koedinger & Corbett, 2006). As students work, they can ask the system for help (Aleven et al., 2003) and receive targeted feedback relative to where they are in the problem solving process. In addition, such systems can also evaluate each step and provide immediate feedback on its own. Such feedback could be more simplistic, such as highlighting an incorrect portion of their problem solving process as in the Andes Physics system (vanLehn et al., 2005). The feedback may also provide targeted feedback or hints as in the Algebra Tutor (e.g. Anderson et al., 1995). Typically, the feedback or hints have multiple levels that increasingly provide more explicit, more targeted support to prevent students from struggling (cf. Vygotsky, 1978). Within our scaffolding we adopt a set of four levels seen in many Cognitive Tutor hints. In addition, the final hint is typically a "bottom-out hint" that tells the student exactly what to do on the current step (cf. Aleven & Koedinger, 2000).

In our work, we will leverage data-mined models, in part, to evaluate students' scientific inquiry skills. We use these models to augment the Inq-ITS inquiry learning environment with automated, real-time scaffolding of data collection skills. In the next chapter, we provide an overview of the learning environment and describe how students engage in inquiry within the system.

# 3 Inq-ITS Inquiry Environment

In this work, assessment, tracking and real-time scaffolding of students' data collection skills were studied and implemented within The Inq-ITS System (**Inq**uiry **I**ntelligent **T**utoring **S**ystem, www.inq-its.org), formerly known as Science Assistments (Gobert et al., 2012). This system, an extension of ASSISTments (Razzaq, et al., 2005), aims to enable automatic assessment of inquiry skills and provide support for middle school students as they engage in inquiry using interactive simulations across several domains such as Physical, Life, and Earth Science. Each Inq-ITS activity acts as performance assessment of inquiry skill; the actions students take within the simulation and work products they create are the bases for assessment (Gobert et al., 2012).

Inq-ITS is similar to the other microworld/simulation-based discovery environments discussed in Section 2.3 in that the computer-based activities structure students' exploration and share a goal in bootstrapping the acquisition of content knowledge. However, Inq-ITS' approach differs from those approaches in several ways. First, it emphasizes the *assessment of inquiry skills* identified by national and state frameworks (see Section 2.2). These skills include hypothesizing, designing and conducting experiments, interpreting data, and communicating findings (National Research Council, 1996, 2000, 2011). Second, it aims to *scaffold* these inquiry skills in real-time as students conduct their investigations. Thus, the system aims to provide students with supports so they do not flounder or engage in unproductive, haphazard inquiry behaviors (Gobert, Buckley & Horwitz, 2006; Gobert & Schunn, 2007). Third, it aims to *track the development* of inquiry skills across several science domains. Finally, the simulation environments in which students conduct inquiry target domain-specific concepts defined in the Massachusetts Curricular Frameworks content standards for middle school Science (Massachusetts Department of Education, 2006). This approach is commensurate with a

conceptualization of science inquiry described by Kuhn et al. (2000, p.497), "students investigate a set of phenomena- virtual or real – and draw conclusions about the phenomena."

The inquiry assessment activities all have a similar look-and-feel. They are designed to enable a moderate degree of student control, less than in purely exploratory learning environments (Amershi & Conati, 2009), but more than in classic model-tracing tutors (Koedinger & Corbett, 2006) or constraint-based tutors (Mitrovic, Mayo, Suraweera, & Martin, 2001; Mitrovic, 2003). More specifically, activities provide students with a driving question and require students to conduct an investigation using a simulation and inquiry support tools (van Joolingen, 1999) to address that question. These inquiry support tools include a hypothesizing widget, a data analysis widget, and graphs and tables for automatically displaying and summarizing data. The tools not only help students explore the simulation and keep track of their progress, but also enable easier assessment of inquiry since these tools make students' thinking explicit. Students' experimentation is also structured into different stages, providing both organizational support to students (Quintana et al., 2004; Guzdial, 1995), and assessment opportunities that elicit demonstration of skill (Gobert et al., 2012).

This dissertation aims to realize the goal of assessing, scaffolding and tracking the development of one subset of skills related to designing and conducting experiments (National Research Council, 1996, 2000, 2011). These two skills are designing controlled experiments and testing stated hypotheses. Towards this goal, we developed our techniques and conducted research within two simulations for physical science, Phase Change and Free Fall. In the following sections, we describe the flow of these activities and interface design in detail. This will give a sense of how the above design decisions for Inq-ITS was concretely implemented.

## 3.1 Version 1 of the Phase Change Inquiry Activities

The Phase Change activities aim to promote understanding about the melting and boiling properties of ice, a concept identified in the Massachusetts Curricular Frameworks (2008). Students learn about the topic by engaging in semi-structured scientific inquiry activities with a simulation. A typical activity is structured as follows. Students are first given an explicit goal to determine how one of four factors (container size, heat level, substance amount, and cover status) affects various measurable outcomes (melting point, boiling point, time to melt, and time to boil). For example, one activity asks students to "find out how the size of the container affects each of the dependent variables. Students then engage in inquiry several inquiry processes: formulating hypotheses, collecting and interpreting data, warranting claims, and communicating findings, in order to address the goal.

To support students' experimentation, the inquiry processes are structured into different stages: "observe", "hypothesize", "experiment", and "analyze data". Students begin in the "hypothesize" phase and use the hypothesis construction widget (Figure 1) to generate testable hypotheses. The widget is set of pulldown menus that act as a template for a hypothesis. For example, a student may state: "If I change the container size so that it decreases, the time to melt increases." If students are not yet ready to form a hypothesis, they can instead elect to first explore the simulation in the "observe" stage. After stating at least one hypothesis, they then move to the "experiment" stage to collect data for testing their hypotheses.

The "observe" stage and "experiment" stage (Figure 2) are similar. In the "experiment" stage, students are shown the Phase Change simulation and graphs that track changes of the substance's temperature over time. Students collect data (trials) by changing the simulation's

variable values, and running, pausing and resetting the simulation. Two inquiry support widgets are also provided to support planning of which trials to run next, a data table tool which shows the data collected by the student, and hypothesis widget which shows all of their stated hypotheses. These tools aim to help students plan which experiments to run next. The "observe" phase, in contrast, hides the inquiry support tools so that students can focus specifically on the simulation. As mentioned earlier, this gives students the opportunity to explore the simulation if they are not yet ready to formulate a hypothesis.

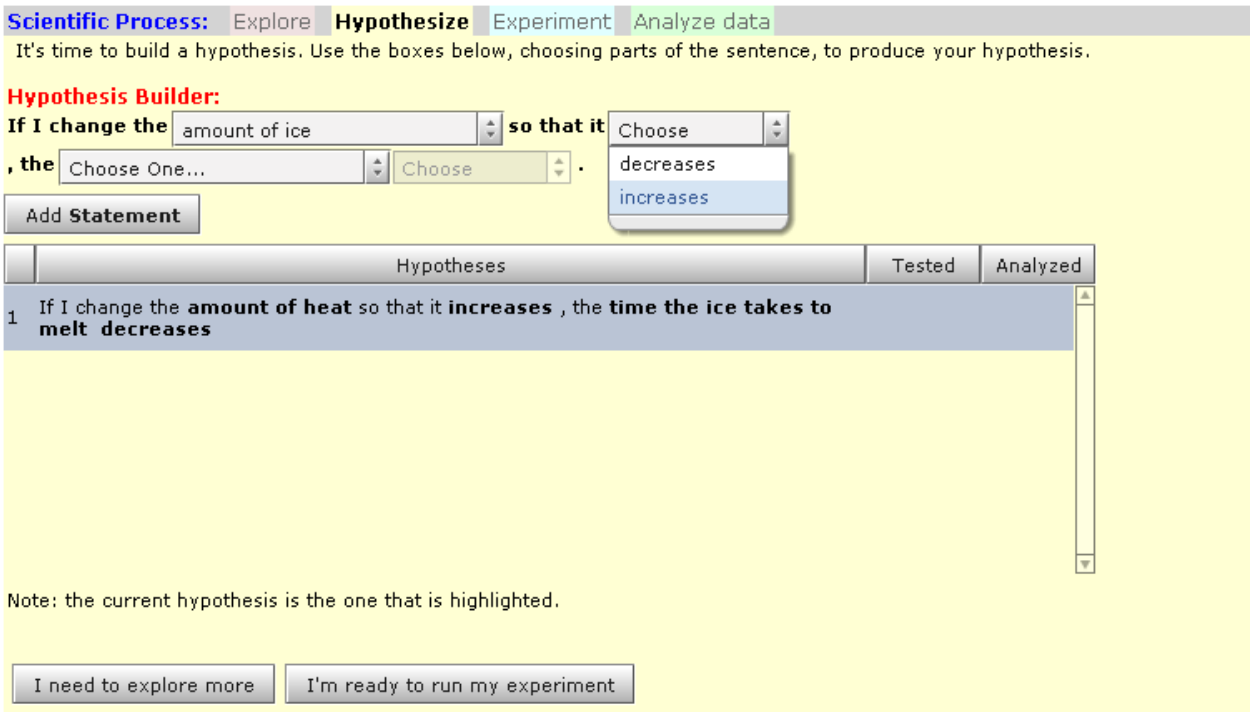


**Figure 1.** Hypothesizing tool for Version 1 of the Phase Change Inquiry Activities.

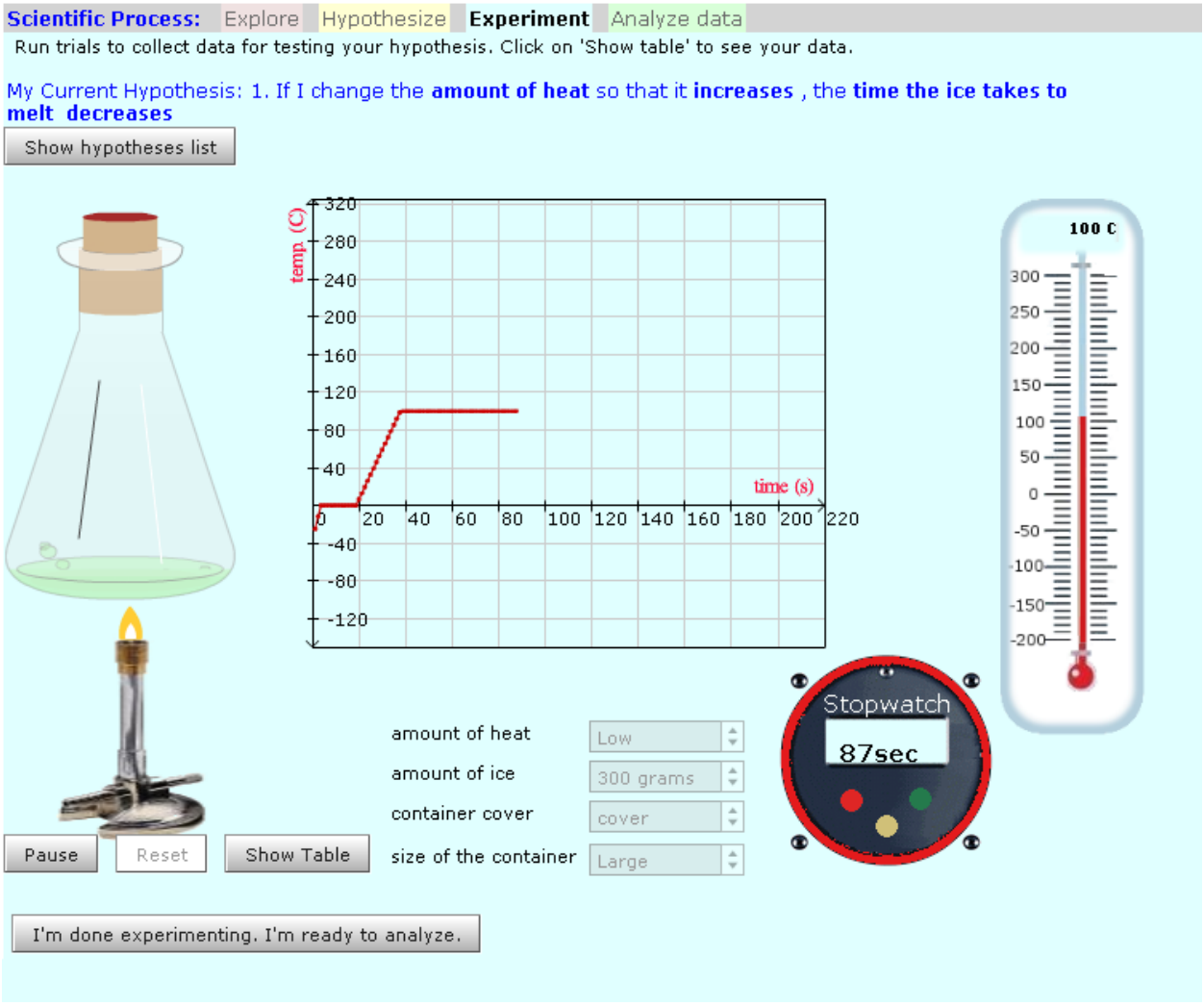


**Figure 2. "**Experiment stage" for Version 1 of the Phase Change activities.

Finally, in the "analyze data" stage (Figure 3), students are shown the data they collect and use the data analysis tool to construct an argument based on their data to support or refute their hypotheses. Here, similar to hypothesizing, students use pulldown menus to construct an argument whether their hypotheses were supported based on the data they collected. If students feel they need more data to conduct their analyses, they can return to the "experiment" stage to collect more data.

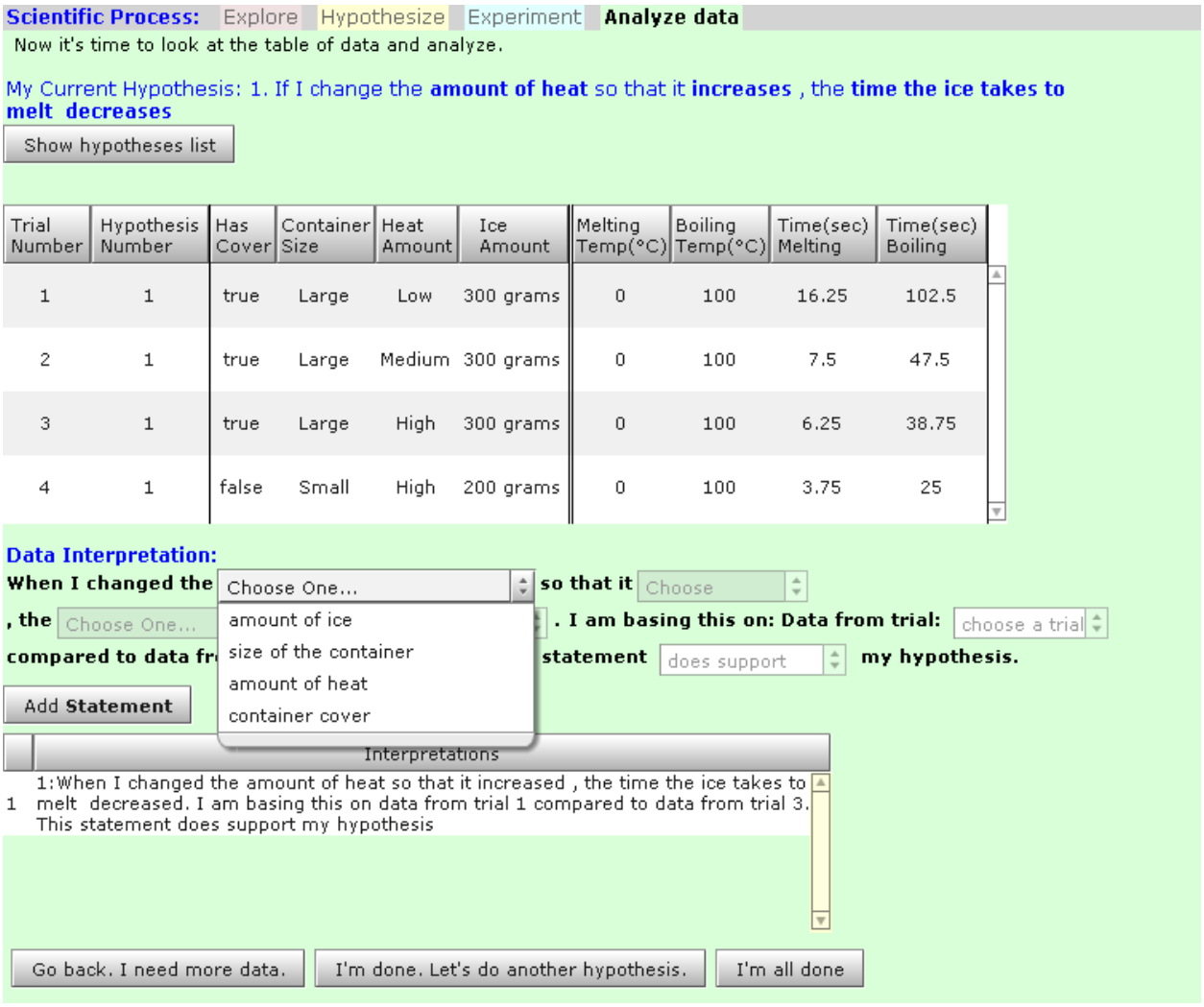


**Figure 3.** "Analyze Data stage" for Version 1 of the Phase Change activities.

In this first version of the Phase Change activities, students had some freedom to navigate between inquiry phases. The ways in which students can navigate between inquiry stages in shown in Figure 4. For example, while in the hypothesizing phase (Figure 1), students can elect to explore the simulation more before formulating any hypotheses by moving to the "observe" phase. Alternatively, they can choose to specify one or more hypotheses. Within the "experiment" phase (Figure 2), students can run as many experiments as desired to collect data for any one or all of their hypotheses. Within the "analysis" phase students also have several options. As they

construct their claims, students could decide to go back and collect more data or, after constructing claims based on their data, they could decide to create additional hypotheses, thus starting a new inquiry loop. In this version of the Phase Change activities, students were required to engage in at least two full inquiry cycles. In other words, after students completed one analysis, they were required to go back to the "hypothesize" phase to begin a new inquiry cycle for the same activity.

Of particular importance to this work is the "experiment" stage of inquiry, because this is where skill at designing controlled experiments and testing stated hypotheses is demonstrated. In this stage students could engage in either systematic or haphazard data collection behavior. Specific to the "hypothesize" and "experiment" phases, students acting in a systematic manner (Buckley, et al., 2010) collect data by designing and running controlled experiments that test their hypotheses. They also may use the table tool and hypothesis viewer in order to reflect and plan for additional experiments. These systematic behaviors are representative of the "designing and conducting experiments" skills (National Research Council, 1996) we aim to assess with machine-learned detectors. In contrast, students acting haphazardly in our environment may construct experiments that do not test their hypotheses, not collect enough data to support or refute their hypotheses, design confounded experiments, fail to use the inquiry support tools to analyze their results and plan additional trials (cf. de Jong, 2006), or collect data for the same experimental setup multiple times (Buckley, Gobert & Horwitz, 2006; Buckley, et al., 2010).

After testing this version of the phase change activities with students, we implemented changes based on feedback from students, teachers, and our own observations. These changes aimed to increase elicitation of students' inquiry skills and to improve the clarity of the inquiry task. We describe the changes in the next section.

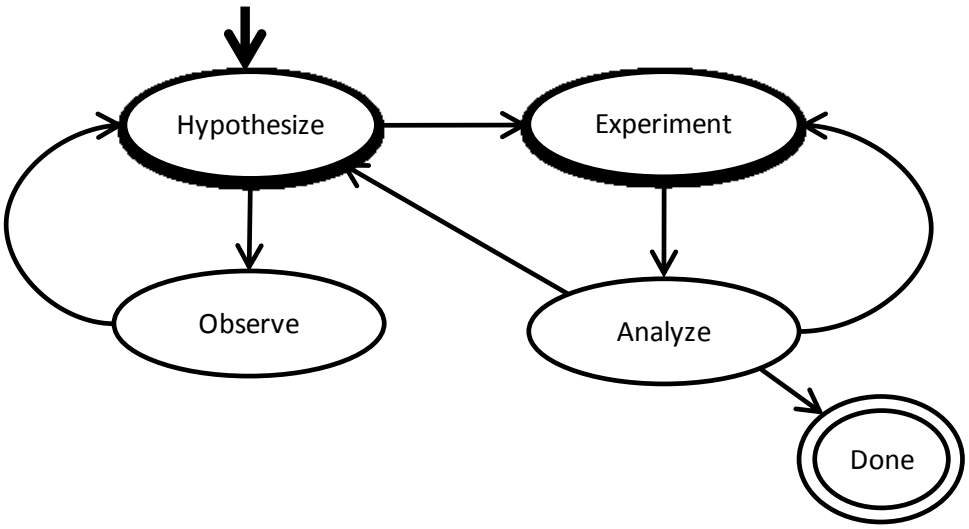


**Figure 4.** Phase Change v.1 paths through inquiry stages

## *3.2 Version 2 of the Phase Change Inquiry Activities*

In the second version of the Phase Change activities, we aimed to improve the usability and clarity of the activities. The overall structure of the activities remained the same in terms of keeping the stages of inquiry separate. However, the interface itself and each stage of inquiry (observing, hypothesizing, experimenting and analyzing data) were revised. We describe these changes in detail below.

In terms of high-level design, we aimed was to make the organization of the activity and the goals more salient to students. We did so by implementing several changes. First, as before, students begin in the "hypothesize" stage. However, before students hypothesize, a dialog box appears describing what the goal of the activity is, and how they should begin addressing the goal. In fact, such instructional dialog boxes appear the first time a student enters a new stage of inquiry. Second, as shown in Figure 5, the stages of inquiry were enlarged and highlighted so students could more clearly identify in which stage of inquiry they were working. Finally, unlike

the prior version, the overarching goal of the activity was always present. For example, in Figure 5, the goal of the activity is to "determine how one variable you choose affects the boiling point of ice". Underneath that goal, a subgoal appears to describe what the student should do in this stage of inquiry. This subgoal description is a summary of the instructions in the initial dialog box. Students can click on "more" or the question mark to re-open the dialog box with the more explicit instructions. The aim of these changes was to provide more orientation to the students as they engaged in inquiry (Quintana et al., 2006).

Changes to each stage of inquiry were also made to make students' thinking more explicit, and to provide more support for students' inquiry (Quintana et al., 2006). For example, in the original version of the Phase Change activities, students could add as many hypotheses as they liked before experimenting. In this new version, students could only specify a single hypothesis (see Figure 5). Anecdotally, we observed students were confused by the original interface and process of adding multiple hypotheses. As such, we simplified this interface. We do note that even though a goal is given to students, as in the prior version they may elect to not follow the goal and specify any hypothesis they wish. In addition, the pulldown menus now mixed together independent and dependent variables, making it possible for students to state an untestable hypothesis. Thus, the hypothesize phase now acts an assessment of hypothesizing skill (Gobert et al., 2012).

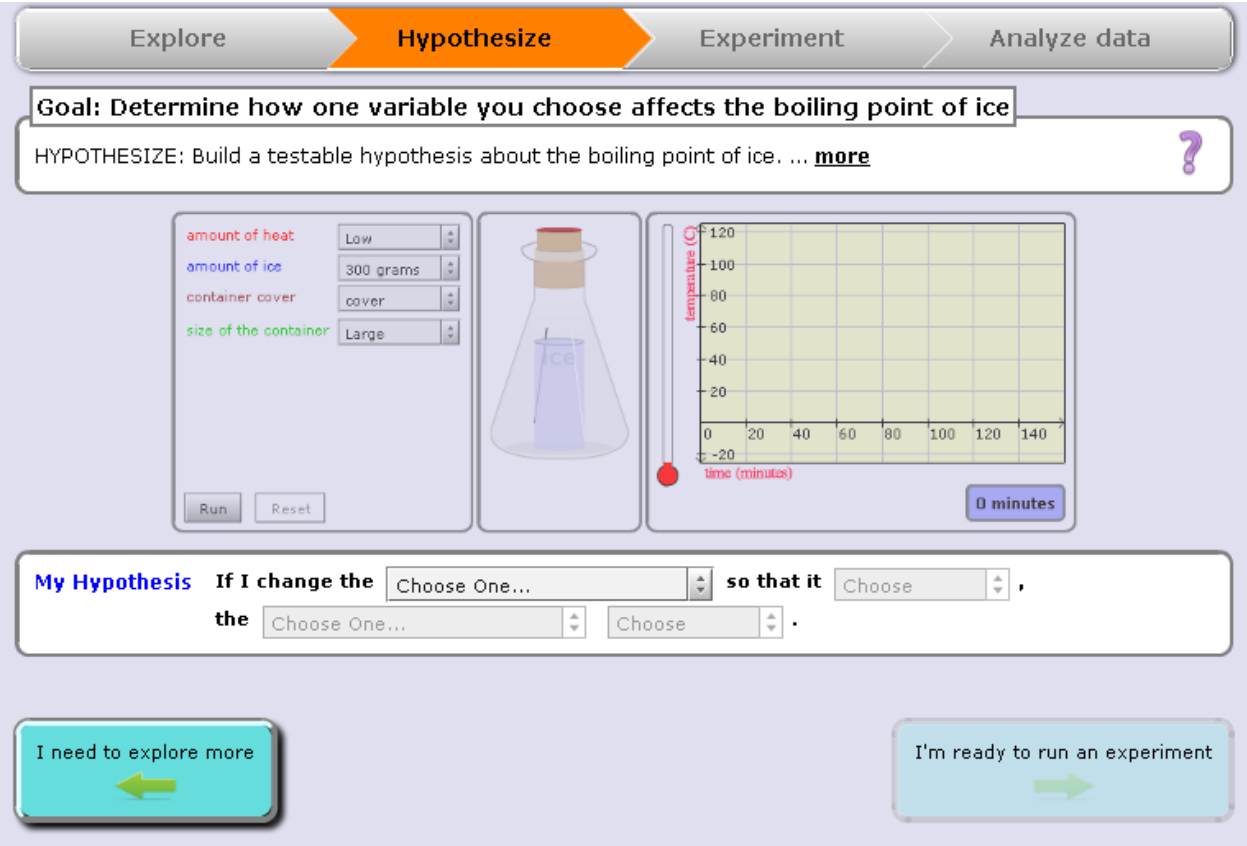


**Figure 5.** Version 2 of the Phase Change Activities - Hypothesize stage of inquiry.

The "experiment" stage of inquiry was also revised. As in the prior version, students still design and run trials for collecting data. However, as shown in Figure 6, the layout of the components has been modified to improve organization and visibility of the graphs and thermometer. In addition, the data table is now permanently visible, so students are always able to see the data they collected. As mentioned earlier, this stage of inquiry is the primary focus of the dissertation, because the skills of interest, designing controlled experiments and testing stated hypotheses, are demonstrated here.

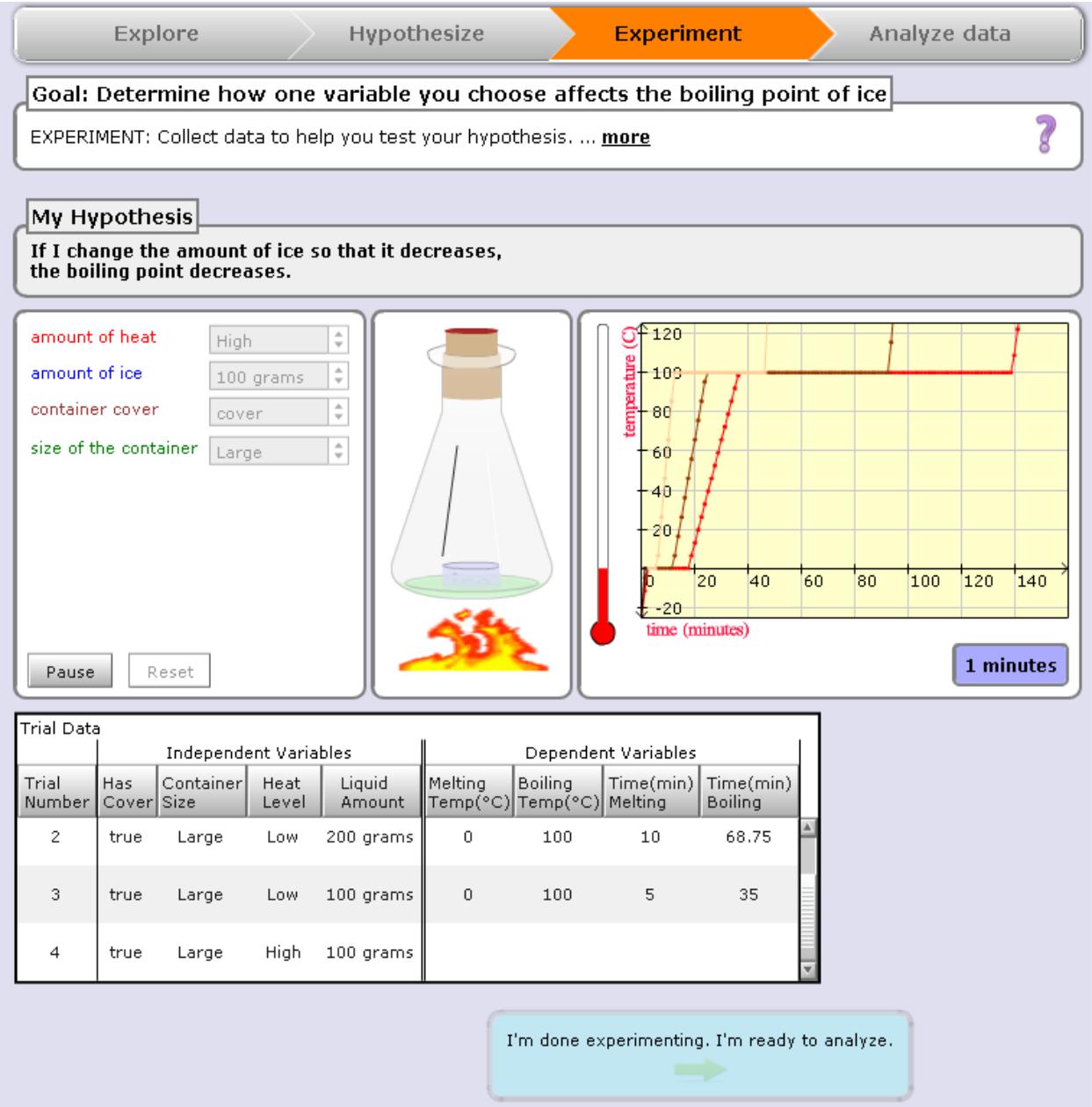


**Figure 6.** Version 2 of the Phase Change activities - Experiment Stage of Inquiry

Modifications to the "analyze data" stage were also made. Most notably, students now dragged and dropped trials they collected into an "evidence" table to demonstrate that they knew which trials warranted the claim they specified in the drop down boxes (see Figure 7). In addition, they only specified a single analysis whereas in the prior version they could specify as

many analyses as they desired. This was done to improve clarity of the task. As in the prior version, if students felt they needed to collect more data to form their analysis, they could return to the "experiment" stage. Finally, unlike the prior version, once students finished their analysis the task, the activity ended thus completing one full inquiry cycle. In the original version, students were made to complete at least two full inquiry cycles. This change was made for two reasons. First, we again aimed to improve clarity of the task. Second, we found that many students would simply skip through the second inquiry loop without testing any additional hypotheses.

All of the physical science activities were updated to have a similar look-and-feel to this new version of the Phase Change activities. As part of this dissertation, we examine demonstration of inquiry skill within a second set of physical science activities pertaining to Free Fall. We discuss these activities next.

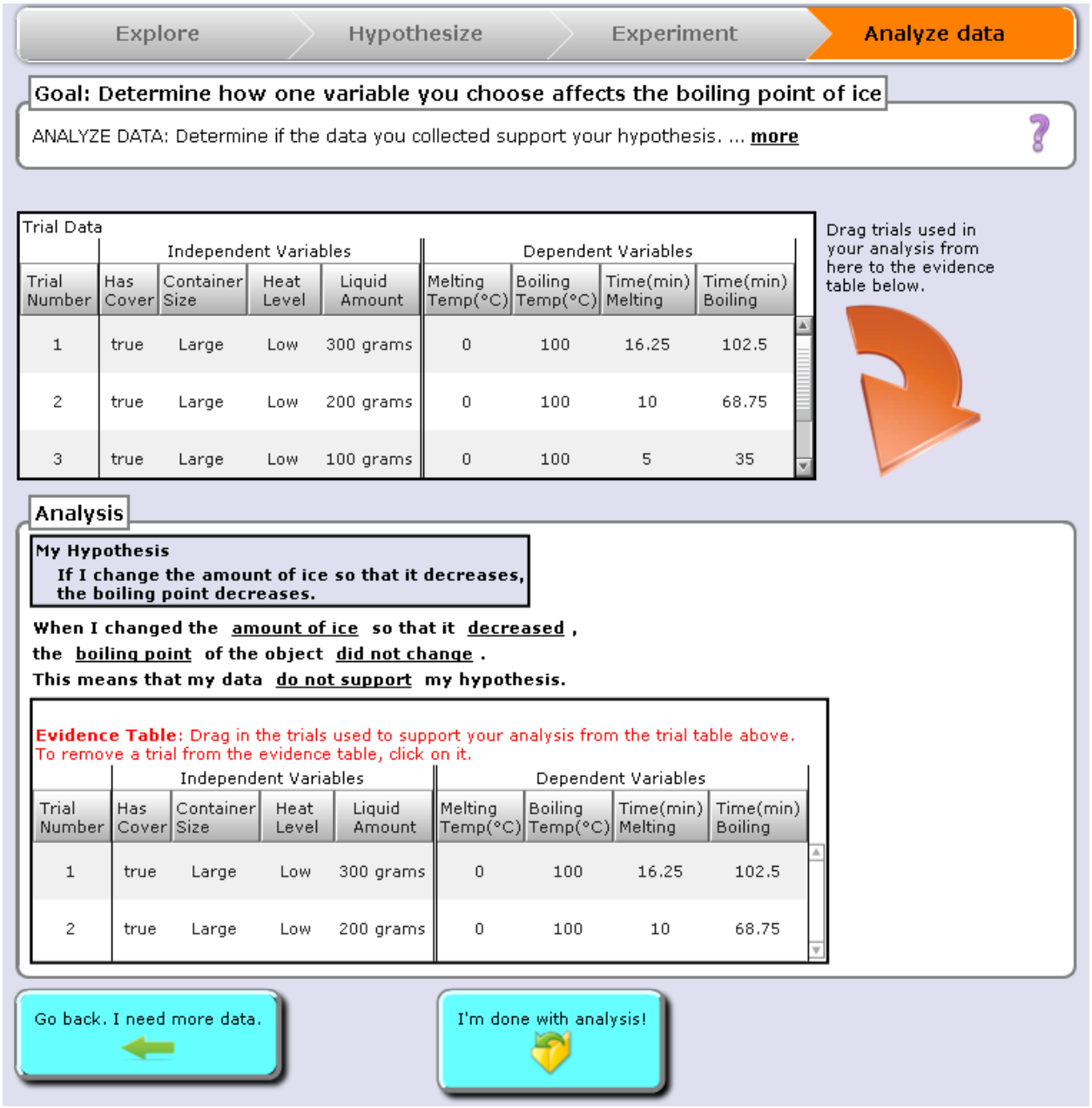


**Figure 7.** Version 2 of the Phase Change Activities – Analyze Data stage of inquiry

## *3.3 Free Fall Inquiry Activities*

The Inq-ITS Free Fall activities aim to foster understanding about factors that influence the kinetic, potential and mechanical energy of a ball when it is dropped. The two factors students could change were the ball's mass and starting height. The look-and-feel of these activities is structurally the same as the second version of Phase Change. In addition, the same interaction constraints were used (see Section 3.2 for more details on these constraints). However, there were some notable differences between the activities. First, the number of factors the student could manipulate was smaller. There were 2 independent/manipulable variables in Free Fall versus 4 in Phase Change. Second, the number of dependent/observable variables the student could observe was larger, 6 in Free Fall versus 4 in Phase Change. Finally, students were shown three graphs in the "experiment" phase (see Figure 8) that tracked how the observable variables changed over time when the student ran the simulation.

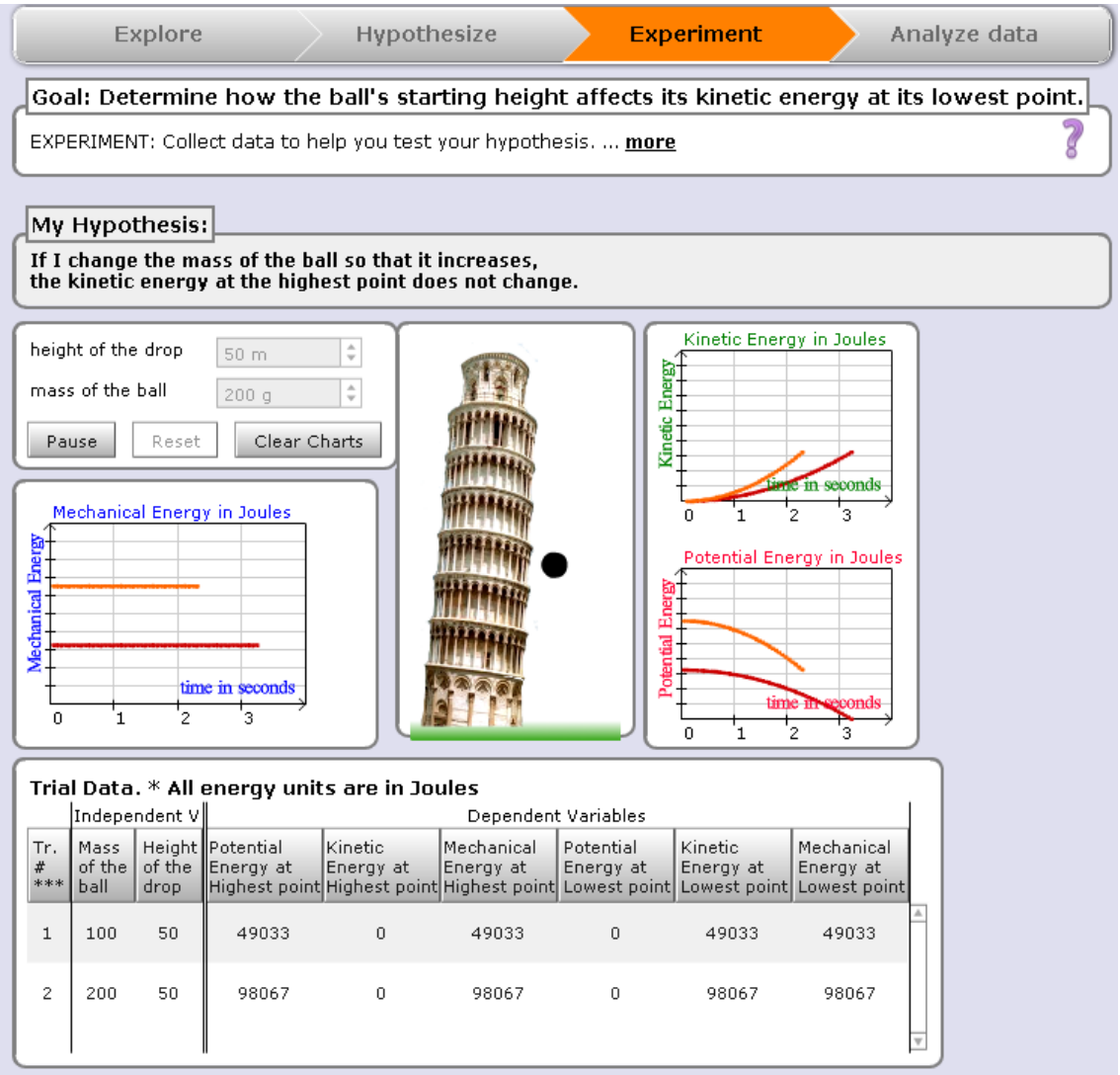


**Figure 8.** Free Fall simulation in the experiment stage of inquiry

## *3.4 Next Steps*

Inq-ITS aims to assess, scaffold and track inquiry skills. This occurs within simulation-based activities like the examples presented for Phase Change and Free Fall. Students' low-level interactions with the simulations and inquiry support tools form the basis of this assessment, scaffolding, and tracking. In initial versions of the physical science activities, the assessment of inquiry was just starting to take shape. The models for data collection skills consisted of

simplistic knowledge-engineered rules that were not validated. There was also no strong, validated methodology on how to aggregate students' practice attempts in order to track inquiry skills over time. Finally, feedback given to students was preliminary and untested in terms of its efficacy.

In the following chapters, we describe the results for improving the assessment and proficiency estimation of the two data collection skills, designing controlled experiments and testing stated hypotheses, within physical science activites for Phase Change and Free Fall. In particular, we describe our results towards developing data-mined detectors of the inquiry skills that were validated against unseen student data. These detectors assess whether or not students demonstrate inquiry skills during an inquiry activity, and trigger scaffolding when students do not demonstrate skill. In addition, we present our work towards validated skill estimation models that aggregate students' practice attempts over the activities, and predict whether students have mastered the two inquiry skills. Finally, we show how our scaffolding approach improves learning and transfer of inquiry skill across domains.

# 4 Development of Data-Mined Detectors for Assessing Data Collection Skills

Real-time scaffolding and estimation of the two data collection skills of interest, designing controlled experiments and testing stated hypotheses, cannot be realized without a means of automatically evaluating students' data collection within simulations. As discussed earlier, prior work used knowledge engineering (cf. Koedinger & MacLaren, 2002) approaches to construct models of inquiry behavior (Koedinger, Suthers & Forbus, 1998; Buckley et al., 2006, 2010; Levy & Wilensky, 2006; McElhaney & Linn, 2008, 2010; Schunn & Anderson, 1998). We instead, developed validated, data-mined detectors (models) for this same purpose.

For the two data collection skills of interest, the knowledge-engineered approaches used previously (Koedinger et al., 1998; Harrison & Schunn, 2004; McElhaney & Linn, 2008, 2010; Gobert & Koedinger, 2011) may fail in cases where students may exhibit a variety of data collection behaviors and 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.

In our approach, student interactions with the learning software are used as a basis for creating models, like the past knowledge engineering approaches. As with Schunn and Anderson (1998), we also aim to evaluate and quantify students' skills, and determine how well the detectors predict the inquiry skills of interest. This approach is different, however, in that rules are not prescribed a-priori. Instead, given student data, human-classified labels, and a feature set derived from student data, data mining techniques are leveraged to *build* models (rules) of skill demonstration.

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. The models can also represent boundary conditions – and the fuzziness at the edges of boundary conditions – more appropriately than knowledge engineering approaches. Finally, the accuracy and generalizability of machine learning approaches are easier to verify than for knowledge engineering, since machine learning is amenable to standard methods for predicting how well models will generalize to new data like cross-validation (cf. Efron & Gong, 1983) and using held-out test sets (cf. Witten & Frank, 2005). Thus, this approach facilitates concrete determination of model goodness.

In this chapter, we discuss our work towards building, refining, and validating data-mined detectors (models) that discern whether students design controlled experiments and test their stated hypotheses within the Phase Change and Free Fall physical science activities. In our approach, we leveraged "text replay tagging" of log files (Baker, Corbett & Wagner, 2006; Sao Pedro et al., 2010a; Montalvo et al., 2010). This approach enables human coders to classify clips,

textual sequences of low-level student actions gleaned from log files, within the phase change environment. Text replay tagging is an extension of text replay approach originally developed in Baker et al. (2006), a method that has been shown to achieve good inter-rater reliability, and has led to models that can successfully predict a wide array of student behaviors and skills (Baker & de Carvalho, 2008; Baker et al., 2010a; Wixon et al., 2012). In text replays and text replay tagging, human coders are presented "pretty-printed" versions of log files with important data features emphasized to simplify the coding process. The two differ in that text replays only permit labeling a clip with a single category, whereas text replay tagging allows multiple tags to be associated with one clip. For example, within our domain, a clip may be tagged as involving designing controlled experiments, involving testing the correct hypothesis testing, both, or neither.

This chapter is segmented as follows. First, we describe how we translated raw student interactions into data suitable to develop data-mined models (Sao Pedro et al., 2010a; 2013b). Then, we describe our first attempt at constructing detectors from those data and present results on their validity (Sao Pedro et al., 2010a; 2013b). From there, we show how we took our "lessons learned" from the initial approach and applied new techniques that led to better predicting models, particularly by considering construct validity (Sao Pedro et al., 2012a). As part of the model redesign, we also show how these models with improved construct validity can also predict with less data, providing credence for using them to trigger scaffolding (Sao Pedro et al., 2012a). Finally, we present results on how well these new models performed at predicting skill within the Free Fall activities, testing their ability to generalize to a new topic and set of students. (Sao Pedro et al., 2013a).

## 4.1 Text Replay Tagging Methodology: Development of the Initial Version of the Data Collection Skill Detectors

As shown in Figure 9, there are several steps in generating and validating detectors of these skills using text replay tagging. A data mining approach necessitates the use of training/testing data from which models can be derived and validated. As such, the first step in the text replay tagging process is to collect student data. These data are students' low-level interactions within the interface (discussed in more detail below). Next, students' actions are segmented into *clips*. A clip contains all the information necessary for a human coder or a model to determine whether the actions indicate skillful demonstration of designing controlled experiments and/or testing stated hypotheses.

In order to train and test a data-mined model that evaluates students' data collection, two additional pieces are needed: labels that state whether a clip is indicative of skill demonstration, and a feature set to summarize clips. Human coders apply tags (labels) to *text replays*, "pretty-prints" of clips that are the "ground truth" of whether or not skill was demonstrated. The feature set is a set of attributes (predictor variables) that characterize clips in some way. Some example features are the number of trials collected by students and the number of hypotheses they made; the features considered in building models are discussed in more detail in Section 4.1.4).

With these two pieces of information in place, clip labels and summary features, a machine learning algorithm then learns which combination of features and feature values best predict demonstration skill, given the data. The models resulting from the algorithm then takes the place of the human coder and can identify whether or not the clip (a segment of students' actions) indicate demonstration of skill at designing controlled experiments and/or testing stated hypotheses. The final step is to validate the models, meaning that we see how well their

predictions match the human coders' ground truth labels on clips not used in the model construction process. If the models and human coders agree highly, then we have some assurance they truly can assess students' data collection.

In next sections, we describe how we went from unprocessed logs to the construction and validation of the initial version of our skill detectors. These detectors were built in the context of the first version of the Phase Change activities (Section 3.1).

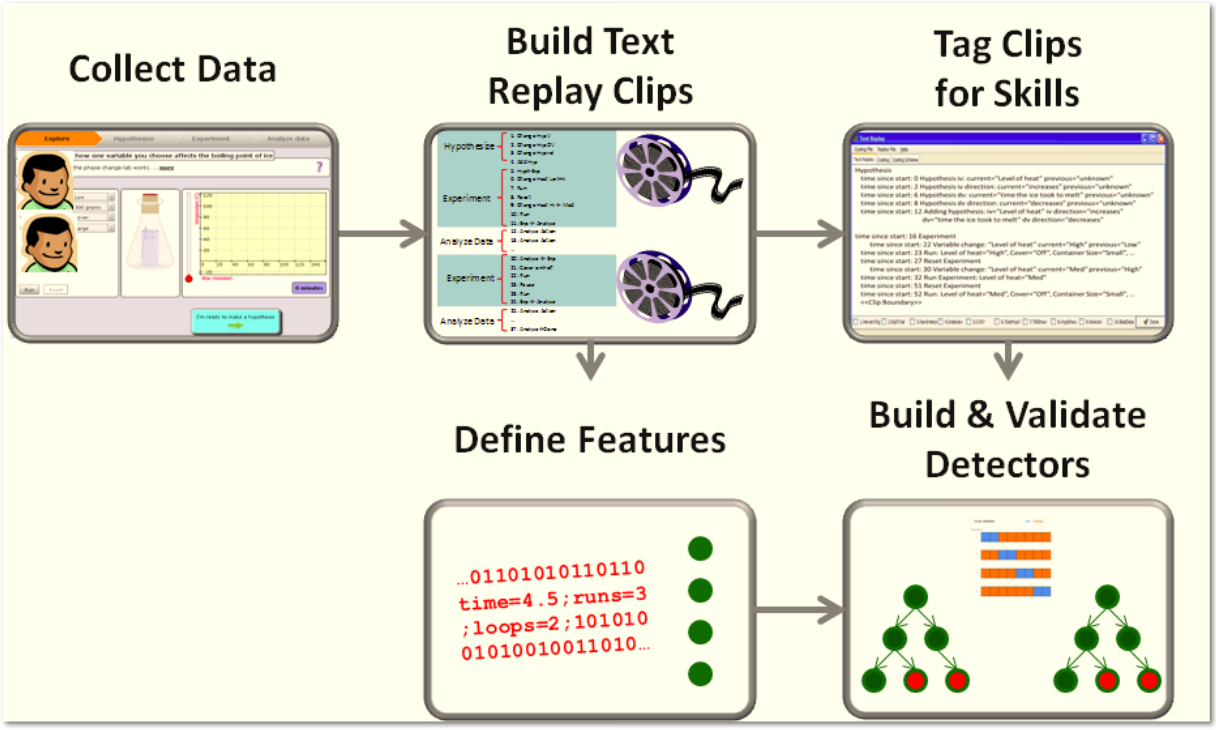


**Figure 9.** Overview of the text replay tagging process that enabled the construction of validated, machine-learned skill detectors. These detectors are used to assess whether students design controlled experiments or test their stated hypotheses during their experimentation.

### 4.1.1 Data Collection and Log File Generation

As mentioned, the first step in the process is to collect students' interaction data as they conduct inquiry within Inq-ITS activities. We collected interaction data from 148 eighth grade students ranging in age from 12-14 years from a public middle school in suburban Central Massachusetts.

Students belonged to one of six class sections and had one of two science teachers. They had no previous experience using simulations within Inq-ITS.

Students engaged in four inquiry activities within the first version of the phase change (see Section 3.1 for more information about these activities). Each inquiry activity had the goal of determining how each of the four factors related to phase change (size of container, amount of substance, container cover, and amount of heat) affected the outcomes (melting point, boiling point, time to melt, time to boil). As students conducted their experiments, they did not receive any feedback on their experimentation processes.

As students engaged in the activities, all of their fine-grained interactions within the learning environment were recorded, resulting in log files of their work. These interactions are the basis of assessing whether or not students designed controlled experiments and tested their stated hypotheses in the activities. An example of an unprocessed log file for a student is shown in Table 1. In developing our models, we looked specifically at student actions from the "hypothesize" and "experiment" phases of inquiry because this is where data collection skills are demonstrated. Logged actions included low-level widget interactions from creating hypotheses, designing experiments, showing or hiding support tools (the data table or hypothesis list), running experiments, and transitioning between inquiry activities (i.e. moving from hypothesizing to experimenting). Looking more deeply, the following data were recorded for each action:

- *Action:* A unique action ID
- *Time:* The action's timestamp, in milliseconds
- *Activity:* The unique ID of the activity in which the action took place
- *Student:* The ID of the student working on the activity

- *Widget:* A unique name associated with a graphical widget / system component associated with the activity
- *Who:* The entity who initiated the action, the student or the system
- *Variable:* The unique aspect of the inquiry problem that the widget / system component changes. Examples include individual components of the hypothesis and values that can be changed for the Phase Change simulation.
- *Value:* current value for the variable, if applicable.
- *Old Value*: previous value for the variable, if applicable.
- *Step Name*: A unique marker describing the action taken by the user. This is akin to a problem solving step in Cognitive Tutors (cf. Corbett & Anderson, 1995; Koedinger & Corbett, 2006). In particular, this information helped simplify and standardize the development of our clip generation and feature distillation software.

To give a concrete example, action 62955 in Table 1 indicates that the student changed the value of the "Level of heat" variable from "Low" to "Medium". In all, 27,257 unique student actions for the four Phase Change activities were captured. These actions served as the basis for generating clips containing actions related to data collection, discussed next.

**Table 1.** Unprocessed log file segment for a student engaging in inquiry within a single activity

| Action | Time | Activity | Student | Widget | Who | Variable | Value | Old Value | Step Name |
|---|---|---|---|---|---|---|---|---|---|
| 62934 | ...5669 | 147212 | 85240 | variable_containerSize | system | Container Size | Large | null | INIT_SET_IV |
| 62935 | ...5669 | 147212 | 85240 | variable_coverStatus | system | Cover Status | Cover | null | INIT_SET_IV |
| 62936 | ...5669 | 147212 | 85240 | variable_substanceAmount | system | Amount of Substance | 300 grams | null | INIT_SET_IV |
| 62937 | ...5669 | 147212 | 85240 | variable_heatLevel | system | Level of heat | Low | null | INIT_SET_IV |
| 62938 | ...5684 | 147212 | 85240 | hypothesis.iv | student | iv | Level of heat | null | SPECIFY_IV_HYPOTHESIS |
| 62939 | ...5691 | 147212 | 85240 | hypothesis.iv.dir | student | iv.heatLevel.direction | increases | null | SPECIFY_IV_DIRECTION_HYPOTHESIS |
| 62940 | ...5704 | 147212 | 85240 | hypothesis.iv.dir | student | iv.heatLevel.direction | decreases | increases | SPECIFY_IV_DIRECTION_HYPOTHESIS |
| 62941 | ...5707 | 147212 | 85240 | hypothesis.dv | student | heatLevel.dv | time to melt | null | SPECIFY_DV_HYPOTHESIS |
| 62942 | ...5722 | 147212 | 85240 | hypothesis.iv.dir | student | iv.heatLevel.direction | increases | decreases | SPECIFY_IV_DIRECTION_HYPOTHESIS |
| 62943 | ...5731 | 147212 | 85240 | hypothesis.dv.dir | student | heatLevel.dv.timeMelting.direction | decreases | null | SPECIFY_DV_DIRECTION_HYPOTHESIS |
| 62944 | ...5737 | 147212 | 85240 | hypothesis.add | student | | | | ADD_HYPOTHESIS |
| 62945 | ...5740 | 147212 | 85240 | stage:hypothesize->experiment | student | | | | CHANGE_STAGE_HYPOTHESIZE_EXPERIMENT |
| 62946 | ...5757 | 147212 | 85240 | variable_heatLevel | student | Level of heat | High | Low | CHANGE_IV |
| 62947 | ...5760 | 147212 | 85240 | variable_heatLevel | student | Level of heat | Low | High | CHANGE_IV |
| 62948 | ...5763 | 147212 | 85240 | run | student | | | | RUN |
| 62949 | ...5763 | 147212 | 85240 | PhaseTable.cvs.column | student | | | | SELECT_TRIALS |
| 62950 | ...5764 | 147212 | 85240 | simulation | system | state | run | | |
| 62951 | ...5777 | 147212 | 85240 | simulation | system | state | complete | | SIM_COMPLETE |
| 62952 | ...5781 | 147212 | 85240 | reset | student | | | | REVERT_IVS |
| 62953 | ...5781 | 147212 | 85240 | simulation | system | state | reset | | |
| 62954 | ...5781 | 147212 | 85240 | simulation | system | state | ready | | |
| 62955 | ...5784 | 147212 | 85240 | variable_heatLevel | student | Level of heat | Medium | Low | CHANGE_IV |
| 62956 | ...5786 | 147212 | 85240 | run | student | | | | RUN |
| 62957 | ...5786 | 147212 | 85240 | PhaseTable.cvs.column | student | | | | SELECT_TRIALS |
| 62958 | ...5787 | 147212 | 85240 | simulation | system | state | run | | |
| 62959 | ...5793 | 147212 | 85240 | simulation | system | state | complete | | SIM_COMPLETE |
| 62960 | ...5797 | 147212 | 85240 | reset | student | | | | REVERT_IVS |
| 62961 | ...5797 | 147212 | 85240 | AppManager | system | state | reset | | |
| 62962 | ...5797 | 147212 | 85240 | AppManager | system | state | ready | | |
| 62963 | ...5807 | 147212 | 85240 | variable_heatLevel | student | Level of heat | High | Medium | CHANGE_IV |
| 62964 | ...5809 | 147212 | 85240 | run | student | | | | RUN |
| 62965 | ...5809 | 147212 | 85240 | simulation | system | state | run | | |
| 62966 | ...5814 | 147212 | 85240 | simulation | system | state | complete | | SIM_COMPLETE |
| 62967 | ...5818 | 147212 | 85240 | showDataTable | student | | | | DATA_TABLE_DISPLAY |
| 62968 | ...5854 | 147212 | 85240 | showHypotheses | student | | | | HYPOTHESES_LIST_DISPLAY |
| 62969 | ...5880 | 147212 | 85240 | showHypotheses | student | | | | HYPOTHESES_LIST_DISPLAY |
| 62970 | ...5884 | 147212 | 85240 | stage:experiment->analyze | student | | | | CHANGE_STAGE_EXPERIMENT_ANALYZE |
| | | | | | … Actions for Analysis Phase of Inquiry… | | | | |
| 62982 | ...5979 | 147212 | 85240 | stage:analyze->hypothesize | student | | | | CHANGE_STAGE_ANALYZE_HYPOTHESIZE |
| 62983 | ...5979 | 147212 | 85240 | HypTable.analyzed.column | student | row:1 | true | null | |
| 62984 | ...5996 | 147212 | 85240 | stage:hypothesize->experiment | student | | | | CHANGE_STAGE_HYPOTHESIZE_EXPERIMENT |
| 62985 | ...6004 | 147212 | 85240 | stage:experiment->analyze | student | | | | CHANGE_STAGE_EXPERIMENT_ANALYZE |
| | | | | | … Actions for Analysis Phase of Inquiry… | | | | |
| 62988 | ...6007 | 147212 | 85240 | submit | student | | | | |

### 4.1.2 Constructing Clips from Log Files

Clips contain all the actions necessary to determine demonstration of data collection skill, either by a human or an automated detector. They are the granularity at which students will be assessed. Prior work delineated clips based on a pre-specified length of time, e.g. 20 second intervals (Baker & de Carvalho, 2008; Baker et al., 2010a). However, for our work, we found that we needed to include significant periods of experimentation to enable determination of whether a set of actions reflected demonstration of designing controlled experiments and/or testing stated hypotheses. As such, clips were not delineated by time, but instead by inquiry stages (Figure 4). Clips included all actions from the "hypothesize" stage, to provide information about which hypotheses were stated, and all actions from the "experiment" stage to provide information about which trials the student design and ran.

Earlier, we mentioned that in the Phase Change activities students could engage in several inquiry loops within an activity (see Figure 4). This means that several clips could be generated for a single activity, and several evaluations of students' data collection could occur in an activity. More specifically, clips could begin at different points, depending on how students navigated through inquiry phases. First, a clip could begin at the start of a full inquiry loop when a student enters the "hypothesize" phase. This phase could be entered in two ways, either by starting the activity, or by choosing to create more hypotheses in the "analyze" phase, thereby starting a new inquiry loop. A clip could also begin in the middle of a full inquiry loop if a student chose to go back to the "experiment" phase to collect more data while in the "analyze" phase. A clip always ended when the "experiment" phase was exited.

As an example, consider the action sequence for the student's activity shown in Table 1. This activity yielded two clips, one clip containing actions 62934 through 62970, and another containing actions 62982 through 62985, when the student navigated back to the "hypothesize" phase from the "analyze" phase. As such, there are two opportunities to assess this student's data collection in this activity.

We applied this clip generation procedure to the student interaction data collected for the Phase Change activities (Section 4.1.1). This yielded 1,503 clips from the database of all student actions. To generate data-mined models that evaluate students' data collection, we needed to generate labels which act as "ground truth" of whether or not skill was demonstrated in a clip. We describe this process next.

### 4.1.3 Tagging Text Replays of Clips with Demonstration of Skill

Part of the text replay tagging process requires that clips be labeled with behaviors or demonstrations of skill. As such, there are two steps which must be performed before tagging clips: defining the possible tags (labels) to assign to clips, and defining what the text replays (pretty-prints of clips) should look like. We discuss the tags first.

Originally, nine tags (or classification labels) were identified (Sao Pedro et al., 2010a, 2013b), corresponding to systematic and haphazard data collection behaviors of interest. In line with the text replay tagging approach, any or all of these tags could be used to classify a clip. These tags were: "Designed Controlled Experiments", "Tested Stated Hypothesis", "Used Data Table to Plan", "Used Hypothesis List to Plan", "Never Changed Variables", "Repeat Trials", "Non-Interpretable Action Sequence", "Indecisiveness", and "No Activity". These were chosen based on systematic and haphazard behaviors identified in previous work on inquiry-based learning (cf. Buckley et al., 2006, 2010; de Jong, 2006). We also added one extra category for unclassifiable clips, "Bad Data", for a total of 10 tags.

The text replays were designed to highlight key aspects of students' hypothesizing and data collection in a clip. For example, students' full experimental designs are displayed when they run a trial, and fully stated hypotheses are shown (Figure 10). The idea is to make it easier for a human coder to identify systematic and/or haphazard inquiry behavior, rather than look at raw log files (e.g. Cocea & Weibelzahl, 2009). When designing text replays, we noticed that it may be difficult for coders to properly label skill for latter clips in an activity (recall that students' interactions could result in several

clips per activity). For example, it would be difficult to accurately tag a clip in which a student transitioned from "analyze" back to "experiment" without seeing students' associated actions from the previous "hypothesize" phase. As another example, data collected in a previous inquiry cycle were sometimes utilized by students to test a hypothesis in a later inquiry cycle; not seeing these previous could lead to incorrect tagging. Without showing the full history for coding, it would not be possible for coders to recognize the student's sophisticated inquiry behavior. To compensate, our text replays contained clips representing the actions for testing the current hypothesis, and cumulative data including actions performed when testing previous hypotheses or collecting previous data. In other words, we coded clips while taking into account actions from earlier clips from the same activity.

To support text replay tagging, we developed a new tool implemented in Ruby (Figure 10, Sao Pedro et al., 2010a; Montalvo et al., 2010). This tool enabled the classification of clips using any combination of the tags defined in Section 4.3. The tool displays one text replay at a time, consisting of the current clip and all relevant predecessor clips. Within our approach, a human coder chooses at least one but possibly several tags to classify the clip. As previously mentioned, our analyses focused on two skills associated with data collection skill, "Designed Controlled Experiments" and "Tested Stated Hypothesis". We tagged a clip as "Designed Controlled Experiments" if the clip contained actions indicative of students trying to isolate the effects of one variable. "Tested Stated Hypothesis" was chosen if the clip had actions indicating attempts to test one or more of the hypotheses stated by the student, regardless of whether or not the experiments were controlled. Thus, a clip could be tagged as demonstrating either skill, both skills, or no skills.

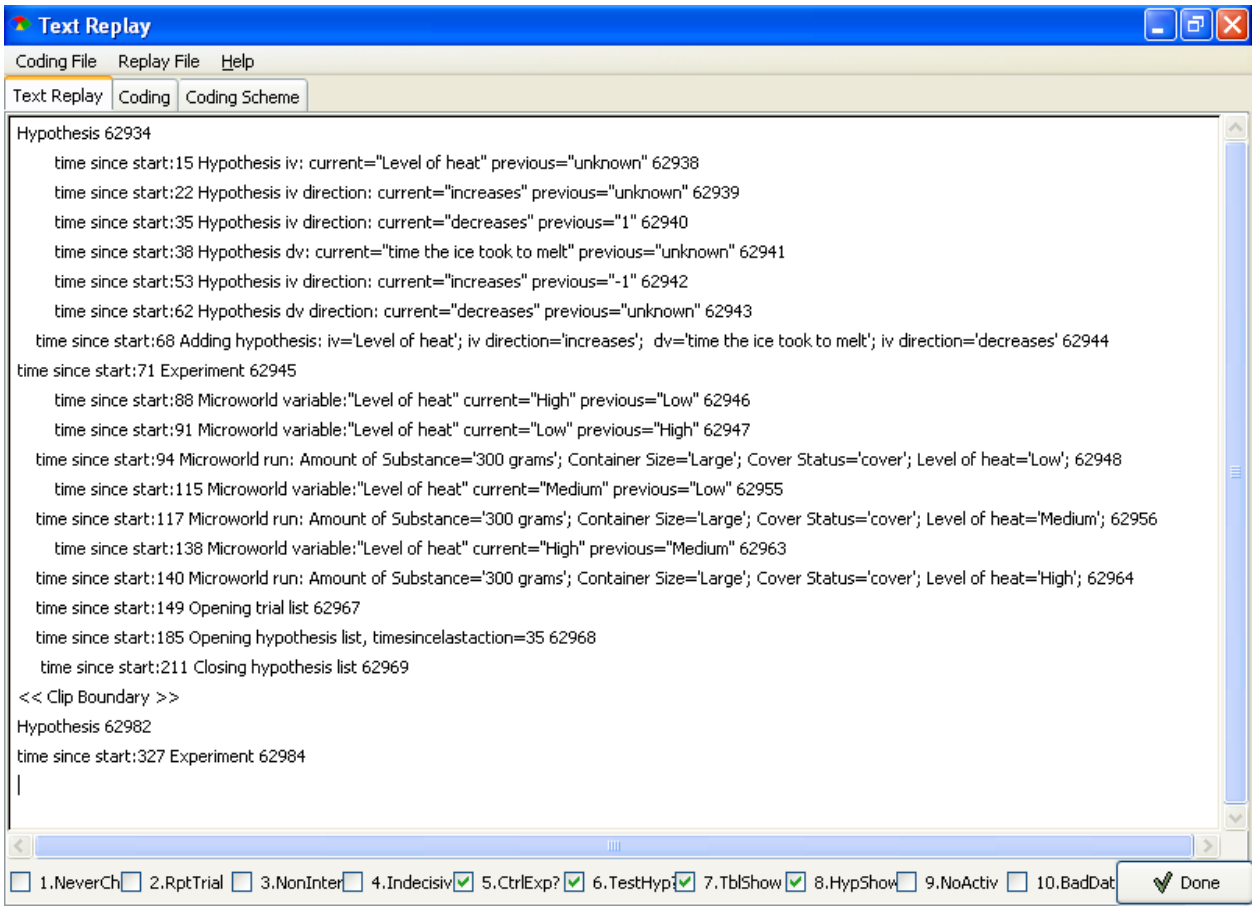


**Figure 10.** Text Replay Tagging Tool with an example text replay corresponding to the action sequence displayed in Table 1. This clip, the second clip generated for the activity, was tagged as involving designing controlled experiments, testing stated hypotheses, and using the data table and hypothesis list to plan which experiments to run next.

To give a better idea of how skill demonstration 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" behavior. To tag this behavior, the human coder focuses primarily on the trials run by the student. At a high level, the student engaged in two experimentation cycles, as indicated by the "<<Clip Boundary>>". This means that the text replay in Figure 6 corresponds to the second clip for the activity. In the first experimentation cycle, he then ran a total of three trials as indicated by the "microworld run" statements at time 94s, 117s, and 140s. For each trial, the student changed only the "level of heat" variable in a successive manner, comparing a 'low' level to 'medium', and then 'medium' to 'high'. The

student spent 55 seconds doing so. In the second experimentation cycle, the student did not collect any additional data or specify any additional hypotheses. This is seen in the listing of "Hypothesis" followed by "Experiment", indicating they transitions between inquiry stages, with no additional actions under those headings. Due to the consistency in manipulating only one variable at a time between trials, we would label this as demonstrating the "designing controlled experiments" behavior.

When tagging clips we noticed that students' experimentation patterns could vary greatly. In some cases, students collected only a single trial of data whereas others collected many trials. Students may have focused on only one variable, or changed all four several times as part of their experimentation process. No matter what process a student has followed, a human coder needs to judge, on the whole, whether their data collection is indicative of the two skills or not. We next describe the process by which we generated a corpus of clips for training and testing our skill detectors, and the process by which we achieved consistency in identifying skill from student interaction data.

#### 4.1.3.1 Clip Tagging Procedure

To build detectors in the context of the Phase Change activities, two human coders tagged a subset of the data collection clips (Section 4.1.2) to generate a corpus of hand-coded clips for training and validating our detectors. The subset contained one randomly chosen clip (e.g. first clip, second clip, etc.) for each student-activity pair, resulting in 571 clips. This ensured a representative range of student clips were coded. The human coders tagged the same first 50 clips to test for agreement, ensuring the skills could be consistently identified. The remaining clips were split for each to code separately. Each coder independently tagged about 260 clips each in three to four hours.

Agreement for the 50 clips tagged by both coders was high overall. Since each clip could be tagged with one or several tags, agreement was determined by computing separate Cohen's Kappa

values for each tag. Over all ten possible clip tags (see Section 4.1.3), there was an average agreement of $\kappa = 0.86$. Of specific importance to this work, there was good agreement on the designing controlled experiments, $\kappa = .69$, and perfect agreement between coders for testing stated hypotheses ($\kappa = 1.00$). The high degree of agreement was achieved in part through extensive discussion and joint labeling prior to the inter-rater reliability session. Even though the agreement on designing controlled experiments was lower than the other behaviors, all Kappas were at least as good as the Kappas seen in previous text replay approaches leading to successful behavior detectors. For example, Baker et al. (2006) reported a Kappa of .58, and Baker et al. (2010a) reported a Kappa of .80 when labeling "gaming the system" behavior in clips from two different learning environments.

The human coders tagged 31.2% of the clips as showing evidence of designing controlled experiments; 34.4% were tagged as showing evidence of collecting data to test specified hypotheses. Planning behaviors involving the data table and hypothesis list were relatively rarer. Only 8.2% and 3.5% of the clips were tagged as exhibiting planning using the data table tool and hypothesis list viewer, respectively.

Next, we describe the feature distillation process. Features were distilled that, when combined with the classification labels from text replay tagging, yielded data instances for machine learning.

### 4.1.4 Feature Distillation

Seventy-three features were selected based on features used in previous detectors of other constructs (e.g. Walonoski & Heffernan, 2006; Baker, et al., 2008b,c), and previous work that identified indicators of systematic and haphazard inquiry behavior (e.g. de Jong, 2006; Buckley, Gobert & Horwitz, 2006; Gobert, et al., 2010; McElhaney & Linn, 2008, 2010). These features are aggregated over significant portions of students' inquiry, such as the experimental setups students designed within an activity, rather

than step or transaction-level data, unlike in many prior models of student behavior (e.g. Beck, 2005; Walonoski & Heffernan, 2006; Baker & de Carvalho, 2008; Amershi & Conati, 2009; Cetintas et al., 2010; Baker et al., 2010a).

The classes of features, summarized in Table 2, included: variables changed when making hypotheses, hypotheses made, total trials run, incomplete trials run (runs in which the student paused and reset the simulation), complete trials run (runs in which the student let the simulation run until completion), simulation pauses, data table displays, hypothesis list displays, variable changes made when designing experiments, and the total number of all actions (any action performed by a student). For each feature class, we computed a count each time the action was taken as well as timing values, similar to approaches taken by Walonoski and Heffernan (2006) and Baker et al. (2008b,c). Timing features included the minimum, maximum, standard deviation, mean and median. We also included a feature for the activity number associated with the clip since students may exhibit different behaviors for each of the activities.

Two additional feature counts specifically related to systematic data collection were also computed. The first was a unique pairwise controlled trials count using the Control of Variables Strategy (CVS), a count of the number of unique trials in which only one factor differed between them. This was similar to the approach taken by McElhaney and Linn (2010) to assess CVS, except they computed a pairwise CVS count for *adjacent* trials (i.e. trial *n* and trial *n+1* demonstrate CVS). Our count, on the other hand, tallied any pairwise CVS trials. This choice was made because students had the opportunity to view their previous trials in the data table, and could judge if they had adequate data or needed to run more trials. The second feature was a repeat trial count (Buckley, Gobert & Horwitz, 2006; Gobert, et al., 2010), the total number of trials with the same independent variable selections. These authors hypothesized that repeating trials is indicative of haphazard inquiry. It is worth noting that repeat trials were not included in the CVS count; that count only considered unique trials.

We computed feature values at two different levels of granularity. Recall that within a phase change activity, a student could make and test several hypotheses, thus generating multiple clips for that activity. Feature values could thus be computed *locally*, considering only the actions within a single clip, or *cumulatively*, taking into account all actions within predecessor clips for an activity. For each data instance, we computed two values for each feature, one local and one cumulative, with the aim of comparing the effectiveness of each feature set in predicting data collection behavior.

**Table 2.** Summary of all 73 distilled features used to build detectors.

| | | Time information | | | | | |
|---|---|---|---|---|---|---|---|
| Feature Classes | Count | Total | Min | Max | Mean | SD | Median |
| All actions | X | X | X | X | X | X | X |
| Hypothesis variable changes | X | X | X | X | X | X | X |
| Hypotheses added | X | X | X | X | X | X | X |
| Data table use | X | X | X | X | X | X | X |
| Hypothesis list use | X | X | X | X | X | X | X |
| Simulation variable changes | X | X | X | X | X | X | X |
| Simulation pauses | X | X | X | X | X | X | X |
| Total trials run | X | X | X | X | X | X | X |
| Incomplete trials run | X | X | X | X | X | X | X |
| Complete trials run | X | X | X | X | X | X | X |
| Repeated trials | X | | | | | | |
| Unique pairwise CVS trials | X | | | | | | |
| Activity number | X | | | | | | |

**Table 3.** Example instances used to build and validate data-mined models with local (Loc) feature values, cumulative (Cu) feature values, and labels from text replay tagging. The row in boldface and italics corresponds with the clip coded via text replay tagging in Figure 10. For the boldface clip, note that since no additional actions were taken in the second clip, the local feature values are zero. However, since actions were performed in the first clip, the cumulative features are nonzero.

| Student | Clip | Activity | Loc: Action Count | Loc: Action Total Time | ... | Loc: Repeat Trial Count | Cu: Action Count | Cu: Action Total Time | ... | Ctrlled Exps? | Test Hyps? | Plan Data | Plan Hyp List? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ... | | | | | | | | | | | | | |
| 81850 | 2 | 3 | 7 | 120 | | 0 | 15 | 155 | | N | Y | N | N |
| 81843 | 2 | 1 | 0 | 0 | | 0 | 7 | 69 | | N | N | N | N |
| ***85240*** | ***2*** | ***1*** | ***0*** | ***0*** | ... | ***0*** | ***22*** | ***208*** | ... | ***Y*** | ***Y*** | ***Y*** | ***Y*** |
| 78382 | 2 | 3 | 13 | 79 | | 0 | 33 | 224 | | Y | Y | N | N |
| 85238 | 2 | 3 | 8 | 145 | | 0 | 16 | 219 | | N | Y | N | N |
| ... | | | | | | | | | | | | | |

Cumulative and local feature values were computed for the corpus tagged clips generated from students' experimentation with the Phase Change activities (Section 4.1.3.1). Combining the feature values with the labels yields a dataset from which skill detectors can be built and validated. An excerpt of this dataset is shown in Table 3. Next, we describe our first attempt at building and validating the skill detectors using these data.

### 4.1.5 Development of First Detectors and Validation Approach

Data-mined skill detectors were generated and validated using the corpus of hand-coded clips and summary features (an example of these data appear in Table 3). The first version of these detectors were developed within RapidMiner 4.6 (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006) using the following procedure. First, redundant features correlated to other features at or above 0.6 were removed. Then, detectors were constructed using J48 decision trees, an open-source implementation of the C4.5 decision tree algorithm (Quinlan, 1993), with automated pruning to control for over-fitting. More specifically, two algorithm parameters were set to control for over-fitting; the minimum number of instances per leaf (*M*) was set to 2, and the confidence threshold for pruning (*C*) was set to 0.25.

This technique was chosen for three reasons. First, J48 decision trees have led to successful behavior detectors in previous research (Walonoski & Heffernan, 2006; Baker & de Carvalho, 2008). Second, decision trees produce relatively human-interpretable rules. Finally, such rules can be easily integrated into our environment to evaluate demonstration of inquiry skills, in real-time.

To validate our detectors, we employed six-fold student-level cross-validation. In this process where models (in this case detectors) are built using one group of students' data and tested on new students whose data were not in model construction (e.g. Pardos et al., 2011). This ensures models will remain accurate when applied to completely new students. Thus, the detectors were trained on five randomly selected groups of students' clips and tested on a sixth group of students' clips, repeating the process such that each student's data were used to test the models exactly once.

The goodness of detectors was assessed using two metrics, A' (Hanley & McNeil, 1982) and Kappa. 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' approximates the area under the ROC curve in signal detection theory. A' is also equivalent to the Wilcoxon statistic (Hanley & McNeil, 1982). Under this view, a model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. In these analyses, A' was computed at the level of clips, rather than students, using the AUC (area under the curve) approximation. Statistical tests for A' are not presented in this paper. An appropriate statistical test for A' in data across students would be to calculate A' and standard error for each student for each model, compare using *Z* tests, and then aggregate across students using Stouffer's method (cf. Baker et al., 2008a,b). However, the standard error formula for A' (Hanley & McNeil, 1982) requires multiple examples from each category for each student, which is infeasible in the small samples obtained for each student (a maximum of four) in our text replay tagging. Another possible method, ignoring student-level differences to increase example counts, biases undesirably in favor of statistical significance.

Second, we used Cohen's Kappa (κ), which assesses whether the detector is better than chance at identifying the correct action sequences as involving the category of interest. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly.

A' and Kappa were chosen because, unlike accuracy, they attempt to compensate for successful classifications occurring by chance (cf. Ben-David, 2008). Thus, we can achieve a better sense of how well our detectors can classify given our corpus' unbalanced labels, with between 4% and 34% of instances labeled as positively demonstrating one of the behaviors. We note that A' can be more sensitive to uncertainty in classification than Kappa, because Kappa looks only at the final label whereas A' looks at the classifier's degree of confidence in classifying an instance.

### 4.1.6 Analysis of Machine-Learned Classifiers

In our analyses, we determined if machine-learned detectors could successfully identify the two data collection skills of interest: testing hypotheses and designing controlled experiments. As part of this goal, we compared whether detectors built with features computed using the current and all predecessor clips (cumulative features) achieved better prediction than those built with features looking solely at the current clip (local features), as measured by A' and Kappa. We hypothesized that cumulative features would yield better detectors, because the additional information from previous clips may help more properly identify systematic behavior. For example, since students could re-use previous trials to test new hypotheses, actions in subsequent clips may be more abbreviated. Thus, taking into account previous actions could provide a richer context to identify and disambiguate behavior.

Separate detectors for each skill were generated from each feature set, resulting in four different detectors. Separate detectors were built for each skill, as opposed to one detector to classify all behaviors/skills, for two reasons. First and most important, the different behaviors were not necessarily

mutually exclusive; they could be demonstrated simultaneously in each clip. Building a single classifier only would allow for finding a *single* behavior within a clip. Second, the number of instances available for training and testing were slightly different for each behavior. This occurred, because for a small set of clips in the corpus that both human coders tagged, there was disagreement as evidenced by the imperfect inter-rater reliability measures (see Section 4.1.3.1). Within that set, we only used clips where the two coders were in agreement for a specific behavior.

The confusion matrices capturing raw agreement between each detector's prediction and the human coders' tagging under student-level cross-validation are shown in Table 4. Overall, we found that the detectors were good overall and that there were no major differences between detectors built using cumulative versus noncumulative attributes as measured by A' and Kappa. The designing controlled experiments detector using cumulative attributes (A' = .85, $\kappa$ = .47) performed slightly better than the detector built with non-cumulative attributes (A' = .81, $\kappa$ = .42). The hypothesis testing detector built with cumulative attributes (A' = .85, $\kappa$ = .40) had a slightly higher A', but slightly lower Kappa than the non-cumulative detector (A' = .84, $\kappa$ = .44). We do note, however, that these detectors appear to bias towards inferring that a student is not demonstrating skill. This is indicated by recall values ranging from 51% to 63% for cumulative attribute-based detectors and 30% to 53% for local attribute-based detectors (shown in Table 4). Thus, these detectors are most appropriate for use in fail-soft interventions, where students assessed with low confidence (in either direction) can receive interventions that are not costly if misapplied. Overall, the performance of these detectors as measured by A' and Kappa, is comparable to detectors of gaming the system refined over several years (e.g., Baker & de Carvalho, 2008; Baker et al., 2010a).

**Table 4.** Confusion matrices for each behavior's cumulative and non-cumulative attribute-based detector tested under six-fold student-level cross-validation.

| | | Designing Controlled Experiments | | Testing Stated Hypotheses | |
|---|---|---|---|---|---|
| | | True N | True Y | True N | True Y |
| Cumulative Features | Pred N | 325 | 65 | 303 | 79 |
| | Pred Y | 63 | 111 | 71 | 117 |
| | | A'=.85, K=.47 | | A'=.85, K=.40 | |
| | | Pc=.64, Rc=.63 | | Pc=.62, Rc=.60 | |
| | | True N | True Y | True N | True Y |
| Local Features | Pred N | 343 | 87 | 331 | 93 |
| | Pred Y | 45 | 89 | 43 | 103 |
| | | A'=.81, K=.42 | | A'=.84, K=.44 | |
| | | Pc=.66, Rc=.51 | | Pc=.71, Rc=.53 | |

Note: Pc = Precision, Rc = Recall

### 4.1.7 Inspecting the Data Collection Models More Closely

In this section, we show the final model of each behavior, generated using all hand-coded clips. We focus on the detectors built from cumulative features, due to the slightly better performance obtained using this method. Since we showed that student under cross-validation performance of each detector was acceptable, we construct 'final' models using all the hand-coded clips. The same procedure of removing correlated features at the 0.6 level and building decision trees is used.

Prior to constructing these final detectors, we again removed correlated features at the 0.6 level, reducing the number of features from 73 to 25. Most features were time-based, though some count-based features remained: repeat trials, incomplete runs, hypotheses added, data table and hypothesis list uses, and the total number of actions. Additionally, how many activities the student had completed so far remained. These features were used to construct the three decision trees for each behavior. The resulting

trees for the designing controlled experiments and testing hypotheses detectors were wide and complex. Portions of the decision tree for the designing controlled experiments detector is shown in Figure 5.

The designing controlled experiments tree had 46 leaves and 91 nodes and used nearly all of the features remaining after correlation filtering. The root note, as shown in Figure 5, was "median time spent changing simulation variables" feature, and indicated that if variables were not changed (i.e. median time = 0), then the clip did not exhibit behavior in line with designing controlled experiments (in this case, no experiments were designed at all). The next level down branched on "minimum time running trials <= 1", a proxy for the number of runs. If this value were zero, it would mean that the clip contained no simulation runs, an indicator that the clip did exhibit the behavior. However, if this value were one, it would mean that simulation runs occurred, but in at least one case was very quick, one second or less. This potentially indicates a student realizing the trial is unnecessary, in turn a potential indicator of inquiry skill.

As a result of this complexity, hierarchy under this branch is complex and tries to disentangle if the clip exhibits systematic behavior using several features primarily associated with simulation runs and microworld simulation variable changes. On the other hand, if the minimum time running trials was greater than 1, the lower branches involved time features associated with running simulations, the number of hypotheses made, the number of times the data table was displayed, and the median time for all actions. Overall, the tree is predominantly based on time features associated with number of runs, number of actions, and simulation variable changes to distinguish behavior. It is worth noting, incidentally, that this tree did not contain the pairwise unique CVS count feature since it was pruned during the "remove correlated features" step. It could be that the tree would be more compact had this feature been considered. However, the features included captured the designing controlled experiments behavior successfully, given the good cross-validated A' and Kappa achieved by this detector.

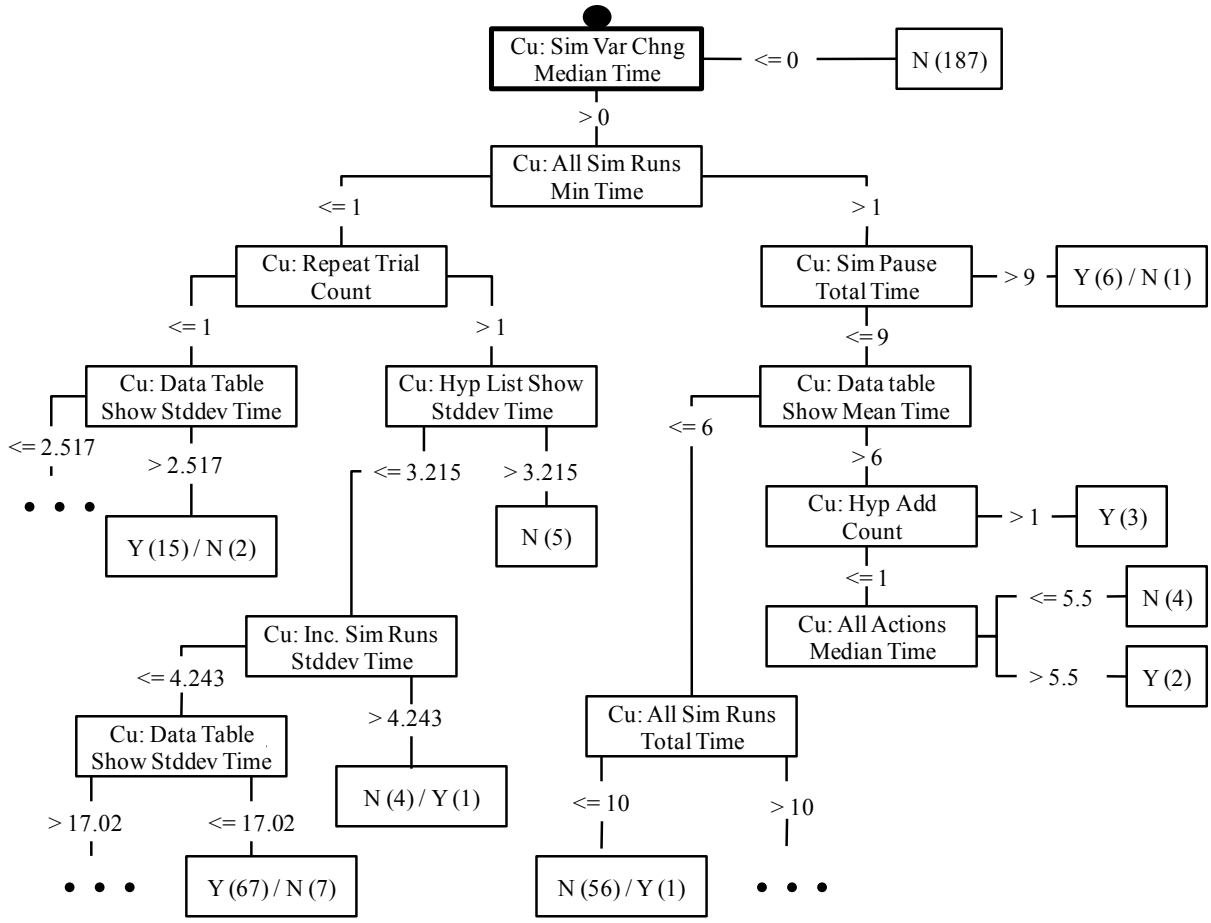


**Figure 5.** Portion of the decision tree for the designing controlled experiments behavior.

### 4.1.8 Discussion on the First Version of Detectors

Overall, our results for classifying student behavior with the initial set of detectors were very promising. We can distinguish a set of trials in which a student designed controlled experiments from a set of trials in which students did not design controlled experiments 85% of the time. We can also distinguish a set of trials in which a student tested their stated hypotheses from a set of trials in which they did not 85% of the time. Furthermore, the associated Kappa values, ranging from .40 to .47, indicate that each of these detectors is better than chance.

Given this level of cross-validated performance, the detectors appear to be sufficiently accurate to be used to select students for scaffolding *when they finish their experimentation*. However, it is

important to note that the detectors bias towards false negatives, as evidenced by the fact that clips tagged as "Y" were only correctly identified between 51% and 63% of the time under student-level cross-validation. As such, any scaffolds administered based on the detectors' classification should be fail-soft. For example, if the system detects that a student is not designing controlled experiments, the student can be provide a small reminder about how to correctly design controlled experiments.

There are two other important caveats to these results. First, we aim to use the detectors for two purposes: (1) evaluation of students' data collection for assessment purposes, and (2) evaluation of students' data collection to drive automatic scaffolding. Up to this point, they have been validated at the level of a full data collection, the *clip level.* To reiterate, this means they have been validated to discern which students designed controlled experiments and/or tested their stated hypotheses at the time they completed their data collection (thus, signifying they are ready to move to the "analyze data" stage of inquiry in our learning environment). The detectors have not been validated to be used sooner, to correctly identify which students are off-track before they finished their data collection. This latter point is of importance if we aim to provide scaffolding that proactively responds to students' experimentation as they are working (Gobert et al., 2012).

Second, though the detector worked well as measured by agreement with humans' labels of clips tagged at the clip level, we noticed the models did not contain features considered theoretically important to demonstration of skill (see Section 4.1.7, Chen & Klahr, 1999; Buckley et al., 2006; McElhaney & Linn, 2010). For example, a feature that tracked the number of controlled comparisons a student made in their dataset was not present in the final model (Sao Pedro et al., 2013b), though this is a known indicator of skill at designing controlled experiments (Chen & Klahr, 1999; McElhaney & Linn, 2010). In addition, other features that were not considered important were present in the model. As an example, a feature related to showing or hiding the hypothesis list appeared in the model (Sao Pedro et al., 2013b). We believe that, despite predicting the labels well, the detectors did not fully reflect the

skills, meaning that they had low construct validity, due to poor feature selection. In turn, we hypothesize that believe this may negatively have impacted their predictive performance.

Thus, we re-engineered new models and tested if selecting features to increase their construct validity also increased their predictive capability. We also evaluated them to see how quickly they could be used to accurately trigger scaffolding, *before* students complete their experimentation. We describe our approach to building these new detectors and our evaluation of them in the next section.

## 4.2 Second Version of the Detectors: Does Improving their Construct Validity Improve their Predictive Ability, Even with Less Data?

Feature selection, the process of pre-selecting features before running a data mining algorithm, can improve the performance of data mining algorithms (cf. Witten & Frank, 2005). Several automated approaches exist for finding optimal feature sets such as filtering redundant features (Yu & Liu, 2003), conducting heuristic searches (cf. Pudil, Novovicova, & Kittler, 1994), using genetic algorithms (Oh, Lee, & Moon, 2004), and clustering (Bernardini & Conati, 2010). These procedures, though powerful, may yield sets that domain experts would not intuitively expect to align with the target class (construct). An alternative is to select features that specifically improve models' construct validity.

This alternative is motivated by our prior work in developing automated detectors of two scientific inquiry skills, designing controlled experiments and testing stated hypotheses, within a science microworld (Sao Pedro et al., 2013b). To build them, we first filtered features that correlated highly with each other, and then constructed J48 decision trees (see Section 0). The resulting detectors worked well under student-level cross-validation. However, upon inspecting them more closely, we noticed some features considered theoretically important to the constructs (Chen & Klahr, 1999; McElhaney & Linn, 2010; Buckley et al., 2006) were eliminated at the filtering step. Also, other features without theoretical

justification remained. We believe this feature selection process may have yielded a feature set that did not represent all aspects of skill demonstration, which in turn may have negatively impacted their predictive performance.

Thus, we explore here whether selecting features with the goal of increasing a model's construct validity and interpretability can also improve a model's predictive ability. We do so by comparing two types of detectors for each skill. One type is built with an automated feature selection strategy used in our original detectors (Sao Pedro et al., 2013b). The other type is built using a combination of manual selection and statistics to select successful features that theoretically align more closely with skill demonstration. We compare the predictive performance of the two types of detectors against a held-out test set in two ways. First, we compare the detectors' ability to predict skill demonstration at the level of a full data collection cycle. This enables us to measure how well the detectors can be used for assessing performance, or for identifying which students need scaffolding when they claim to finish collecting data. In addition, it is useful to have detectors that can identify a student's lack of skill as quickly as possible so the software can "jump in" and support the student as soon as they need it to prevent frustration, floundering, or haphazard inquiry (Gobert et al., 2012). Thus, the second way we compare detectors is to determine how much student data is needed before inquiry skill demonstration can be accurately predicted. The faster detectors can make valid inferences, the faster the system can help the students who need it.

### 4.2.1 Data Sets for Constructing, Comparing and Validating Detectors

Clips were generated from 148 suburban Central Massachusetts middle school students' interactions within a sequence of four Phase Change activities. These students and activities were the same ones used to build the first version of the skill detectors (Sao Pedro et al., 2013b). Text replay tagging (Baker,

Corbett & Wagner, 2006) was again employed to tag new clips. The following training, validation and test data sets were created:

- *Training Set (601 clips).* Initially, two human coders tagged 571 clips for training and cross-validating the detectors in (Sao Pedro et al., 2013b; See Section 4.1.3.1). Since several clips could be generated per activity, a single, randomly chosen clip was tagged per student, per activity. This ensured all students and activities were equally represented in this data set. Inter-rater reliability for the tags was high overall ($\kappa$=.69 for designing controlled experiments, $\kappa$=1.0 for testing stated hypotheses). By chance, the stratification yielded few first clips, clips representing students' first data collection within an activity. To have a more representative training set, an additional 30 randomly selected first clips were tagged. In total, 31.4% of the clips were tagged as designing controlled experiments, and 35.6% as testing stated hypotheses.
- *Validation Set (100 clips).* A special set of clips was tagged by one human coder for engineering detectors with improved construct validity (described in more detail later). This set contained 20 randomly chosen first clips, 20 randomly chosen second clips, up through fifth clips. Clips were not stratified by student or activity. More stringent student or activity-level stratification was not used, because all students and activities were used to build the training set. Stratification would not remove biases already present in this data set. In total, 34.0% were tagged as designing controlled experiments, and 42.0% as testing stated hypotheses.
- *Held-out Test Set (439 clips).* A human coder tagged all remaining first through fourth clips in the data set for comparing detectors. This set did not contain fifth clips because only 2 remained in the tagged corpus. First clips in which one or no simulation runs occurred were also excluded, because demonstration of the inquiry skills requires that students run the simulation at least twice (Gobert et al., 2012). Such clips would trivially be identified as not demonstrating either skill and could bias our comparisons. This set had 64.7% tagged as designing controlled experiments and

61.0% as testing stated hypotheses. Note that the data distribution of the skill demonstration was different in the held-out test set than the other data sets. This occurred due to random chance, but provides an opportunity to conduct stringent validation, since the base rates will be different in this data set than the other data sets.

Feature sets computed over clips, combined with text replay tags, form the basis for training and testing the detectors. Since the aim of this work is to compare models built from different feature sets, we discuss the feature generation and selection processes in more detail in the following section.

### 4.2.2 Feature Selection and Detector Construction

Our original designing controlled experiments and testing stated hypotheses detectors considered 73 features associated with a clip (Sao Pedro et al., 2013b, see Section 4.1.4). Feature categories included: variables changed when making hypotheses, full hypotheses made, simulation pauses, total simulation runs, incomplete simulation runs (paused and reset before the simulation finished), complete simulation runs, data table displays, hypothesis list displays, variable changes made when designing experiments, and total actions (any action performed by a student). For each category, counts and timing values (min, max, standard deviation, mean and mode) were computed. In addition, the specific activity number associated with the clip was also included. A pairwise repeat trial count, the number of all pairs of trials with the same independent variable values (Buckley et al., 2006), was also included, as was a unique pairwise controlled trial count, the number of non-repeated trials in which only one independent variable differed between them (cf. Chen & Klahr, 1999). All features were computed cumulatively, taking into account actions in predecessor clips, as in Sao Pedro et al. (2013b). For example, given the actions shown in Figure 11, the total number of runs for clip 2 would be 5 (assuming no more runs had occurred after action 40).

We added five additional features to this set which seemed to have face validity as potential predictors of the demonstration of each skill, giving a total of 78 features. In specific, we added *adjacent* counts for unique controlled trials and repeats. These are counts of successive trials (e.g. trial 2 vs. 3, 3 vs. 4) in which only one variable was changed (controlled) or all variables were the same (repeated). Since the controlled trials counts excluded repeat trials, we added two additional counts for controlled trials that did allow them, one pairwise and one adjacent. Finally, we added a feature to count when simulation variables explicitly stated in hypotheses were changed.

Two different approaches for feature selection over this set were employed to form detectors. The first approach removed correlated features prior to building detectors (RCF detectors). The second approach involved selecting features geared at improving construct validity (ICV detectors). These procedures are discussed below.

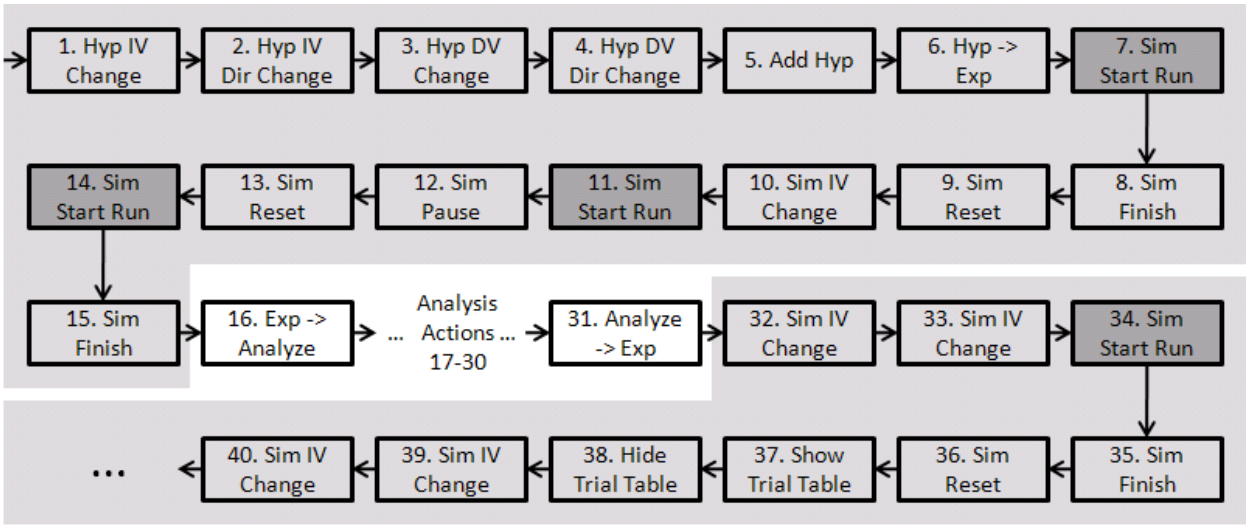


**Figure 11.** Example sequence of student actions for a phase change activity. Two clips (shown in light grey) would be generated since the "Experiment" stage was entered twice.

### 4.2.3 Removed Correlated Features (RCF) Detector Construction

The original models in Sao Pedro et al. (2013b) were built in RapidMiner 4.6 as follows. First, redundant features correlated to other features at or above 0.6 were removed. Then, J48 decision trees, a Java-based implementation of C4.5 decision trees with automated pruning to control for over-fitting (Quinlan, 1993), were constructed (see Section 0) for more explicit information about the decision tree parameters). The RCF detectors of each skill developed in this paper were built using this same process. However, they instead were built from the new feature set (78 features), and the enhanced training corpus.

The initial remove correlated features procedure eliminated 53 features. Of the 25 remaining features, 19 were timing values associated with the following feature classes: all actions, total simulation runs, incomplete simulation runs, simulation pauses, data table displays, hypothesis table displays, variables changed when making hypotheses, full hypotheses made, and simulation variable changes. The remaining 6 features were activity number and counts for the following feature classes: all actions, incomplete simulation runs, data table displays, hypothesis list displays, full hypotheses created, and adjacent repeat trials count (one of the new features added). RCF detectors for designing controlled experiments and testing stated hypotheses were then built based on this set of 25 features. Their performance will be discussed later in the Results section.

We note that this procedure eliminated some features which are considered theoretically important to both constructs. For example, counts for controlled trials, total simulation runs, and simulation variables stated in hypotheses changed were all filtered. These features are important, because they reflect theoretical prescriptive models of how data should be collected to support or refute hypotheses. Constructing controlled trials is seen as a key procedural component in theory on designing controlled experiments (cf. Chen & Klahr, 1999). Similarly, running trials and changing values of the variables explicitly stated in the hypotheses both play roles in determining if hypotheses are supported.

In addition, some features remaining did not immediately appear to map to theory on these constructs, such as the number of times that the student displayed the hypothesis viewer or data table. As discussed previously, we hypothesize these RCF detectors will not perform as well as detectors, because the remaining features do not theoretically align as well with demonstration of skill. Next, we describe how we selected features to yield detectors with improved construct validity (ICV detectors), which may in turn improve predictive performance.

### 4.2.4 Improved Construct Validity (ICV) Detector Construction

We selected features for the new detectors with increased construct validity (ICV) using a combination of theory and search. We first sought to understand how individual features related to the constructs. This was done by identifying which features had linear correlations to demonstration of skill at or above 0.2. Several features did so with both skills: all actions count, total run count, complete run count, variable changes made when designing experiments, changes to variables associated with stated hypotheses when designing experiments, adjacent and pairwise controlled experiments counts (both with and without considering repeats), and pairwise and adjacent repeat trials counts. An additional feature correlated with designing controlled experiments, the number of simulation pauses. From this set of 11 features, the counts for controlled trials, repeat trials, and changing variables associated with stated hypotheses are all features used by others to directly measure procedural understanding associated with skill demonstration (Chen & Klahr, 1999; McElhaney & Linn, 2010). The other features, though not directly related, may also help distinguish procedural understanding. Thus, we kept all 11 features for the next round of feature selection.

From here, we reduced the feature set further by performing separate manual backwards elimination search (cf. Witten & Frank, 2005) for each construct as follows. Features were first ordered in terms of the theoretical support for them by a domain expert. Then, features were removed one at a

time, starting with the one with the least theoretical support. From this candidate feature set, a decision tree was constructed using the training set. The resulting model's predictive performance was then tested on the *validation set* of 100 clips. If the candidate model yielded better performance than its predecessor, it was kept. If it did not, the candidate was rejected and another feature with low theoretical support was removed to form a new candidate set. This process was repeated, removing one feature at a time, until performance no longer improved.

Predictive performance was measured using A' (Hanley & McNeil, 1982) and Kappa (κ). Briefly, A' is the probability that when given two clips, one labeled as demonstrating skill and one not, a detector will correctly identify which clip is which. An A' of 0.5 indicates chance-level performance, 1.0 indicates perfect performance. Cohen's Kappa (κ) assesses if the detector is better than chance at identifying demonstration of skill. κ of 0.0 indicates chance-level performance, 1.0 indicates perfect performance. When comparing two candidate models, the model with higher κ was preferred. However, if A' decreased greatly and κ increased slightly, the model with higher A' was chosen. If two models yielded the same values, the model with fewer features was chosen.

The best ICV detectors of each construct performed well over the validation set. The best designing controlled experiments ICV detector had 8 features (total run count and pause count were removed) and had A'=1.0 and κ=.84. The best testing stated hypotheses ICV detector had 5 features: variable changes made when designing experiments (both related and unrelated to stated hypotheses), unique pairwise controlled trials, adjacent controlled trials with repeats considered, and complete simulation runs. Its performance on the validation set was also strong (A'=.96, κ=.77).

### 4.2.5 Results: Comparing Predictive Capabilities of Detectors

Having created these two sets of detectors (RCF and ICV), we now can study whether selecting features more theoretically aligned with the two skills will yield better detectors than more traditional approaches. There are two key questions we address. First, which detectors predict best overall? Second, how quickly can detectors identify the two inquiry skills? Performance will be compared against the *held-out test set* only, rather than using cross-validation over all datasets. This was done for two reasons. First, the entire training set was used to select features for the ICV detectors. Using the full training set enabled us to understand the relationships between individual features and demonstration of skill more thoroughly. Second, the search procedure for building ICV detectors likely overfit them to the validation set data.

#### 4.2.5.1 Comparing Detectors' Overall Performance

We compared detectors' performance at classifying demonstration of skill in the held-out test set, labeled at the clip level. As a reminder, this comparison measures how well the detectors can be used for assessing performance, or identifying which students need scaffolding when they claim to be finished collecting data. Detectors are compared using A' and Kappa (κ). These were chosen because they both try to compensate for successful classification by chance (Ben-David, 2008), and have different tradeoffs. A' can be more sensitive to uncertainty, but looks at the classifier's degree of confidence; κ looks only at the final label, leading to more stringent evaluation. We note that statistical tests comparing models' A and κ are not performed. This is because students contribute multiple clips in the test set, and thus independence assumptions are violated. Meta-analytical techniques do exist to handle this (e.g. Fogarty et al., 2005), but our data did not have enough data points per student to employ them.

As shown in Table 5 the detectors with improved construct validity (ICV) detectors outperformed the removed correlated features (RCF) detectors within the held-out test set. For designing controlled experiments, both the RCF (A'=.89) and ICV (A'=.94) detectors were excellent at distinguishing this construct. However, the ICV detector was better at identifying the correct class (RCF κ=.30 vs. ICV κ=.45). Both detectors seem to bias towards labeling students' experimentation as "not designing controlled experiments", as indicated by lower recall rates than precision rates (RCF recall=.46, precision=.90 vs. ICV recall= .58, precision=.95). This suggests that more students would receive scaffolding than necessary upon finishing data collection.

Upon inspecting the results for designing controlled experiments more closely, we noticed a large number of first clips with exactly two simulation runs had been misclassified. These kinds of clips comprised 26.7% of the held-out test corpus. When filtering these out (leaving 322 clips), the performance of the ICV detector was substantially higher (ICV A'=.94, κ =.75, recall=.83). The RCF detector's performance was also higher (RCF A'=.90, κ =.44, recall=.56), but did not reach the level of the ICV detector. The implications of this will be discussed later.

For the testing stated hypotheses skill, the ICV detector again showed a substantial improvement over the RCF detector. The ICV detector was around ten percentage points better at distinguishing between the two classes (RCF A'=.82 vs. ICV A'=.91). Furthermore, κ and recall were much higher for the ICV detector than the RCF detector (RCF κ =.24, recall=.44 vs. ICV κ =.70, recall=.86). The ICV detector is therefore quite good at selecting the correct class for a clip, and has much less bias towards labeling students' experimentation processes as "not testing stated hypotheses".

Though not shown in Table 5, the ICV and RCF detectors were also compared to our original detectors (Sao Pedro et al., 2013b), which used the original 73 features and had correlated features removed. Performance on the held-out test set was slightly worse than the RCF detector described here for designing controlled experiments (A'=.86, κ=.28, recall=.42), but slightly better for testing stated

hypotheses (A'=.83, κ=.30, recall=.49). The new ICV detectors still outperform these detectors by a substantial amount. In sum, these findings support the idea that improving construct validity can lead to better overall prediction of systematic inquiry. Next, we determine if the ICV detectors can infer demonstration of skill with fewer actions.

**Table 5.** Confusion matrices and performance metrics for detectors' overall predictions.

| Designing Controlled Experiments | | | | | Testing Stated Hypotheses | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | RCF Detector | | ICV Detector | | | RCF Detector | | ICV Detector | |
| | True N | True Y | True N | True Y | | True N | True Y | True N | True Y |
| Pred N | 140 | 153 | 146 | 118 | Pred N | 142 | 149 | 146 | 37 |
| Pred Y | 15 | 131 | 9 | 166 | Pred Y | 29 | 119 | 25 | 231 |
| | Pc = .90, Rc = .46 | | Pc = .95, Rc = .58 | | | Pc = .80, Rc = .44 | | Pc = .90, Rc = .86 | |
| | A' = .89, K = .30 | | A' = .94, K = .45 | | | A' = .82, K = .24 | | A' = .91, K = .70 | |

* Pc = precision; Rc = recall

### 4.2.5.2 Comparing Detectors' Performance Predicting with Less Data

The analyses here determine if detectors can predict demonstration of skill labeled at the clip level using less information. Again, these comparisons enable us to determine which detectors are more suitable for identifying which students need support *as they conduct their data collection*. Given our learning environment and approach, there are several ways to define "less information". We chose to look at simulation runs because they are the grain size at which we aim to activate scaffolding. In considering simulation runs, we also had to consider the clip number. Recall that several cycles of data collection could occur in an activity (each cycle represents a clip). Predictive performance could be impacted by the clip number under consideration, because later clips contain all actions associated with predecessor clips. Thus, we compare each detector on predicting skill demonstration labeled at the clip level using actions up to the $n^{th}$ run within the $m^{th}$ clip, for varying numbers of runs and clips.

This approach required new sets of feature values to accommodate the fewer actions. Feature values were computed using all actions from clips 1..*m*-1 ($m > 1$), and all actions in the *m*th clip, up to and including the *n*th "sim start run" action (actions in dark grey in Figure 11). As an example, the feature values for the action sequence in Figure 1 for clip 2 and two runs would be computed using all actions 1-16 from the first clip, and actions up to and including the second "sim start run" (actions 31-38) in clip 2. Note that the notion of a "full run" actually spans several actions (e.g. actions 11-13 in Figure 11), given that the student could let the simulation run to completion, pause the simulation, or reset it. The "sim start run" action was chosen (rather than "sim finish" or "sim reset") to denote the boundary due to considerations for how we would scaffold students. In particular, we may want to prevent students from collecting of data unhelpful for the subsequent stage of inquiry, where they analyze data. Having the detectors classify skill demonstration at the point where students try to run the simulation enables such an intervention.

We compare detectors' performance using less data by comparing predictions for a given clip-run combination against the ground truth labels at the clip level. The number of clips was varied from 1 to 4, and the number of runs was varied from 1 to 5. A' and κ were computed per combination. Our expectation is that as the number of runs considered increases (and correspondingly the number of actions considered increases), A' and κ will increase. However, since many clips had fewer than five simulation runs, performance metrics may plateau as the number of runs increases. This may occur because no additional information would be available to improve predictions.

As shown in Table 6, the ICV detectors match or outperform the RCF detectors, when both detector variants are given less data on student performance. For clip 1, neither detector performed well for one or two runs ($\kappa \cong 0.0$). This finding associated with one run matched expectations because demonstration of skill can only be identified after two or more runs (cf. Chen & Klahr, 1999). For runs 3-5 on the first clip, the RCF detector had A' ranging from .73 to .76, whereas the ICV detector had A'

ranging from .93 to 1.0. The RCF detectors' κ remained at chance levels ranging from .06 to .07. The ICV detectors' κ values were better but still low, ranging from .16 to .20.

The designing controlled experiments detectors' poor performance on first clips may be due to misclassifications of such clips with exactly two runs (see Section 4.2.5.1). To see if ignoring such clips would impact detectors' ability to classify with less data, we removed them from the test set and re-computed our performance metrics. With only first clips with at least three runs, both detectors' performance using fewer actions, up to the first and second run, remained very low. However, when using actions up to runs 3-5, the ICV detector (run 3: A'=.99, κ=.42; run 4: A'=1.0, κ=.65; run 5: A'=.91, κ=.47) outperformed the RCF detector (A'=.70-.79, κ=.06-.11 for the same values). Additionally, three runs was the level at which the ICV detector could perform as well as classifying when considering all actions in the first clip (ICV all actions A'=.89, κ=.50).

For later clips within an activity, both detectors reach predictive performance equivalent to considering all actions (the "all" columns Table 6) after a single run. However, the ICV detectors outperform the RCF detectors. For example, when looking at clip 2 / run 2, the ICV detector performs better (A'=.97, κ=.82) than the RCF detector (A'=.95, κ=.59). Thus, once students have begun their second data collection cycle within an activity, the ICV detectors can better judge who needs scaffolding after the first run.

For testing stated hypotheses, the ICV detector again matched or outperformed the RCF detectors as shown in Table 7. For first clips, the RCF detector had A' values ranging from .63 to .70, and κ values at chance levels. However, the ICV detector performed well at this skill for first clips (ICV all actions A'=.89, κ=.52), a difference from designing controlled experiments. In fact, it could properly identify skill demonstration after just the second run (ICV clip 1, run 2 had A'=.84, κ=.37). By the third run, predictive performance was on par with a detector that could consider all actions. For later clips, the ICV detector outperformed the RCF detector at all run levels. For example, when predicting using

actions up to the second run for clip 2, the RCF detector had A'=.92 and κ=.44. Though this performance is good, the ICV detector performed much better with A'=.95 and κ=.75. Thus overall, the ICV detectors can be used to classify testing hypotheses skill as early as the second run in the first clip, and are better at classification in later clips than the RCF detectors are.

**Table 6.** Designing controlled experiments performance over *n*-runs and *m*-clips

| Designing Controlled Experiments | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RCF Detector | | | | | | | ICV Detector | | | | | | |
| Runs | | 1 | 2 | 3 | 4 | 5 | All | Runs | 1 | 2 | 3 | 4 | 5 | All |
| Clip Num | 1 | .79 (.00) | .69 (.01) | .75 (.06) | .76 (.07) | .73 (.06) | .71 (.05) | 1 | 1.0 (.00) | 1.0 (.04) | 1.0 (.16) | 1.0 (.20) | .93 (.16) | .93 (.16) |
| | 2 | .92 (.39) | .95 (.59) | .94 (.59) | .95 (.61) | .95 (.61) | .95 (.61) | 2 | .98 (.66) | .97 (.82) | .97 (.85) | .97 (.85) | .97 (.85) | .97 (.85) |
| | 3 | .84 (.22) | .89 (.33) | .89 (.33) | .89 (.33) | .89 (.33) | .89 (.33) | 3 | .95 (.51) | .93 (.59) | .94 (.66) | .94 (.66) | .94 (.66) | .94 (.66) |
| | 4 | .89 (.57) | .84 (.46) | .84 (.46) | .84 (.46) | .84 (.46) | .84 (.46) | 4 | 1.0 (.90) | .99 (.79) | .99 (.69) | .99 (.69) | .99 (.69) | .99 (.69) |

* Each entry is in the format A' (K)

**Table 7.** Testing stated hypotheses performance over *n*-runs and *m*-clips

| Testing Stated Hypotheses | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RCF Detector | | | | | | | ICV Detector | | | | | | |
| Runs | | 1 | 2 | 3 | 4 | 5 | All | Runs | 1 | 2 | 3 | 4 | 5 | All |
| Clip Num | 1 | .70 (.01) | .66 (.06) | .63 (.02) | .63 (.01) | .66 (.04) | .65 (.04) | 1 | 1.0 (.00) | .84 (.37) | .86 (.49) | .91 (.54) | .89 (.53) | .89 (.52) |
| | 2 | .91 (.40) | .92 (.44) | .90 (.39) | .90 (.39) | .90 (.39) | .90 (.39) | 2 | .93 (.68) | .95 (.75) | .93 (.73) | .93 (.73) | .95 (.75) | .95 (.75) |
| | 3 | .88 (.50) | .87 (.47) | .87 (.47) | .87 (.47) | .87 (.47) | .87 (.47) | 3 | .93 (.86) | .89 (.79) | .87 (.76) | .88 (.76) | .89 (.79) | .89 (.79) |
| | 4 | .89 (.47) | .91 (.57) | .91 (.57) | .91 (.57) | .91 (.57) | .91 (.57) | 4 | .90 (.90) | .90 (.79) | .90 (.79) | .90 (.79) | .90 (.79) | .90 (.79) |

* Each entry is in the format A' (K)

### 4.2.6 Discussion of Second Version of Detectors with Increased Construct Validity

In this section, we investigated whether selecting features based on construct validity improves the predictive capabilities of machine-learned detectors that identify two scientific inquiry data collection skills as students experiment within the Phase Change activities (Gobert et al., 2012). These skills are designing controlled experiments and testing stated hypotheses. To explore this, we compared two types of detectors. One type used an automated approach, removing inter-correlated features (RCF detectors).

Another used a partially manual approach to select features theoretically aligned with the skills, thereby increasing construct validity (ICV detectors). Models' predictive performance was compared against a held-out test set in two ways. We predicted skill demonstration at the level of a full data collection cycle, the grain size at which behavior was labeled. We also predicted behavior at a finer grain size, microworld simulation runs, a grain size containing less information.

The results showed that improving construct validity can yield models with better overall predictive performance, even with less data. The ICV detector for testing stated hypotheses reached much higher performance levels than the RCF detector. The ICV detector can effectively be used to assess students' data collection or trigger scaffolding when students finish data collection, given its high A'=.91 and κ=.70 values. It also can be used after as few as two runs on students' first data collection to provide fail-soft interventions that are not costly if misapplied. This is evidenced by A' values at or above .84, and κ at or above .37 found when increasing the number of simulation runs (thereby increasing the number of actions available) to make predictions.

The ICV detector for designing controlled experiments also outperformed its RCF counterpart. However, both the ICV and RCF detectors performed poorly when they inferred behavior for students' first data collection within an activity. We discovered this was due, in part, to poor classification of first cycles containing exactly two simulation runs. When ignoring such cycles, the ICV detector's performance improved substantially while the RCF detector remained poor. It could be applied in as few as three runs on students' first data collection. We believe the ICV detector failed on this case because the training set did not contain enough cases of this kind (see Section 4.2.1 for more details). This issue may be alleviated by adding more of these training clips and re-engineering the ICV detector following our procedure.

These findings suggest that the detectors can be used to reliably assess students' data collection when they complete their experimentation, and to trigger scaffolding in as few as two or three runs

(depending on the detector). An important issue remains, however. This is the degree to which these detectors can predict systematic behavior for a different student sample and within other physical science microworlds (e.g. Gobert et al., under review). Currently, we have no reason to believe our detectors are over-fit to particular kinds of students since each student was represented equally within our training corpus, and we validated the detectors against a held-out test set. However, we only used students from one school, and it is possible that our detectors are over-fit to some aspect of learning and performance at this school. Furthermore, they were built using data from interactions within the Phase Change environment. Thus, it is also possible that the detectors are over-fit to the particular physical science domain. These questions of *generalizability* are of particular importance since we aim to test the efficacy of an automated scaffolding approach to help students as they collect data with a new set of students and within a different physical science topic, Free Fall. Thus, in the next section, we describe our approach for validating these detectors' use for a new cohort of students and the second physical science topic, Free Fall.

## *4.3 Detectors' Applicability to a new Science Topic*

Data mining/learning analytics is a powerful approach for building predictive models ("detectors") of student behavior (e.g. Baker et al., 2008c; Blikstein, 2011), and the demonstration of somewhat ill-defined skills (e.g. Ghazarian & Noorhosseini, 2010; Sao Pedro et al., 2013b) within interactive learning environments. This approach makes validation of models easier, because processes like cross-validation and validating against held-out test sets exist that can estimate how the models will generalize to new students and tasks not used to build them. Such estimates are important because can they provide assurance that models will correctly identify students who lack skill or engage in undesirable learning behaviors, enabling the system to provide accurate, real-time feedback (Siemens, 2012). Within

interactive environments, the estimates can also assure that models of skill demonstration can be reused for new tasks or domains and students, paving the way for reliable, scalable performance-based assessment.

In educational domains, validation is often done at the *student-level*, where models are built using one group of students' data and tested on new students whose data were not in model construction (Pardos et al., 2011; Baker et al., 2008c). This ensures models will remain accurate when applied to completely new students. It is possible, though, that this method may fail to identify specific instances when models do not predict adequately, particularly if some other aspect of those cases, other than the student, is not taken into account. We explore this topic here in the context of measuring how well two models (detectors) of scientific inquiry skill demonstration built and validated for one science topic (Sao Pedro et al., 2013b; Sao Pedro et al., 2012a) can predict demonstration of those skills for a new science topic and a new set of students. Few have tested model effectiveness on different topics (Baker et al. (2008c) is an exception), but validating at this level is essential if models will be used beyond the topics in which they were originally developed.

In our approach, we first take this new topic and student sample, and construct a test set stratified at the student level, where students are equally represented in the test set. When doing this, we find that there is an imbalance in the nature of behaviors demonstrated by students. In particular, there is an imbalance in the number of trials collected by students in this set, a factor which could influence predictive performance of our detectors (cf. Sao Pedro et al., 2012a; see Section 4.2.6). To address this, we construct a second test set, this time stratifying over the number of trials, to ensure a greater balance in student behaviors. We show that utilizing this different kind of stratification can unveil a different performance profile than conducting student-level validation alone, revealing new insights on the predictive limitations of the detectors.

To concretize this idea further, recall that in Section 4.2 (Sao Pedro et al., 2012a), we found that both detectors with improved construct validity could predict skill demonstration for unseen clips quite well. The designing controlled experiments model was quite good at distinguishing a clip in which the skill was demonstrated from a clip in which skill was not (A' = .94). It also matched the human coder's labels better than chance ($\kappa$ = .45). Similarly, the testing stated hypotheses model performed very well, A' = .91, $\kappa$ = .70. These findings meant that the data-mined models could adequately identify skill demonstration in the Phase Change inquiry activities for this sample of students.

Though their performance for Phase Change is encouraging, these metrics do not provide a measure of their generalizability to other science topics, because the models were built solely from data for the Phase Change activities. Furthermore, the model construction procedure in (Sao Pedro et al., 2012a) used the same students in the training/validation clip sets as in the test set. Thus, we aim here to explore the generalizability of these models to a new Physical Science topic, Free Fall. To do so, we collected data from new students who conducted inquiry Free Fall activities, tagged their resulting clips with skills, and re-measured the models' predictive performance. Using data from a different science topic enables us to estimate model transfer to different topics (Baker et al., 2008c). Using new students also enables us to estimate how well these models (detectors) will work for a broader range of students.

### 4.3.1 Data Sets for Estimating Detector Generalizability

We collected data from 292 eighth grade students' interactions with the Free Fall activities. None of these students were part of the original data set used to construct the models. Students attended three different schools in suburban Central Massachusetts. All had prior experience conducting inquiry in Inq-ITS for topics other than Free Fall. Students engaged in at most five different Free Fall activities. As per

the text replay tagging process, clips were distilled to cull out student actions relevant to hypothesizing and collecting data. In total, 1580 clips were generated.

Since tagging all clips would be time consuming, we selected a representative set of clips. One approach for selecting clips for the test set is to apply student-level stratification when choosing clips to code, so that each student is equally represented in the data set. We note that this is distinct from student-level cross-validation, where students are distributed to either training or test folds, e.g. (Pardos et al., 2011; Baker et al., 2008c). Equally representing all students in a test set, and using students different than those used for model construction provides more assurance that such models will work for new students. In our work, this stratification was performed as follows:

- *Student-stratified test set (291 clips)*: One clip per student was randomly selected and tagged by a human coder. Only clips in which a student ran the simulation at least once were considered for selection. One student did not appear in this set, because they had no clips with at least one run. In this set, 90.0% of the clips and 87.6% of the clips were tagged as designing controlled experiments and testing stated hypotheses, respectively.

During the clip selection process, we noticed that a disproportionate number of clips had exactly 3 simulation runs. As shown in Table 8, 70.4% of all clips distilled had 3 simulation runs, and 74.6% in the student-level test set. We looked at simulation runs because this is the grain size at which we assess students' data collection skills and activate scaffolding in the live system (Sao Pedro et al., 2012a). Though these percentages may reflect actual student behavior, it is possible that some aspects of the models' performance may not be captured by stratifying solely in terms of the student. In particular, the models' performance may be impacted by different numbers of simulation runs. To address this, we

constructed a second test set that ensures clips with a given number of simulation runs are equally represented. This stratification is described below:

- *Run-stratified test set (245 clips)*: To generate a test set that balances the number of runs per clip, we determined an optimal number of clips to have per stratum. Given the distribution in Table 1, we used runs = 5, 49 clips, as the base. We then randomly select 49 clips for each stratum with exactly 2 simulation runs, 3 runs, etc. The final stratum was for clips with more than 5 runs. As in (Sao Pedro et al., 2012a), we do not consider clips with fewer than 2 simulation runs, because demonstration of skill requires at least two trials to be collected. In this set, 93.1% of the clips and 83.3% of the clips were tagged as designing controlled experiments and testing stated hypotheses, respectively. Students' work could be represented more than once in this test set.

We note it would be more optimal to stratify over both runs and students, but too few clips would have been available for testing. In the next section, we present our models' predictive performance against these two held-out test sets.

**Table 8.** Counts of Clips Tagged for Free Fall by Number of Simulation Runs

| Simulation Runs | # Clips Distilled in Total | # Clips in Student Strat. Test Set | # Clips in Run Strat. Test Set |
|---|---|---|---|
| < 2 | 167 | 20 | 0 |
| 2 | 91 | 18 | 49 |
| 3 | 1112 | 217 | 49 |
| 4 | 102 | 15 | 49 |
| 5 | 49 | 10 | 49 |
| > 5 | 59 | 11 | 49 |
| Total: | 1580 | 291 | 245 |

### 4.3.2 Results: Estimating the Generalizability of the Detectors

We estimate how well the two inquiry skill assessment models built for one science topic, Phase Change, can predict skill demonstration for another topic, Free Fall, and a new sample of students. Generalizability is estimated by measuring how well the models predict skill demonstration in two held-out test sets containing clips pertaining to Free Fall activities. In the first test set, clips were randomly chosen via student-level stratification. Given our interest in understanding how well the models work at finer grain-sizes (Sao Pedro et al., 2012a) and the earlier finding that clips with exactly 3 simulation runs were over-represented, we constructed a second test set. This set had clips randomly chosen to ensure a balanced number of clips with a given number of simulation runs. Performance is again measured using A' and Kappa (κ). These metrics compensate for successful classification by chance which is important given the imbalance in clip labels.

#### 4.3.2.1 Student Stratification Performance

As shown in Table 9, both detectors performed quite well at predicting clips in the student stratified test set. The designing controlled experiments detector could distinguish a clip in which skill was demonstrated from a clip which it was not at a rate of A' = 90%. It also highly agreed with the human coder's ratings, $\kappa = .65$. Performance for the testing stated hypotheses detector was also high, A' = .91, $\kappa = .62$. These findings imply that the detectors built for Phase Change generalize to another Physical Science topic, Free Fall, and to an entirely new student sample, under student-level stratification.

Recall this set has exactly one randomly chosen clip per student. Furthermore, as shown in Table 9, a majority of these clips had exactly 3 runs. Though a majority of students may run exactly three trials, providing credence to being able to use the detectors as-is to assess students, the detectors'

performance may differ based on the number of trials collected. We turn next to performance on the run-level stratification test set.

**Table 9.** Overall Performance Predicting Skill Demonstration within Free Fall on the Student-Stratified Test Set

| **Designing Controlled Experiments** | | | **Testing Stated Hypotheses** | | |
|---|---|---|---|---|---|
| | True N | True Y | | True N | True Y |
| Pred N | 26 | 20 | Pred N | 21 | 7 |
| Pred Y | 3 | 242 | Pred Y | 15 | 248 |
| | Pc = .99, Rc = .92 | | | Pc = .94, Rc = .97 | |
| | A' = .90, K = .65 | | | A' = .91, K = .62 | |

* Pc = precision; Rc = recall

### 4.3.2.2 Run Stratification Performance

As shown in Table 10, the performance profile on the run stratified test set was different than on the student stratified test set. Though the performance of the testing stated hypotheses detector remained high (A'=.78, κ=.59), performance dropped for the designing controlled experiments detector, particularly for raw agreement with labels (e.g. κ) (A' = .84, κ = .26). We inspected these results more closely by recalculating the metrics for each stratum of 49 clips. As shown in the bottom of Table 10, when model confidence is not taken into account (κ), the designing controlled experiments detector had very low agreement with human labels for all run-levels (κ = .08 - .17) with the exception of clips with exactly 3 simulation runs (κ = 1.00). The testing stated hypotheses detector fared better on agreeing with human labels on all strata (κ = .40 - .78) except for clips with exactly 4 simulation runs (κ = .00). When model confidence is taken into account (A'), both models could distinguish clips that demonstrated skill

from those that did not fairly well on each strata. The only exception was the designing controlled experiments detector for on clips with at least 5 simulation runs (A' >= .61).

In summary, both detectors performed well under student-level validation. However, under run-level validation, the testing stated hypotheses detector remained strong while the designing controlled experiments detector's performance suffered. In the next section, we discuss the implications of these finding on generalizability.

**Table 10.** Performance Predicting Skill Demonstration within Free Fall on the Run-Stratified Test Set

| **Designing Controlled Experiments** | | | | | **Testing Stated Hypotheses** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | True N | True Y | | | | True N | True Y | | |
| Pred N | 16 | 60 | | | Pred N | 22 | 5 | | |
| Pred Y | 1 | 168 | | | Pred Y | 19 | 199 | | |
| | Pc = .99, Rc = .74 | | | | | Pc = .91, Rc = .98 | | | |
| | A' = .84, K = .26 | | | | | A' = .78, K = .59 | | | |
| Runs | A' | K | Pc | Rc | Runs | A' | K | Pc | Rc |
| 2 | 1.00 | .08 | 1.00 | .18 | 2 | .84 | .71 | .89 | .94 |
| 3 | 1.00 | 1.00 | 1.00 | 1.00 | 3 | .88 | .40 | .91 | .98 |
| 4 | $ | .00 | 1.00 | .98 | 4 | .84 | .00 | .90 | 1.00 |
| 5 | .66 | .14 | 1.00 | .66 | 5 | .70 | .51 | .89 | .98 |
| >5 | .61 | .17 | .97 | .76 | >5 | .79 | .78 | .98 | .98 |

* Pc = precision; Rc = recall
$ = A' could not be computed because only one class label was present

### 4.3.3 Discussion and Summary on Applying the Detectors to Free Fall

We investigated whether data-mined models that assess two inquiry process skills for activities in one science topic (Sao Pedro et al., 2012a) could be reused as-is for assessing those same skills for a new

topic and new student sample. To explore this, we collected a new set of student interactions for the topic, employed text replay tagging (Baker et al., 2006; Sao Pedro et al., 2013b), in which student interactions (clips) were labeled by humans with skill demonstration, and measured our models' ability to predict those labels. The overarching goal of this process is to measure the degree to which these models can enable scalable, reliable performance-based assessment of the inquiry skills as students conduct inquiry within simulations (Gobert et al., 2012).

Central to detector evaluation was choosing the clips to code that would yield good estimates of model performance, since coding all student interactions would be too laborious. One approach was to represent the new students equally in the held-out test set. We noticed that when we stratified this way there was an imbalance in clips for an important kind of student interaction indicative of skill, the number of times students ran the simulation. As such, we constructed a different held-out test set that ensured an equal representation over the number of simulation runs.

Under student-level stratification we found that the assessment models of each skill performed quite well in this new domain and new sample of students. These findings provide evidence that the models can be applied as-is to new topics without retraining (Baker et al., 2008c). Under run-level stratification, a different performance profile for the models emerged. The testing stated hypotheses assessment model still maintained high performance providing even stronger evidence of its generalizability. However, performance for the designing controlled experiments detector decreased. This model worked best for clips with exactly three simulation runs, the most prominent kind of clip; performance on other clips was poorer. Though performance was poorer, if the distribution of clips with given numbers of runs (Table 8) is representative of the student population we aim to assess, this model still can be used to assess in the new topic. This implies that the detectors can be used, as-is, for assessment in Free Fall for our later work presented in this dissertation which involve the same student sample.

This analysis showed how different kinds of stratification in such a test set can reveal limitations on the performance of data mined models. In particular, the ways in which a model will be used should be considered when considering generalizability. In our work, we aim for our models to be reusable to assess all students, trigger scaffolding (Gobert et al., 2012), and work regardless of how much data the student collected (Sao Pedro et al., 2012a). Thus it was essential for us to consider performance in the new simulation at the run-level since this is the granularity at which we aim to assess student work and provide scaffolding. Stratifying on other variables such as the total number of student actions or the specific inquiry activity in question (cf. Baker et al., 2008c; Ghazarian & Noorhosseini, 2010) may reveal other differences in performance. Considering these additional points may provide more evidence to the reusability of data-mined models in different contexts or reveal limitations in the models that can be addressed to improve performance in specific cases.

At this point, we have established that the detectors can be used to assess data collection in the Phase Change activities (Sao Pedro et al., 2012a) and the Free Fall activities. We have also shown that the Phase Change activities can be used, as is, to detect when students' data collection is off-track after a number of simulation runs (Sao Pedro et al., 2012a). We discuss the implications of these results, limitations and future work regarding the detectors in the next section.

## *4.4 Summary and General Discussion on the Data-Mined Detectors*

Despite the recognized importance of learning scientific inquiry skills (National Research Council, 1996, 2000; Kuhn, 2005a), hands-on inquiry activities are seldom used in schools because assessing authentic inquiry cannot be done reliably or easily (cf. Alonzo & Aschbacher, 2004; Gotwals & Songer, 2006). This difficulty arises in part because students may follow a large number of productive and unproductive paths as they engage in inquiry (cf. de Jong, 2006; Buckley et al., 2010). In other words, it is hard to evaluate students' inquiry because of the complexity and "ill-definedness" of what it means to

demonstrate skill in authentic tasks. Towards facilitating assessment of inquiry skills, we have presented data mining-based detectors of two inquiry skills associated with data collection, designing controlled experiments and testing stated hypotheses. This approach, which requires text replay tagging (Baker et al., 2006; Montalvo et al., 2010; Sao Pedro et al., 2010a, 2013b) as a first step and 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.

In terms of evaluating students' data collection as a whole, our final detectors built from a procedure aimed at increasing construct validity did so very well. The testing stated hypotheses detector could distinguish students' clips (e.g. a full data collection cycle) in Phase Change, the topic in which they were originally built, very well (A' = .91). It also had high agreement with human labels, $\kappa = .70$. When applying this detector to a new set of students and new topic, Free Fall, performance predicting students' clips remained high, (A' = .91, $\kappa = .62$). The implication is that this detector has a degree of generalizability to new students and topics, and can be used to automatically evaluate demonstration of this skill in both sets of activities. The designing controlled experiments detector also performed well, but not as strongly as its kin. When considering clips with at least three runs, this detector could distinguish clips very well (A' = .94) and agreed highly with human judgment ($\kappa = .75$). It also transferred well to Free Fall (A' = .90, $\kappa = .65$). The implication is that this detector can also be used to automatically evaluate demonstration of skill only if a student has collected at least three trials of data.

In terms of using these detectors to drive scaffolding, we found that by increasing construct validity, the detectors could be applied, as is, to drive scaffolding as students collect data. This is evidenced by reasonable A' and $\kappa$ values measured at a finer grain-size, the level of individual simulation runs, than the grain-size at which the detectors were built (a full data collection cycle). In

other words, the detectors can be used to trigger scaffolding sooner, rather than waiting for students to finish their data collection, without needing to re-tag and retrain detectors to work at this level. More specifically, the testing stated hypotheses detector had A' >= .84 and κ >= .37 for at least two simulation runs. This means that fail-soft interventions can be applied to scaffold students in as few as two simulation runs. The designing controlled experiments detector had A' >= .91 and κ >= .42 for at least three simulation runs. This means that fail-soft interventions can be applied to scaffold students in as few as three simulation runs. The choice of "simulation run" to use as the finer grain size was important. It was chosen, because we envisioned that the best point to scaffold students would be to do so when they try to collect data. As such, when we apply these detectors in the actual scaffolding system, they will be triggered when students run a new trial (see Section 4.2). In general, grain size and use of the detectors, whether for scaffolding (run or clip level in our domain) or for overall assessment (clip level in our domain), are both important to consider when evaluating detectors' applicability in a learning environment.

A central part of our approach to build detectors was increasing their construct validity, meaning that we aimed to generate models that made rational sense to a domain expert. This was realized through judicious feature engineering and by empirically comparing automated approaches for feature selection with a more hands-on approach requiring domain expertise. Our results showed that increasing construct validity improved detectors' predictive performance and enabled them to be reused for multiple purposes: evaluation of skill and driving scaffolding. There are some limitations to this interpretation. For example, we only compared our new, human-in-the-loop-based approach against a more simplistic (yet frequently used) automated feature selection approach, removing intercorrelated features (cf. Witten & Frank, 2005). As such we did not compare our approach to more sophisticated automated approaches (e.g. Oh, Lee, & Moon, 2004; Bernardini & Conati, 2010). In addition, we only used a single data mining algorithm to generate detectors, J48 decision trees. Different data mining algorithms may have

yielded different results regarding the impact of construct validity. Finally, we did not perform an empirical test to determine if increasing construct validity impacts the broader generalizability of detectors across students / contexts.

As mentioned, a key ingredient in improving construct validity is engineering predictive, domain-relevant features. This was performed by incorporating features that theoretically aligned to the constructs of interest, e.g. demonstration of data collection skills (Chen & Klahr, 1999; Koedinger, Suthers & Forbus, 1999; Buckley et al., 2006; McElhaney & Linn, 2010). As such, our results on predictive capability and generalizability are contingent on the initial set features engineered; there is no guarantee we computed all possible relevant features for our domain. Thus, including other types of features has the potential to improve prediction and generalizability. For example, recall that under run-level stratification (balancing clips based on the number of times students ran the simulation), we found that the detectors' performance was hindered. When looking more closely, we found that clips that were misclassified primarily fell under a branch of the decision tree with features reflecting domain complexity (the number of variables changeable by the student). One possible way to improve generalizability would be use ratio-based features (e.g. percent of pairwise controlled trials over all possible pairs of trials) instead of a raw counts (Sao Pedro et al., 2012a) for handling domain complexity. There are other classes of features we could consider as well. For example, Baker et al. (2008c) analyzed student timing and counts in relation to averages and standard deviations *across students*, whereas we only computed features *within* students. Our feature set also did not capture any temporal relationships between student actions, meaning that the order in which students performed actions was not considered. To take this into account, we could either attempt to engineer new features which do so, or use a process mining approach (cf. Köck & Paramythis, 2011; van der Aalst, 2011) or machine learning-based plan generation algorithm (Gu, Wu, Tao, Pung, & Lu, 2009).

An important factor not considered in constructing detectors was determining the amount and nature of training data needed to generate effective models. For example, though we considered the impacts of different kinds of stratification (student vs. simulation runs) on *measuring the performance* of detectors, we did not compare kinds of stratification *to select training data* to build models. To elaborate, we mentioned earlier that there was an imbalance on the number of training examples for clips with two simulation runs (see Section 4.2.1). This (in addition to feature engineering) may have caused the poorer performance of the designing controlled experiments detector on predicting this kind of clip. Adding more of these clips and re-employing our procedure may improve its predictive capability. Similarly, we could try rebuilding our detectors by combining all data from the Phase Change and Free Fall activities, and employing student-level and/or run-level cross-validation (cf. Efron and Gong, 1983). In general, empirical work is needed to address the role of stratification to generate appropriate training data for building generalizable detectors.

Another important factor not considered was investigating the role that the data mining algorithm itself plays in generating generalizable detectors. In our analyses, we compared and constructed behavior detectors using only J48 decision trees.[3] This algorithm was chosen based on previous successes in generating behavior detectors using text replays (Baker & de Carvalho, 2008; Baker et al., 2010a). However, other data mining algorithms may have produced better predicting or more generalizable detectors. Therefore, building detectors using other machine learning algorithms will be a valuable area of future work.

Overall, our work offers several contributions towards assessing process-based scientific inquiry skills, and analyzing the efficacy and generalizability of data-mined models. Of importance, we showed that machine learning/educational data mining methods can be employed, in combination with human classifications, to produce verifiable models of low-level inquiry behavior/skills within an inquiry

[3] We also tried building detectors using stepwise regression, but found that models' performance was not as good as decision-tree based models. These results have not been published.

environment. This provides several advances over prior methods (e.g. Schunn & Anderson, 1999; Dean Jr. & Kuhn, 2006; Kuhn & Pease, 2008; McElhaney & Linn, 2008, 2010) and enables scalable computer-based performance assessments for inquiry skills; we will discuss these broader implications in more detail in Chapter 8. Second, we explored the importance of considering construct validity when selecting features. We found that selecting features taking this into account yielded better detectors than selecting features using a more atheoretical approach, by removing inter-correlated features. Third, we described a general process for validating detectors at finer grain-sizes than they were trained and built. For our domain, the finer grain-size was the level of individual simulation runs. We found that detectors with improved construct validity could correctly infer behavior at the finer grain-size. This means we can reuse the ICV detectors as is to trigger scaffolding sooner, without needing to re-tag and retrain detectors to work at this level. In general, grain size and use of the detectors, whether for scaffolding (run or clip level in our domain) or for overall assessment (clip level in our domain), are both important to consider when evaluating detectors' applicability in a learning environment. Finally, like prior work (Baker et al., 2008c; Ghazarian & Noorhosseini, 2010) we measured the transferability of models built for one task to a new task and new set of students. In our case, we applied data mining to assess students' inquiry skills within physical science simulations. Though we have increased evidence of models' generalizability, we note that the look-and-feel and task structure of the physical science activities were generally similar. For other science domains like biology, the nature of the experimentation process may differ; further research is needed to determine if our models will generalize to entirely new types of tasks and science domains (cf. Gobert et al., 2012).

Going forward, we make heavy use of these detectors for several purposes. In Chapter 5, we use the detectors (in part) as the basis for evaluating students' demonstration of skill in the Phase Change and Free Fall activities. These evaluations become observations from which we construct latent models of student understanding across the topics. The detectors also enabled us to develop automated scaffolds

to aid students as they collect their data. We discuss these extensions to the learning environment in Chapter 6. Finally, the detectors' evaluations are used in part to measure the efficacy of our scaffolding approach. The results of an efficacy study on determining the impacts of scaffolding on students' acquisition and transfer of inquiry skills is as part of Section 6.3, and Section 7.5.2. Next, we turn to construction of latent skill models that track students' progress across activities.

# 5 Estimating Skill at Data Collection: Development and Validation of the First Version of the Bayesian Knowledge Tracing (BKT) Models

Given the ability to identify demonstration of inquiry skill using our detectors, it is possible to predict students' latent proficiency at each skill as they conduct inquiry over several activities and topics. Being able to predict proficiency provides at least three key benefits. First, proficiency estimates can be leveraged to provide scaffolding to students[4]. Second, the estimates can be triangulated with other performance data, such as paper-and-pencil tests or other performance-based tasks (cf. Corbett & Anderson, 1995), to get a handle on how well knowledge and skills learned within the environment externalize and transfer to other kinds of activities. Third, such estimates can enable "discovery with models" analyses (cf. Baker & Yacef, 2009) that can shed light on the relationships between performance in the environment and scaffolding. This last point is of particular importance since we aim to utilize such proficiency estimates, in part, to determine the efficacy of providing automated scaffolding to students as they collect data (Chapter 6).

To determine skill proficiency, a summary measure is needed that combines all of the students' chances demonstrate inquiry skill. In this work, we determine the feasibility of using a Bayesian Knowledge-Tracing (BKT) framework, a classic approach for modeling learning within intelligent tutoring systems (Corbett & Anderson, 1995), for this task. BKT was chosen due to its success at estimating skill in other learning environments that provide explicit learning support (e.g. Corbett & Anderson, 1995; Koedinger & Corbett, 2006; Baker et al., 2008a,b; Ritter et al., 2009; Feng et al., 2009; Baker et al., 2010b; Pardos et al., 2010). BKT has also been shown to predict student future performance

[4] Though we have the mechanisms for doing so, this scaffolding is out of scope for the current work. Instead, we use information solely from the detectors. See Chapter 6 for more information.

on par with or better than competing algorithms[5] (Gong, Beck & Heffernan, 2010; Baker et al., 2011; Pardos et al., 2011). Finally, the BKT framework separates performance within the learning environment from latent knowledge, facilitating the prediction of performance on tasks outside the learning environment requiring the same skills (e.g. Baker et al., 2010b).

In the following sections, we describe our initial procedure for constructing and evaluating these predictive models. We begin with an initial study that determined the feasibility of using the classic BKT model described in Corbett & Anderson (1995). Later in this dissertation, we describe an approach to extend this model to incorporate scaffolding and the change of science topics.

## 5.1 Overview of Approach for Building BKT Models

We began by determining whether the Bayesian Knowledge Tracing (BKT) framework would be suitable for estimating students' proficiency at these inquiry skills over time. We did so by comparing Bayesian Knowledge-Tracing (Corbett & Anderson, 1995), a more complex model which assumes learning between practice attempts, with a simpler average-based approach which assumes no learning within the environment. The efficacy of these proficiency models was compared in two ways. First, we compared them on predicting performance within the learning environment, providing a measure of the internal reliability of these estimates. Second, we compared these proficiency models on predicting performance on transfer tasks requiring inquiry skill. These tasks included a paper-style test of inquiry and a "hands-on" assessment in another domain. This enabled us to get a benchmark on the skill estimates' external validity.

This work also enabled us to study the relationship between standardized-test style questions and more hands-on inquiry. Though it has been argued that performance assessments are better suited than

[5] We are not asserting that BKT always outperforms other methods. For example, some alternatives have been shown to outperform BKT (e.g. Pavlik et al., 2009).

standardized-test style questions to assess inquiry skills (cf. Black, 1999; Pellegrino, 2001), rote paper tests are still typically used for assessing inquiry skills (cf. Alonzo & Aschbacher, 2004). Hence, the relationship between the two forms of assessment must be understood if our environment is to be used, in part, as an assessment tool (Gobert et al., 2012).

## 5.2 Data Set for Identifying Skill Demonstration within the Inquiry Activities

The skill demonstration detectors described in Chapter 4 have the capability to distinguish students who design controlled experiments and test their stated hypotheses from those who do not. Recall from earlier that these detectors were built and validated (in part) at the level of a *clip,* a set of actions in an inquiry activity related to students' data collection. Each of these clips, then, can be considered as a *practice opportunity* in which students have a chance to "show what they know". Thus, the detectors provide a means to evaluate students' experimentation.

Building skill proficiency estimates required a full collection of evaluations for all students over all activities. In this section, we applied the original skill detectors (Section 4.1) to the student sample (Section 4.1.1) who engaged in inquiry using the first version of Phase Change (Section 3.1). As a reminder, this sample consisted of 148 eighth grade students' interactions with four Phase Change inquiry activities. The distillation of these interactions yielded 1,503 clips which can be viewed as unique evaluations of students' work.

Because the skill detectors achieved good performance in labeling clips (see Section 4.1.6), they were used to label all students' clips, not just those clips hand-coded by humans. The resulting collection of assessments contained different numbers of clips per student, because students could engage in data collection varying numbers of times within each activity (see Figure 1). In some cases though, students transitioned between inquiry phases, causing new clips to be generated, but performed no actions within

those phases. Clips of this nature were pruned from the collection of assessments since these events did not reflect students' skills or lack thereof. Each of the remaining clips was treated as a practice opportunity.

These assessed practice opportunities served as the basis for producing our two proficiency measures, estimates of how well the student knew each data collection skill. We describe how we computed these estimates below.

## 5.3 Average-Based Proficiency Estimate

The first proficiency estimate, an average-based approach, assumed no learning occurred between practice attempts. In this approach, we averaged the number of clips in the practice opportunity corpus that positively demonstrated each skill. For example, if a student engaged in 12 data collection activities (clips) and was labeled as designing controlled experiments in 5 out of 12 clips and testing their hypotheses in 8 out of 12 clips, they received .42 and .75, for each respective skill proficiency estimates. This approach serves as a reasonable baseline because it can be expected that students who possess the inquiry skills of interest will demonstrate them at each practice opportunity.

## 5.4 Bayesian Knowledge-Tracing Proficiency Estimate

The second proficiency estimate employed Bayesian Knowledge-Tracing (BKT) (Corbett & Anderson, 1995). BKT is a two-state Hidden Markov Model (a simple Bayes Net (Reye, 2004)), that estimates the probability a student possesses latent skill ($L_n$) after $n$ observable practice opportunities. To concretize this for our work, the observable student performance is whether or not a student demonstrates one of the data collection behaviors. This is determined using the skill 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.

As shown in Figure 12, BKT models are characterized by four parameters, $G$, $S$, $L_0$, and $T$, used in part to estimate latent skill ($L_n$) based on observations of student work ($Performance_n$). We discuss the meaning of these parameters below. At each practice opportunity, a student may either demonstrate ($Performance_n = 1$) or fail to demonstrate ($Performance_n = 0$) inquiry skill. Recall that $Performance_n$ is determined by the detector's evaluation of students' data collection in a clip. Similarly, the BKT model assumes that students can either know a skill ($L_n = 1$) or not know a skill ($L_n = 0$). Thus, there is a clear distinction between latent knowledge ("knowing" how to do something) versus the observable of performance (actually doing it). A student who does not know a skill generally will fail to demonstrate skill at inquiry. But, there is a certain probability ($G$, the Guess parameter) that the student will demonstrate skill, despite not knowing it. Correspondingly, a student who knows a skill generally will demonstrate it, but there is a certain probability ($S$, the Slip parameter) that they will not succeed in demonstrating the skill. At the beginning, each student has an initial probability $L_0$ of knowing each skill, and at each practice opportunity, the student has a certain probability $T$ of learning the skill, irrespective of whether or not they demonstrated it. These four parameters provide a means for both estimating latent knowledge and predicting whether or not students will demonstrate skill.

The classic BKT framework also carries other important assumptions. As already mentioned, latent knowledge is assumed to be binary as is demonstration of skill. Second, the four parameters are assumed to be the same for all students. Third, it is assumed that students do not forget their knowledge once it is learned. Finally, the model assumes that all skills are independent and that one BKT model is fit per skill.

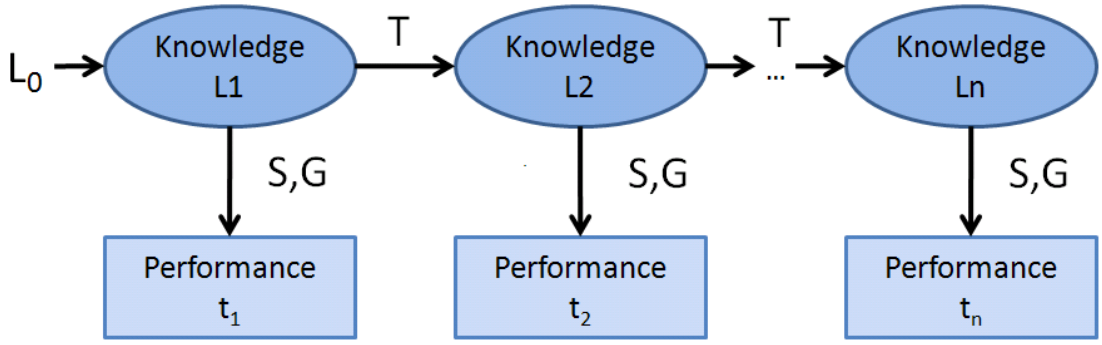


**Figure 12.** Classic Bayesian Knowledge Tracing model (Corbett & Anderson, 1995) for a skill, e.g., knowing how to design controlled experiments. The model estimates the likelihood the student knows a skill ($L_n$) after $n$ observable practice opportunities. It does so using four parameters: $L_0$ 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.

Using these four parameters, the probability that a student knows the skill, $P(L_n)$ and the estimate that a student will demonstrate that skill in their next practice opportunity $P(Demonstrate\_Skill_n)$ can be computed. The equations for these calculations are:

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

where

$$P(L_{n-1}|Clip_n) = \begin{cases} \dfrac{P(L_{n-1}) * \big(1 - P(S)\big)}{P(L_{n-1}) * \big(1 - P(S)\big) + \big(1 - P(L_{n-1})\big) * P(G)}, Demonstrated\_Skill?\,(Clip_n) \\ \dfrac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + \big(1 - P(L_{n-1})\big) * (1 - P(G))}, \sim Demonstrated\_Skill?\,(Clip_n) \end{cases}$$

and

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

In the equations above, we compute the estimate that the student knows the skill after a practice attempt at time $n$, $P(L_n)$, by first applying Bayes' Theorem. We do so to calculate the probability that the student knew the skill beforehand using the evidence from the current clip, $P(L_{n-1}|Clip_n)$. Then, we take into account the possibility that the student learned it in during this practice opportunity, $T$. Note that in the $P(L_n)$ equation, $P(L_{n-1}|Clip_n)$ is replaced by one of the two equations depending on whether or the skill

was demonstrated in clip *n*. In addition, we can also estimate the likelihood that a student will demonstrate the behavior $P(Demonstrate\_Skill_n)$ at a practice opportunity *n,* based on the probability of previously knowing the skill, i.e. $P(L_{n-1})$, and how likely a guess or slip is for this skill. $P(Demonstrate\_Skill_n)$ is an a-priori estimate of demonstrating skill since it depends on the prior estimate of knowing the skill, $P(L_{n-1})$.

An example application of these equations for a students' performance profile associated with the designing controlled experiments skill is shown in Table 11. In this example, the student generated 9 clips over all activities (and thus engaged in 9 data collections). Each clip was labeled as demonstrating skill or not using the designing controlled experiments detector. From there, estimates for $P(L_n)$ and $P(Demonstrate\_Skill_n)$ can be found by applying the equations.

**Table 11.** Example Student Practice Profile with BKT Estimates. This student engaged in 9 data collection activities, and their final estimate of knowing this skill is P(Ln) = .999.

| Designing Ctrl'd Exp's Practice Opportunities | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Final |
|---|---|---|---|---|---|---|---|---|---|---|
| $P(L_{n-1})$ | 0.077 | 0.387 | 0.819 | 0.97 | 0.795 | 0.965 | 0.769 | 0.959 | 0.994 | 0.999 |
| P(Demonstrate_Skill) | 0.191 | 0.429 | 0.761 | 0.877 | 0.742 | 0.873 | 0.723 | 0.869 | 0.895 | |
| Observable: Demonstrated Skill? | Yes | Yes | Yes | No | Yes | No | Yes | Yes | Yes | |

BKT Model: $\{L_0 = .077, G = .132, S = .100, T = .038\}$

The parameters for the BKT models for each skill can be learned from student performance data. In our work, we learned values for the four parameters ($L_0$, $G$, $S$, $T$) as follows. After applying the detectors to all clips to generate a set of observables, we used a brute force search to find the best fitting estimates of the four parameters over these data. In this approach, all potential parameter combinations of values at a grain-size of 0.01 were tried for each skill, i.e. $Params = (L_0, T, G, S) \in \{(0.01, 0.01, 0.01, 0.01), (0.01, 0.01, 0.01, 0.02), (0.01, 0.01, 0.01, 0.03), \ldots, (0.01, 0.01, 0.02, 0.01), \ldots, (0.99, 0.99, 0.30,$

0.30)}. We chose the parameter set yielding the lowest sum of squares residual (*SSR*) between the likelihood of showing a behavior in clip *n*, $P(Demonstrate\_Skill_n)$, and the actual data from clip *n*, $Observed\_Demonstration_n$. In other words, we computed

$$SSR = ( Observed\_Demonstration_n - P(Demonstrate\_Skill_n) )^2$$

for all clips, summed these values, and chose the parameter set yielding the lowest sum. After finding the best parameter set *Params*, we repeated the above procedure to find a tighter fit for the parameters, *Params*', by searching around a radius of .01 for each parameter at a .001 grain-size. For example, if the parameters yielding the lowest SSR on the first pass were *Params* = (.05, .03, .04, .06), we would search for tighter fitting parameters by finding the parameters *Params*' yielding the minimal SSR within the set {(0.041, 0.021, 0.031, 0.051), …, (0.059, 0.039, 0.049, 0.069)}.

We also restricted the search by bounding the values of *G* and *S* to avoid the "model degeneracy" problems that arise when performance parameter estimates rise above 0.5 (cf. Baker et al., 2008a,b). When values of these parameters go above 0.5, it is possible to get paradoxical behavior where, for instance, a student who knows a skill is more likely to get it wrong than to get it right. Thus, like Baker et al. (2008a,b), we bound *G* to be less than 0.3 and *S* to be less than 0.1; all other parameters can realize values in (0.0, 1.0).

The brute force approach was chosen since recent investigations (e.g. Ritter et al., 2009; Pardos & Heffernan, 2010) suggested that the Bayesian Knowledge-Tracing parameter space is non-convex (Boyd & Vandenberghe, 2004), meaning that using the Expectation Maximization (EM) algorithm could produce non-optimal parameters. Furthermore, the brute force approach we used has been shown to be computationally tractable for larger datasets than ours (cf. Baker, et al., 2010b). However, we note that this approach may still not guarantee optimal parameters; there has been recent debate as to whether Expectation Maximization functions better or worse than Brute-Force Search and other algorithms.

Results thus far have been inconclusive across studies (cf. Pavlik, Cen & Koedinger, 2009; Gong, Beck & Heffernan, 2010; Baker et al., 2010b, 2011).

It is also worth noting that BKT has been shown to effectively estimate skill even when data on correctness of student responses are noisy (Beck & Sison, 2006). This work showed that BKT models are effective even when observables are biased towards false negatives, as in our case since the skill detectors bias towards labeling students as not demonstrating skill (see Section 4.1.8).

Part of the goal of this research is to measure the degree to which the estimates produced by the aggregating models can predict skill measured by other tasks. We describe the tests we used to estimate models' external validity next.

## 5.5 Transfer Tests

To investigate the degree to which our automated detectors capture knowledge that transfers outside of the phase change environment studied, we developed three transfer assessment batteries. These batteries were administered after students completed the four Phase Change activities. These instruments measure students' understanding of hypotheses and designing controlled experiments (cf. National Research Council, 1996), skills aligned with the inquiry behaviors modeled in this paper. These assessments provided a way to validate our proficiency estimate models of skill within the phase change environment (cf. Corbett & Anderson, 1995; Beck & Mostow, 2005). They also allowed us to study the relationships between our measures of authentic inquiry performance and more traditional measures of inquiry knowledge (cf. Black, 1999; Pellegrino, 2001).

Two assessments utilized multiple choice items, an approach involving items similar to those seen in standardized paper tests of inquiry (cf. Alonzo & Aschbacher, 2004). These items came from several sources: our team, an inquiry battery developed by a middle school science teacher, and an assessment battery on designing controlled experiments developed by Strand-Cary and Klahr (2009).

Items were chosen to be as domain-neutral as possible. The first multiple-choice assessment contained 6 items and measured understanding of hypotheses. These items required students to identify independent (manipulated) variables, dependent (outcome) variables, and a testable hypothesis for different cover stories. The second multiple-choice assessment contained 4 items and measured understanding of controlled experiments. The first item required students to identify the Control of Variables Strategy procedure (cf. Chen & Klahr, 1999). The remaining three items required students to identify the appropriate controlled experiment that makes it possible to test a specific variable's effects on an outcome.

Our third assessment, the ramp transfer test (Figure 13), was designed specifically to measure students' authentic skill at designing controlled experiments. This assessment required students to construct four unconfounded experiments within a different domain, a ramp microworld on determining which factors (ramp surface, ball type, ramp steepness, and ball start position) would make a ball roll further down a ramp (Sao Pedro et al., 2009; Sao Pedro et al., 2010b). For each item, two ramp apparatuses in an initially confounded setup were presented. Students were asked to change the ramp setups in order to test the effects of a given target variable (e.g. ramp surface). A setup was evaluated as correct if they correctly contrasted the target variable while keeping all other extraneous variables the same. Note that this performance assessment focuses exclusively on the design of controlled experiments and does not include full inquiry cycles like the Phase Change activities.

Though the ramp transfer test and multiple choice tests on designing controlled experiments attempt to measure the same skill, their formats are quite different. Therefore, we did not combine scores from the two tests to form a single measure of the skill. This choice also enabled us to analyze if authentic inquiry skill in one domain predicts skill in another domain (the ramp environment) separately from our analysis predicting performance at answering multiple choice questions involving that same skill.

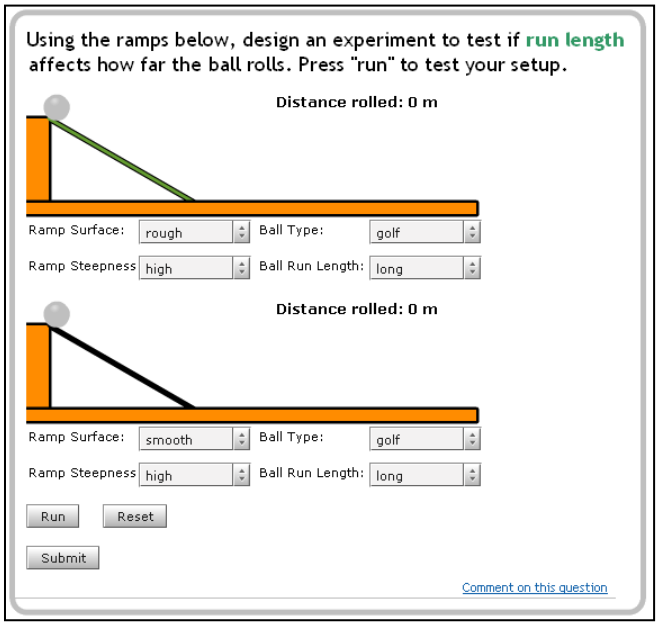


**Figure 13.** Example ramp transfer test question (Sao Pedro et al., 2009, 2010b). Students constructed experiments to determine if any of four dichotomous independent variables: *surface*, *ball type*, *steepness*, and *run length* affected how far a ball rolls down the ramps. Initially, ramp setups could be *unconfounded* (all variables are controlled), *singly confounded* (one variable is not controlled), *multiply confounded* (more than one variable is not controlled), and/or *uncontrasted* (the target variable is unchanged). The setup shown above is uncontrasted and singly confounded because the target variable, run length, is the same for each setup and one extraneous variable, surface, is not controlled.

## *5.6 Results: Validating the Proficiency Estimation Models*

### 5.6.1 Internal Validation: Comparison of Models in Predicting Inquiry Skill Demonstration within the Phase Change Activities

We compared the two estimation models for each skill on how well they could predict if students would demonstrate that skill for a given practice attempt ($Clip_n$) within the phase change environment. To do this, each detector was first used to classify the entire data set, including data not labeled by the human coders. Then, each estimation model was applied to the entire data set to produce proficiency estimates at each practice opportunity. For the average-based model, we estimated skill at the time of $Clip_n$ by computing the average number of times the student demonstrated skills in prior attempts ($Avg_{n-1}$). For the BKT model, we computed $P(Demonstrate\_skill_n)$, the probability that the student would demonstrate

the skill in $Clip_n$, which was based on having previously learned the skill, and on the guess and slip parameters. In these comparisons, we omitted all first practice opportunities from model comparison, as the average-based model is undefined in this context.

Model goodness was determined by computing A' because it is an appropriate metric when the predicted value is binary (skill demonstrated in $Clip_n$ or not), and the predictors for each model are numerical (probabilities of demonstrating skill). However, unlike Baker et al. (2010a, 2011) which computed A' values per student and aggregated them into a meta A', we collapsed over the student parameter. This was done because there was not enough within variance per skill for each student to produce a meaningful within-student A'. We then compared models based on their A' goodness measures.

For each skill, the average-based and BKT models fit comparably well to student performance. There was at most a 2% difference between the two models for each skill: designing controlled experiments (A'=.74 for both), and testing hypotheses (BKT model A'=.79 vs. Avg model A'=.77). Typically, the difference between models can be statistically tested by performing a Z-test comparing A' values (cf. Fogarty, Baker, & Hudson, 2005). However, we could not perform this analysis since we collapsed over students to compute A' values, making all observations non-independent. In practice, the differences between the proficiency estimate models for each skill are small. Additionally, the A' values are reasonably high in all cases, suggesting that either approach would be adequate in estimating proficiency.

Thus, both models perform acceptably at predicting the success of a student's current practice attempt within the phase change environment. using aggregated information from previous practice attempts. Furthermore, these findings suggest that the behaviors represent a latent skill which is stable over time. Surprisingly, though, the BKT model did not outperform the average-based model. We discuss possible reasons why the BKT model did no better than the average-based model in Section 5.7.

### 5.6.2 External Validation: Comparison of Models in Predicting Transfer Test Performance

The proficiency estimate models also enabled us to examine if inquiry skills demonstrated in the phase change environment predicted performance on the transfer tests. This also provides a measure of external validity of these models. As described in Section 5.5, students completed three transfer tests: a multiple-choice test of inquiry on designing controlled experiments, a hands-on test of designing controlled experiments to determine what factors make a ball roll further down a ramp, and a multiple-choice test on testing hypotheses. In our analyses, we determined the degree to which a particular skill demonstrated within the phase change environment predicted performance on the transfer test measuring the same skill. For example, performance on the designing controlled experiments transfer tests was predicted using the two proficiency models' final estimates of that skill in the phase change environment. We computed these final skill estimates as follows. The final skill estimate for the average-based models was computed as the average number of times the student demonstrated skill over all clips. For the BKT models, we used the estimate of student knowledge, $P(L_n)$, at the last clip. We used $P(L_n)$ rather than $P(Demonstrate\_Skill_n)$ for the BKT estimate because the guess and slip parameters used within the microworld may not apply to our transfer tests.

In these analyses, we considered only students who completed both the phase change microworld activities and the inquiry and ramp transfer post-tests. This resulted in 134 students. Prediction strength was determined by computing Pearson correlations ($r$), and checking if they were significantly different from zero (no positive or negative correlation) at the $p < .05$ level. Means and standard deviations for each test and each model's final skill estimates are shown in Table 12.

Both the BKT model and average-based model significantly predicted transfer test performance, with the average-based model predicting performance equally well or better than the BKT model for all tests, as shown in Table 13. For the designing controlled experiments skill, both models achieved relatively modest but statistically significant correlations to each of the transfer tests. The two models each achieved a correlation of $r =$. 26, $p = .003$ to the multiple-choice test. They also achieved similar correlations with the hands-on activity, the ramp environment, $r = .38$, $p < .001$ for the average-based model, and $r = .37$, $p < .001$ for the BKT model.

For the skill of testing stated hypotheses, the average-based model achieved better predictive performance on this skill's corresponding multiple-choice transfer test ($r = .41$, $p < .001$) than the corresponding BKT model ($r = .31$, $p < .001$). The difference between correlations was statistically significant, as determined by a significance test of the difference between two correlation coefficients with correlated samples (Ferguson, 1976, pp.171-172), $t(131) = 2.50$, $p = .01$. This suggests that the average-based model yielded a more valid estimate of this skill.

Given that the models significantly predicted performance on both the multiple choice and authentic transfer measures of designing controlled experiments, we examined next whether the models were better predictors for one type of test format than the other.

**Table 12.** Means and Standard Deviations for estimates of inquiry skill and posttest measures, $N = 134$.

| | *Max* | *M* | *SD* |
|---|---|---|---|
| Average-Based Models | | | |
| Controlled Experiments Average | | 0.25 | 0.26 |
| Testing Hypotheses Average | | 0.30 | 0.29 |
| BKT Models | | | |
| Controlled Experiments BKT Estimate | | 0.32 | 0.38 |
| Testing Hypotheses BKT Estimate | | 0.43 | 0.41 |
| Dependent Measures | | | |
| Inquiry Posttest: Testing Hypotheses | 6 | 2.06 | 1.53 |
| Inquiry Posttest: Controlled Experiments | 4 | 2.09 | 1.18 |
| Ramp Transfer: Controlled Experiments | 4 | 1.59 | 1.70 |

**Table 13.** Correlations between posttest measures and each model's estimate of inquiry skill, $N = 134$.

| | Avg-Based Model *r* | BKT Model *r* | *t* Difference |
|---|---|---|---|
| Dependent Measures | | | |
| Inquiry Posttest: Testing Hypotheses | .41*** | .31*** | 2.50* |
| Inquiry Posttest: Controlled Experiments | .26** | .26** | 0.03 |
| Ramp Transfer: Controlled Experiments | .38*** | .37*** | 0.21 |

Note: The *t* difference between model correlations was computed using a significance of the difference between two correlation coefficients for correlated samples (Ferguson, 1976, pp.171-172).
*$p < .05$; **$p < .01$; *** $p < .001$

### 5.6.3 Comparing the Hands-on and Multiple Choice Transfer Assessments

Here, we address whether a performance assessment better captured inquiry skill than a multiple choice test, a topic of debate within the assessment and science education communities (cf. Black, 1999; Pellegrino, 2001). Recall that both models appeared to have stronger correlations with the ramp transfer test than the multiple-choice measure of designing controlled experiments (Avg-based $r = .26$ for the multiple-choice measure vs. $r = .38$ for the authentic measure, and BKT $r = .26$ for the multiple-choice

measure vs. $r$ = .37 for the authentic measure). Though the correlation appeared to be stronger for both models for the ramp task, neither difference was statistically significant, $t(131)=1.38$, $p=.17$ for the average-based model and $t(131)=1.30$, $p=.20$ for the BKT model. We discuss possible reasons why in the following discussion section.

## 5.7 Discussion on First Version of Proficiency Estimation Models

Since our skill models achieved acceptable classification of inquiry skill demonstration (see Section 4.1.8), comparable to other detectors of other behaviors used in effective interventions (e.g. Baker et al., 2008c; Baker & de Carvalho, 2008), we leveraged them to classify all students' inquiry behaviors, including unlabeled data. This permitted us to generate and compare two estimates of students' authentic inquiry skill proficiency. We compared two approaches for generating these estimates, an average-based method that assumes no learning, and Bayesian Knowledge-Tracing (BKT) (Corbett & Anderson, 1995), in terms of their ability to accurately estimate student skill at each practice opportunity. Both approaches estimated students' skills at each practice opportunity acceptably, as indicated by A' values in the .74 - .79 range. Given this performance, these models could be used to provide formative assessment feedback to teachers on inquiry proficiency for individual students or entire classes.

The models of skill proficiency were also used to study the relationship between skill demonstrated in the phase change environment, and performance on the multiple-choice and authentic transfer measures of inquiry skill. Overall, each model of authentic inquiry skill was significantly, albeit modestly, correlated to its corresponding posttest, i.e. the testing hypothesis behavior correlated with the standardized-test style questions on hypotheses. This provides some external validation of the skill estimates derived from performance within the phase change environment. These transfer findings also support the notion that authentic inquiry skills are not necessarily tied to the domain in which the skills were learned.

We also leveraged the proficiency models to determine if authentic skill at designing controlled experiments in the phase change microworld was assessed more accurately with a hands-on performance assessment than with multiple choice questions. This issue is important to the assessment and science education communities since it is unclear whether multiple-choice tests can adequately measure inquiry (Black, 1999; Pellegrino, 2001). We found no significant differences between the two sets of correlations, namely the correlations between each authentic skill estimate and the hands-on assessment, and the correlations between each estimate and the multiple choice assessment. Therefore, we could not determine whether one transfer test better measured latent inquiry skill than the other. The similarity between correlations may be because the ramp and multiple choice assessments are both domain-neutral and isolate the assessment of a single skill. The phase change microworld, on the other hand, is complex and domain-rich, meaning that domain knowledge may influence inquiry performance (Glaser et al., 1991; Schauble, Klopfer & Raghavan, 1991). A further study which uses two domain-rich measures of inquiry skill (e.g. state change and another physical science domain) could help disentangle this issue. We do note, though, our results are in accordance with earlier findings that multiple choice measures are poor assessments of inquiry performance (Black, 1999; Pellegrino, 2001), since, as previously mentioned, the correlation between authentic skill and the multiple choice test was modest.

Surprisingly, the predictive strength of the average-based estimates of skill was as good as the BKT estimates, and in the case of the multiple-choice test for testing hypotheses, the average-based model was significantly better than the BKT estimate. This finding may have several potential interpretations. First, the data set used here was quite small. It may be that more student data, both in terms of number of students and practice attempts, is required to obtain BKT models, due to the greater number of parameters. Another possibility is that we found relatively little learning occurring in our phase change environment, and Bayesian Knowledge-Tracing assumes that students' skills improve with each practice opportunity, on the whole. This is evidenced by the low learning rate for the best-fitting

BKT models (the learning parameter *T* equaled 0.038 and 0.058 for designing controlled experiments and testing hypotheses, respectively). These findings, though, are not surprising since our environment provided no explicit learning support, and it has been found that students need more time and repeated practice to develop data collection skills in the absence feedback (Dean Jr. & Kuhn, 2007). Thus, BKT may be more appropriate for skill assessment in environments that provide explicit learning support, the context where it is typically used (e.g. Corbett & Anderson, 1995; Koedinger & Corbett, 2006; Baker et al., 2008a,b; Ritter et al., 2009; Feng et al., 2009; Baker et al., 2010b; Pardos et al., 2010). BKT may become more effective for our science inquiry microworlds once we have added explicit scaffolding. We do note, though, another model may be more appropriate than both BKT and the average-based approach. More empirical evidence is needed to test this.

Since building the initial version of the BKT models, we changed the nature of the inquiry activities, developed better detectors, incorporated scaffolding, and collected more data with new students for other inquiry topics (e.g. Free Fall). All of these factors may impact how well BKT can estimate skill proficiency in the learning environment. In the next chapter, we describe our approach to incorporate scaffolding into the system and describe an analysis to address whether that scaffolding helps students acquire and transfer the data collection skills of interest. Part of this analytical work will involve extending the BKT framework to handle some of these changes. These extended models are discussed in detail in Chapter 7.

# 6 The Impacts of Scaffolding on the Acquisition and Transfer of Data Collection Inquiry Skills

Science educators and reseachers agree that cultivating inquiry skills is critical for students to become scientifically literate (e.g. NRC, 1996, 2000, 2011; Kuhn, 2005a) and to be well-poised for the demands of the knowledge-based economy of the 21st century (Clarke-Midura, 2011). As such, it is of importance to understand how best to foster acquisition and transfer of such skills to novel tasks, particularly since students typically have difficulty with inquiry (de Jong & van Joolingen, 1998).

Regarding acquisition, there is an open debate whether students should discover inquiry principals on their own or be provided explicit instruction (cf. Kirschner, Sweller & Clark, 2006; Hmelo-Silver, Duncan & Chinn, 2007) to promote learning and transfer. For example, if inquiry tasks are too open-ended, students often become lost and frustrated, which can further lead to many false starts (Schauble, 1990), misconceptions (Brown & Campione, 1994), and failure to learn the science principles being targeted (Kirschner, Sweller, & Clark, 2006). As a result, teachers spend considerable time scaffolding students' procedural skills (Aulls, 2002) making it difficult to tailor to individual students in real-time within a classroom setting (Fadel, Honey, & Pasnick, 2007).

Regarding data collection skills, some studies have shown that explicitly teaching the Control of Variables Strategy (cf. Chen & Klahr, 1999) can lead to successful acquisition (Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008; Zohar & David, 2008; Sao Pedro et al., 2009), retention (Sao Pedro et al., 2010b), and transfer to novel tasks (Klahr & Nigam, 2004) over pure discovery methods. On the other hand, evidence exists that long-term, repeated practice of these skills can also lead to acquisition (Kuhn, Schauble & Garcia-Mila, 1992; Dean Jr. & Kuhn, 2006; Kuhn & Pease, 2008). Finally, there are "middle ground" approaches in which students have shown learning gains when engaging in structured (scaffolded), but still open-ended hands-on and computerized activities (e.g. White & Frederiksen, 1998).

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). As such, learning inquiry skills in one context (science topic) may not guarantee that skills will be demonstrated in other topics (Thorndike & Woodworth, 1901; Singley & Anderson, 1989). However, several researchers have provided evidence that this is not the case. For example, Glaser et al. (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.

In the present work, we seek to improve both acquisition and transfer of inquiry skills by providing automatic scaffolding to students in real-time as they experiment within a computer-based simulation inquiry learning environment. To do so, we leverage automatic evaluation of inquiry skills, and estimation of latent skill across activities—to provide automated scaffolding to students as they engage in inquiry, and to evaluate the efficacy of the scaffolding. Efficacy is determined by measuring the impact that scaffolding has on the acquisition and transfer of the inquiry (data collection) skills

across two physical science topics, Phase Change and Free Fall. As in White & Fredricksen (1998), scaffolding may strike a balance between direct instruction and discovery learning by providing help only when students need it (cf. Vygotsky, 1978; Wood & Wood, 1996) which may enable them to master and transfer these skills to new domains. Unlike White & Frederiksen (1998), we provide explicit support by giving individualized feedback as students collect data in this environment.

To that end, we address two research questions regarding the efficacy of scaffolding data collection, and its impact on inquiry performance across two domains. These questions are:

1. Does scaffolding data collection within a domain improve students' skill compared to a condition with no data collection scaffolding in the same domain in which skills are learned/practiced? We hypothesize that scaffolding students' data collection skill will result in better usage of these skills as students engage in more practice opportunities within the same domain.

2. Will scaffolding of data collection skill in one domain give students a "leg up" in authentic data collection skill in a second domain when the scaffolds are removed? We hypothesize that honing inquiry skills within phase change will lead to increased and correct use of these skills within the kinematics domain. It may be, though, that as tasks become more difficult (i.e. the topic changes to an unfamiliar topic), learners become novices relative to the task (Bereiter & Scardamalia, 1993). Therefore, in terms of this work, it might mean that as students switch content topics, they become less systematic (cf. van Joolingen et al., 2007).

For the scope of this dissertation, the scaffolding was implemented to support students' data collection within the second version of the Phase Change activities (see Section 3.2), though we aimed to develop scaffolds that were general enough to be used for other physical science topics. We begin by describing

the extensions to Inq-ITS made to incorporate scaffolding. Then, we describe how we used it to define different pedagogical tutorial tactics (Chi, VanLehn, & Litman, 2010), depending on student responses. Finally, we describe a study in which we determine the efficacy of our approach and present our findings.

## 6.1 Extending the Learning Environment to Incorporate Scaffolding

The Inq-ITS system delivers scaffolds and hints to students in a text-based format via a pedagogical agent named Rex, a cartoon dinosaur, shown in Figure 14. Primarily, Rex provides real-time feedback to students as they engage in inquiry. In other words, the system can "jump in" and provide support to students as they work, thus initiating support on its own. Scaffolds for each skill involves multiple levels, with each level providing a more specific hint, similar to Cognitive Tutors (e.g. Anderson et al., 1995; Corbett & Anderson, 1995; Koedinger & Corbett, 2006). Thus, if students continue to struggle, more directed help is provided to the student. Determination of who receives scaffolding is performed using both EDM-based detectors and knowledge engineered rules (Gobert et al., 2012; Sao Pedro et al., 2012a, 2013a, 2013b). We will elaborate on the details of this approach for specific skills and known haphazard inquiry behaviors in the upcoming subsections.

This automated approach using a pedagogical agent was chosen for several reasons. First, prior work suggests that in general, students have difficulty engaging in inquiry without support (Schauble, 1990; de Jong, 2006) and monitoring their progress while doing so (de Jong et al., 2005). Without support, students may flounder or engage in unproductive, haphazard inquiry behaviors (Gobert, Buckley & Horwitz, 2006; Gobert & Schunn, 2007). Second, students may lack the appropriate metacognitive help-seeking skills to recognize when they should ask for help (Aleven & Koedinger, 2000; Aleven et al., 2004). Thus, reacting when students appear to be off-track may benefit students' inquiry. A pedagogical agent was chosen specifically, because they have been shown to benefit learners

(cf., Rickel & Johnson, 1999; Moreno, 2005). They may improve learning indirectly by increasing students' engagement and motivation (Lester et al., 1997; Walker et al., 1994; Dehn & van Mulken, 2000; Lester et al., 1999; Mitrovic & Suraweera, 2000). They may also make learning more interactive and the just-in-time scaffolding they afford can improve learning (Mayer et al., 2003). Finally, providing students with short explanations at just the right time may support learning (Mayer, 2004; Rieber, et al., 2004).

Though Rex can automatically respond to students as they experiment, some on-demand help (Aleven, Stahl, Schworm, Fischer, & Wallace, 2003) that can be initiated by students is also available. Students can ask Rex for more clarification after he gives advice to ask for more help. For example, after Rex "jumps in" and tells the student they are not testing their hypotheses as shown in Figure 14, the student can ask Rex "How do I do that?" by clicking on the button. Students can also play back any hints or scaffolds Rex already gave by clicking on the "Show what I said" button under Rex, also shown in Figure 14.

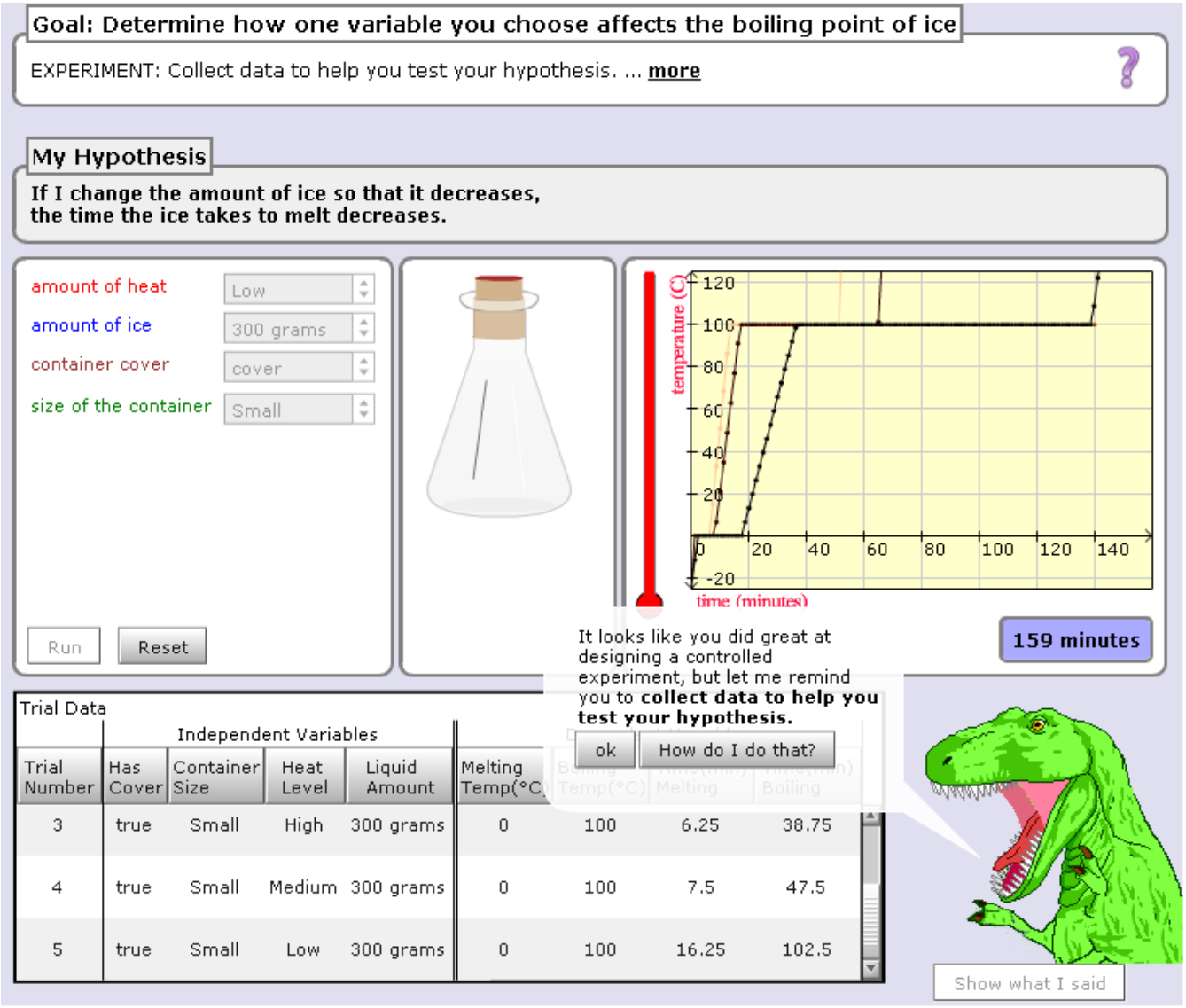


**Figure 14.** Phase Change activity highlighting the pedagogical agent Rex. Here, Rex has responded to a student who appears to be designing controlled experiments, but is not testing their hypothesis. They can continue experimenting or ask Rex for more help, in this case by clicking "How do I do that?"

### 6.1.1 Integrating Evaluations of Students' Data Collection and Scaffolding

Underlying Rex is a flexible scaffolding system that enables authoring of many kinds of scaffolds. Briefly, the scaffolding system was designed to be data-driven, and separate from the software evaluating student skill. This software design approach has three advantages: (1) the scaffolding and evaluation could be developed and tested separately, (2) scaffolding paths can be re-

used between inquiry activities, and (3) it becomes easier to manage, evaluate, and compare different kinds of scaffolding implementations. To enable this kind of flexibility, we implemented a scaffolding framework in which scaffolds are declared using a combination of specialized XML and Lisp scripting.

At a high level, a scaffold definition is comprised of two pieces: (1) a set of constraints to determine if a scaffold should be fired, and (2) a list of text / hint possibilities Rex should give as feedback. In our system, a student receives feedback from Rex if a student takes actions that cause one of the constraints to be violated. For example, one constraint checks to see if a student did not test their stated hypotheses, but did design controlled experiments. If that constraint fires, and it is the first time it fired, the system provides a feedback message to the student shown in Figure 14. In this sense, Inq-ITS acts similarly to constraint-based tutors (cf. Ohlsson & Mitrovic, 2006) since student work is evaluated against a set of constraints and is deemed as haphazard in some fashion if at least one constraint is violated. Conversely, if no constraints fire, then a student is deemed to have demonstrated skill at collecting data.

This approach for scaffolding and evaluation was chosen for two reasons. First, we aimed to provide a degree of flexibility in letting students explore with our activities to be true to scientific inquiry. Second, from an evaluation standpoint, scientific inquiry can be considered an ill-defined domain (cf. Lynch et al., 2009) with no single "right" or "wrong" way to engage in inquiry (Shute, Glaser & Raghavan, 1989; Glaser et al., 1991). As such, we argue it would have been more difficult to specify all possible correct methods of data collection and to react to students' experimentation in a more flexible way than if we evaluated and scaffolded in a forward, step-based approach as performed in model tracing (cf. Koedinger et al., 1998).

Next, we provide some examples of the specific scaffolds we implemented to aid students' data collection. Aside from the two skills of interest we aim to measure (designing controlled experiments

and testing stated hypotheses), we also developed scaffolds for other common, well-defined haphazard behaviors. We describe these in more detail below.

### 6.1.2 Evaluating and Scaffolding Well-Known Haphazard Data Collection Behaviors

As described in Section 2.4, students have difficulty with data collection. We developed scaffolds for some of these difficulties that have been documented. These include: never changing variables while running trials, running only a single trial (Kuhn, Schauble, Garcia-Mila, 1992; Glaser et al., 1991), and running repeated trials (Kuhn, Schauble & Garcia-Mila, 1992; Buckley, Gobert & Horwitz, 2006; Buckley, et al., 2010). Because these behaviors are very well-defined and unambiguous, we authored knowledge-engineered rules to detect them, as opposed to building machine-learned models. To illustrate, a student is deemed to "never change variables" if, when collecting data, they never change any simulation variables and run the simulation three times. A student is deemed to have run only a single trial if they try to finish the experiment task (signifying they are ready to go analyze data) and they have run the simulation only once (Kuhn et al., 1992). Finally, a student is deemed as running repeated trials (Kuhn et al., 1992; Buckley et al., 2006) when they have run three of the exact same experimental setup amongst all trials.

As an example, a sequence of scaffolds associated with running repeated trials is shown in Appendix A. If a student has run three of the exact same experimental setup amongst all trials, Rex jumps in and says: "Did you know you already ran a trial using these variables? The **table** has all the results of your trials, so you don't need to run this trial again." If they persist on running repeated trials, the help becomes more specific; Rex says: "You already ran this trial. The results of it are in the table. You don't need to run it again."

We next turn to the scaffolding and evaluation approach for our two skills of interest, designing controlled experiments and testing stated hypotheses.

### 6.1.3 Evaluating and Scaffolding Designing Controlled Experiments and Testing Stated Hypotheses

In the previous section, some known procedures not conducive to productive inquiry (i.e. haphazard inquiry behaviors) were identified and modeled with knowledge engineered rules due to their straightforwardness. By contrast, we evaluate whether students design controlled experiments and test their stated hypotheses by combining predictions made by data mined detectors (Sao Pedro et al., 2012a, 2013a,b) with knowledge engineered rules to handle the specific edge cases where the detectors do not identify skill demonstration as well. To elaborate, as described in Section 4.4, the designing controlled experiments detectors work well when students have run the simulation at least three times (thus collecting three pieces of data). At this point, the detectors can distinguish a student who has designed controlled experiments when they have completed their data collection from a student who has not A' = 94% of the time. They also could identify the correct class extremely well, $\kappa = .75$. The testing stated hypotheses detector also predicted quite well, without the limitation on the number of trials collected by the student, A' = .91, $\kappa = .70$.

We also found that these detectors could also be used as-is to drive scaffolding in Phase Change, *before* students finished collecting their data (Sao Pedro et al., 2012a; Section 4.2). Given this level of goodness, these detectors could be used to drive scaffolding in Phase Change, before students finished collecting their data. The designing controlled experiments detector could successfully be applied by the student's third data collection with the simulation, and the testing stated hypotheses detector could be applied in as few as two simulation runs. Finally, these detectors have been show to generalize to assess skill within the Free Fall activities (Sao Pedro et al., 2013a), a different science topic from which they

were built (Phase Change) and an entirely different cohort of students. Under student-level stratification, the designing controlled experiments model could distinguish a student who designed controlled experiments from one who did not A' = 90% of the time, and highly agreed with a human coder's ratings, $\kappa = .65$. Performance for the testing stated hypotheses model was also high, A' = .91, $\kappa = .62$.

Though performance of these models is quite high, as mentioned, there are edge cases where the detectors did not perform as well. In particular, the designing controlled experiments detector cannot be applied when students collect only 1 or 2 pieces of data with the simulation. The testing stated hypothesis detector cannot be applied when the student collects only a single trial. In these cases, we authored simple knowledge engineered rules for a single trial (Glaser et al., 1991) as described in Section 6.1.2, and for exactly two trials since evaluating these skills for this case is also well-defined (Chen & Klahr, 1999; Koedinger, Suthers & Forbus, 1998).

While collecting data, students may exhibit skill at testing their hypotheses, designing controlled experiments, both, or neither. To account for these possibilities, we implemented different scaffolding levels for each case of not demonstrating skill as shown in Appendix A, with each level providing more specific help. For example, while collecting data if the system detects that a student is designing controlled experiments but not collecting data to test their hypothesis, Rex will tell them "It looks like you did great a designing a controlled experiment, but let me remind you to collect data to help your test your hypotheses." If students continue to struggle in this fashion, a bottom-out hint is given: "Let me help some more. Just change the [IV] and run another trial. Don't change the other variables. Doing this lets you tell for sure if changing the [IV] causes changes to the [DV]" where IV and DV are specific to the student's hypothesis.

Though our detectors / knowledge engineered rules evaluate whether students are engaging in good data collection, they do not encapsulate *when* scaffolding should be triggered. The scaffolding policy we implemented (cf. Chi, VanLehn, & Litman, 2010) to determine when to trigger scaffolding

works as follows. If a student signals they are finished collecting data by opting to go to the "analyze data" task, their actions are evaluated by the system. When this occurs, if students only ran a single trial or never changed any variables, they are given that scaffolding sequence (see Section 6.1.2) and are evaluated as not having designed controlled experiments and not testing stated hypotheses. If they ran more than one trial, the data mined detectors / knowledge engineered rules are used to evaluate whether students demonstrated the skills. If they have not successfully demonstrated either skill (perhaps neither skill), they are provided with the first scaffolding hint in the sequence (Appendix A) and not permitted to analyze data until they have demonstrated the skills. Once scaffolding is triggered, students may receive more help automatically after they have run the simulation two more times. In other words, the system provides some leeway to let students continue experimenting before providing more targeted support. During this time, as mentioned, students can also ask Rex questions (e.g. how do I design controlled experiments?) for more support. The questions students can ask in response to specific scaffolds are shown in Appendix A.

Scaffolding can also be triggered automatically as students work after they run the simulation two times. On the third time running the simulation, students' actions are evaluated each time they run the simulation, e.g. collect data, to determine if they are off-track. Recall that our detectors are capable of discerning this (see Section 4.2.5.2). If students are deemed to be off-track at designing controlled experiments and/or testing their hypotheses, scaffolding is triggered and the same process applies as describe previously for providing more targeted help.

A goal of this work is to determine if this scaffolding approach is effective in that it helps students acquire and transfer skills across science topics. We now describe a study and analytical approach for determining this.

## *6.2 Method*

### 6.2.1 Participants

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

### 6.2.2 Materials

All materials were administered using the Inq-ITS System. We developed two sets of inquiry-based activities using the Phase Change (Section 3.2) and Free Fall (Section 3.3) inquiry activity templates. As a reminder, the inquiry activities have a similar look-and-feel and structure (Gobert et al., 2012). Students are given explicit goals to conduct investigations with the simulation by formulating hypotheses, collecting and interpreting data, and warranting their claims to address that goal. Under this scheme, we developed two sets of activities per science topic.

Five inquiry-based activities were developed for Phase Change. Three targeted specific concepts relevant to Phase Change, and two enabled students to test their own hypotheses related to Phase Change (subject to the factors they could vary and outcomes they could observe with the simulation). Students first completed the target-specific activities, and then the choose-your-own-hypothesis activities. We elaborate on these as follows:

- *Targeted Concept Inquiry Activities:* We developed three inquiry activities that required students to conduct investigations with the simulation to address a particular content-specific goal. This approach enables that a breadth of independent/dependent variables will be explored. In each activity, students were first given an explicit goal. For Phase Change, these goals were: 1) “See if the amount of heat applied affects the water's boiling point (after the ice melts to water)”; 2) “See

if the amount of ice affects the water's boiling point (after the ice melts to water)"; 3) "See if the amount of ice in the container affects the ice's melting point." After reading the goal, students then engaged in inquiry by hypothesizing, collecting data with the simulation, and analyzing the data they collected using the interface (Section 3.2; Figure 14). We note that students could choose to ignore the experimentation goal presented if they chose. Once complete, students were provided a text box in the learning environment and asked to write a sentence to summarize their experiment.

- *Choose-Your-Own-Hypothesis Inquiry Activities:* We also gave students two activities in which they were instructed to "Come up with your own hypothesis and then test it." After reading this goal, students again engaged in inquiry to investigate their own hypotheses. Once complete, students were again provided a text box and asked to "Describe your investigation and the results of your investigation. Write your answer as if you are explaining it to a friend who didn't do the activity."

Five inquiry-based activities were also developed for Free Fall, three for targeting specific concepts, and two for students to test their own hypotheses. The concept-related goals for Free Fall were: 1) "See if the ball's mass affects its mechanical energy at its lowest point"; 2) "See if the ball's starting height affects its potential energy at its highest point"; 3) "See if the ball's starting height affects its kinetic energy at its lowest point."

### 6.2.3 Procedure

Students were assigned the 5 Phase Change inquiry activities, and two weeks later, the 5 Free Fall activities. Students were allotted approximately two class periods per science topic to complete the activities. Due to time constraints, some students did not finish all the activities in each science topic. Recall that in each activity, students formulated hypotheses, collected data and analyzed data. Prior to beginning the activities, the learning environment randomly assigned students to one of two learning conditions as follows:

- Data collection (DC) scaffolding condition: In this condition, Scaffolds for hypothesizing, and experimenting (data collection) were available to students. As soon as the system detected they are engaging in unproductive inquiry, the system provided feedback via the pedagogical agent. Scaffolds for the "analyze data" task are not given since certain scaffolds were authored for that phase of inquiry which could impact students' data collection processes.
- No data collection (NDC) scaffolding condition: Scaffolds for hypothesizing will be present, but students will not receive scaffolding on data collection.

We highlight that students in the "No Data Collection (DC) scaffolding condition" still received scaffolds on hypothesizing. Furthermore, neither condition received scaffolds on analyzing. This experimental design is chosen for two reasons. First, scaffolding the hypothesis phase guarantees that students formulate a syntactically correct, testable hypothesis when they enter the experiment phase. This tries to ameliorate possible confounds of skill in which students try to collect data for a malformed hypothesis. Second, deciding *not* to scaffold the analysis phase ameliorates the possibility that students gain experimentation skill due to scaffolds in the analysis phase. This phase requires at least one pairwise controlled experiment be run to construct a valid analysis, and such scaffolds could direct the student to go back and collect more data, affecting performance in that phase.

After the randomization process occurred, students in each condition completed the first 4 Phase Change activities (3 for targeting specific Phase Change concepts, and then one activity where students were instructed to devise their own hypothesis). Both groups then completed a fifth Phase Change activity with no scaffolding in any of the inquiry tasks (hypothesizing, data collecting, analyzing data). This enabled us to measure immediate impacts of scaffolding on skill acquisition within the same science topic.

Approximately two weeks later, all students engaged in inquiry within the Free Fall activities. No scaffolding was present for any part of the inquiry process. These activities were used to determine the impacts of scaffolding on transfer of skill across domains.

## 6.3 Results

Our goal is to determine the efficacy of our scaffolding approach that helps students acquire two data collection inquiry skills, designing controlled experiments and testing stated hypotheses. The efficacy of scaffolding is addressed by analyzing two research questions. The first question addresses if providing automated scaffolding promotes acquisition of the skill within that same science topic (Phase Change). The second question addresses if providing scaffolding in the first science topic enables transfer of skill (cf. Thorndike & Woodworth, 1901; Singley & Anderson, 1989) to a second science topic (Free Fall) with scaffolding removed. We hypothesize that more students who received data collection scaffolds (DC condition) will both acquire and transfer skill than students who did not receive these scaffolds (NDC condition).

As mentioned, the detectors / knowledge engineered rules evaluate students' actions, thereby determining whether students demonstrated data collection skill. In general, this evaluation occurs when students change from the "experiment" inquiry task to the "analyze data" inquiry task, after they completed their data collection. However, in the DC scaffolding condition, students' actions are also

evaluated as they work to determine if scaffolding should be applied. Relevant to the analyses we present, if the system believes the student does not know either skill (or both skills) while they experiment and it provides scaffolding, the student is deemed as *not* demonstrating skill. Furthermore, they are considered as not having demonstrated skill for the duration of their data collection even if they eventually do demonstrate skill as a result of the scaffolding. Instead, their next full data collection is considered (e.g. in their next inquiry activity, or if they decide to re-collect data after being in the analyze data task) to determine if they demonstrate skill. This approach to disentangle immediate scaffolding effects is similar to others (e.g. Pardos & Heffernan, 2011). Using this assessment framework, we analyze our two research questions on the efficacy of scaffolding below.

### 6.3.1 Effects of Scaffolding on Data Collection Skill Acquisition within the Same Science Topic

First, we determine whether students who received scaffolding were more likely to demonstrate the skills in the final, completely unscaffolded (fifth) Phase Change activity that both learning conditions completed. In this analysis, we only consider students who had their data collection evaluated in this activity, 268 students. The data are analyzed using two approaches. In the first approach, we compared whether students demonstrated these skills in the unscaffolded Phase Change activity looking solely at the evaluation of their actions in this activity. In other words, did more students in the DC scaffolding condition demonstrate skills in the unscaffolded practice opportunity than students in the NDC condition?

As shown in Table 14, there was a significant difference in whether students designed controlled experiments ($\chi^2(1) = 8.60$, $p = .003$) and tested their stated hypotheses ($\chi^2(1) = 8.60$, $p = .003$) in this activity. In particular, 78.8% of the students in the NDC condition demonstrated both skills, whereas

91.6% in DC condition demonstrated both skills. These findings suggest that our scaffolding approach had an impact on whether students demonstrated these skills in the unscaffolded Phase Change activity.

**Table 14.** Crosstabulations of practice condition, and whether students demonstrated skill in the unscaffolded Phase Change inquiry activity, $n = 268$.

| | Designed Controlled Experiments? | | Tested Stated Hypotheses? | |
|---|---|---|---|---|
| | No | Yes | No | Yes |
| No Data Collection Scaffolding (NDC) | 29 | 108 | 29 | 108 |
| Scaffolding (DC) | 11 | 120 | 11 | 120 |
| | $\chi^2(1) = 8.60$** | | $\chi^2(1) = 8.60$** | |

**p < .01

Our second approach to evaluating scaffolding's impact is to aggregate all the practice attempts together to see if, by the end of the session, students mastered these skills. In this analysis, we aim to see if the different kinds of practice, with or without scaffolding, had differential impacts on acquisition of skill. To address this, we fit classic Bayesian Knowledge Tracing (BKT) models for each skill using students' performance data across all Phase Change activities (Corbett & Anderson, 1995) as we have done previously (Sao Pedro et al., 2012a, 2013b; see Chapter 5). In this process, we fit new BKT models because the previous models were built for use for an entirely unscaffolded version of the activities (Sao Pedro et al., 2013b). In this new scenario, students may receive scaffolding which could affect how well the models predict skill acquisition. The newly fitted BKT models were able to predict performance within the environment very well under student-level cross-validation, A' = .82 for designing controlled experiments and A' = .84 for testing stated hypotheses. This is on par with our original models' performance (see Chapter 5).

We leverage these models to determine if more students in the DC condition mastered each skill than students in the NDC condition[6]. If the scaffolding has an effect, we would expect students in the DC scaffolding condition to have higher final estimates of latent skill ($L_{final}$) than students in the NDC condition at the end of their Phase Change intervention. To compare groups, we again use a Chi-Squared test because there is non-normality of the residuals when attempting to predict mastery. As such, multiple regression cannot be used. Furthermore, 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 0.95.

As shown in Table 15, significantly more students in the data collection scaffolding (DC) condition mastered the skills than the no scaffolding condition by the end of the Phase Change activities. In particular, 115 out of 131 students (87.8%) in the DC condition mastered the designing controlled experiments skill compared to 108 out of 137 (78.8%) students in the NDC condition, $\chi^2(1) = 3.84$, $p = .050$. Similarly, 121 out of 131 students (92.4%) in the DC condition mastered the testing stated hypotheses skill compared to 107 out of 137 (78.1%) students in the NDC condition, $\chi^2(1) = 10.73$, $p = .001$. These findings, combined with our prior findings, suggest that scaffolding enabled acquisition of these data collection skills.

[6] In Chapter 7, we describe the construction and performance of these new BKT models in more detail. We also show how we incorporated scaffolding and changing of science topics into the BKT framework for addressing the same research questions. The classic BKT model is used here, rather than the proposed extensions, since including these factors may bias estimates of skill mastery for the analyses used here.

**Table 15.** Crosstabulations of practice condition, and whether students mastered each skill ($L_{final} > .95$) by the end of the Phase Change inquiry activities, $n = 268$.

| | Designing Controlled Experiments | | Testing Stated Hypotheses | |
|---|---|---|---|---|
| | Not Mastered | Mastered | Not Mastered | Mastered |
| No Data Collection Scaffolding (NDC) | 29 | 108 | 30 | 107 |
| Scaffolding (DC) | 16 | 115 | 10 | 121 |
| | $\chi^2(1) = 3.84$# | | $\chi^2(1) = 10.73$** | |

#p = .05; **p < .01

Though these results are promising, we note that the analytical approach we chose has two important drawbacks. First, it is possible that students in either condition already knew both skills. A second, related drawback is that some students in the DC condition may never have received scaffolding because they knew the skills. Though students were randomly assigned to each learning condition, it is still possible that such students could skew our findings. To address these potential drawbacks, we re-ran both analyses considering *only students who did not demonstrate either skill in their first data collection* opportunity. From the original set of 268 students, 123 students did not design controlled experiments in their first data collection, and 95 students did not test their stated hypotheses.

When analyzing only students who initially appeared to not know the skills, we found a similar pattern of results. As shown in Table 16, 92.9% of students who did not initially design controlled experiments in the DC condition did so in the unscaffolded Phase Change activity compared 58.5% of the students in the NDC condition, $\chi^2(1) = 20.79$, $p < .001$. In addition, 91.7% students in the DC condition tested their stated hypotheses compared to 53.2% of the students in the NDC condition, $\chi^2(1) = 17.69$, $p < .001$. When comparing conditions in terms of mastery using the BKT models, again we found that significantly more students who initially did not demonstrate skill appeared to have benefitted from scaffolding. As shown in Table 17, 85.7% of the students in the DC condition mastered skill by the end of the Phase Change activities compared to 52.8% of the students in the NDC condition, $\chi^2(1) = 16.02$, $p < .001$. For testing stated hypotheses, 90.0% of the students in DC condition mastered this skill

compared to 46.8% of the students in the NDC condition, $\chi^2(1) = 20.11$, $p < .001$. Combining all the results together suggests that overall scaffolding can help students acquire and demonstrate the two data collection skills of interest, even if they originally do not know the skills.

**Table 16.** Crosstabulations of practice condition, and whether students who did not originally demonstrate skill in their first practice attempt eventually demonstrated skill in the last, unscaffolded Phase Change inquiry activity, $n = 123$ for designing controlled experiments, and $n = 95$ for testing stated hypotheses.

| | Designed Controlled Experiments? | | Tested Stated Hypotheses? | |
|---|---|---|---|---|
| | No | Yes | No | Yes |
| No Data Collection Scaffolding | 22 | 31 | 22 | 25 |
| Scaffolding | 5 | 65 | 4 | 44 |
| | $\chi^2(1) = 20.79$*** | | $\chi^2(1) = 17.69$*** | |

***p < .001

**Table 17.** Crosstabulations of practice condition, and whether students who did not originally demonstrate skill in their first practice attempt mastered the skills ($L_{final} > .95$) in the last, unscaffolded Phase Change inquiry activity, $n = 123$ for designing controlled experiments, and $n = 95$ for testing stated hypotheses.

| | Designing Controlled Experiments | | Testing Stated Hypotheses | |
|---|---|---|---|---|
| | Not Mastered | Mastered | Not Mastered | Mastered |
| No Data Collection Scaffolding | 25 | 28 | 25 | 22 |
| Scaffolding | 10 | 60 | 5 | 43 |
| | $\chi^2(1) = 16.02$*** | | $\chi^2(1) = 20.11$*** | |

***p < .001

### 6.3.2 Effects of Scaffolding on Data Collection Skill Transfer

We address our second research question of whether scaffolding has a positive impact on the transfer of inquiry skills (cf. Thorndike & Woodworth, 1901; Singley & Anderson, 1989) when experimenting in a second, unscaffolded physical science domain using a similar analytical approach (cf. Sao Pedro et al., 2012b). In particular, we measured transfer to the Free Fall activities in two ways. First, students may show immediate transfer by demonstrating skill on their *first attempt* at data collection in the new science topic. In other words, students may have more initial skill when starting the activities, and thus

show immediate transfer. Second, practicing and/or being scaffolded in Phase Change may also prepare students to learn skills while practicing in Free Fall (Bransford & Schwartz, Rethinking Transfer: A Simple Proposal with Multiple Implications, 1999). We account for this possibility by using our BKT models to compare each condition on whether or not they mastered each skill by the end of the Free Fall activities. As in our previous analyses, a student is considered as having mastered a skill if the final estimate of their knowledge generated by the BKT model ($L_{Final}$) is above 0.95. Finally, we consider only students who completed the Phase Change activities and completed at least one Free Fall activity, a total of 261 students.

In terms of immediate transfer of skill to Free Fall, as Table 18 indicates scaffolding appeared to positively impact whether students could transfer their skill to this new topic. More specifically, 118 out of 129 students (91.5%) in the data collection scaffolding (DC) condition designed controlled experiments in their first data collection attempt in Free Fall compared to 107 out of 132 (81.1%) of students in the no scaffolding condition (NDC). The difference between the groups was significant, $\chi^2(1) = 5.95$, $p = .015$. Also, significantly more students in the DC condition tested their stated hypotheses (123 out of 129, 95.3%) than the NDC condition (113 out of 132, 85.6%), $\chi^2(1) = 7.15$, $p = .007$. In terms of mastery by the end of the Free Fall activities, scaffolding again appeared to have a positive effect. As shown in Table 19, 96.9% of the students in the DC scaffolding condition mastered the designing controlled experiments skill by the end of the Free Fall activities, compared to 87.1% of the students in the NDC condition, $\chi^2(1) = 8.43$, $p = .004$. Similarly, 96.9% of the students in the DC condition mastered the testing stated hypotheses skill compared to 85.6% of the students in the NDC condition, $\chi^2(1) = 12.36$, $p < .001$. Taken together, these results indicate that scaffolding also has a positive impact on the transfer of inquiry skills across domains.

**Table 18.** Crosstabulations of practice condition, and whether students demonstrated each skill in their first opportunity to collect data in the Free Fall inquiry activities, $n = 261$.

| | Designed Controlled Experiments? | | Tested Stated Hypotheses? | |
|---|---|---|---|---|
| | No | Yes | No | Yes |
| No Data Collection Scaffolding (NDC) | 25 | 107 | 19 | 113 |
| Scaffolding (DC) | 11 | 118 | 6 | 123 |
| | $\chi^2(1) = 5.95$* | | $\chi^2(1) = 7.15$** | |

*p < .05; **p < .01

**Table 19.** Crosstabulations of practice condition, and whether students demonstrated mastery of skill ($L_{final} > .95$) in the Free Fall activities, $n = 261$.

| | Designing Controlled Experiments | | Testing Stated Hypotheses | |
|---|---|---|---|---|
| | Not Mastered | Mastered | Not Mastered | Mastered |
| No Data Collection Scaffolding (NDC) | 17 | 115 | 21 | 111 |
| Scaffolding (DC) | 4 | 125 | 4 | 125 |
| | $\chi^2(1) = 8.43$** | | $\chi^2(1) = 12.36$*** | |

**p < .01; ***p < .001

Recall from Section 6.1.2 that these findings may be biased because students in either condition may already know the skills, and that students in the DC condition may never have received scaffolding. To address this, we again re-ran both analyses using students who did not demonstrate skill in their first Phase Change to get at students who are learning with our system. From the original 261 students, 119 did not design controlled experiments in their first Phase Change attempt and 91 students did not test their stated hypotheses.

When comparing students who initially did not demonstrate skill on their first opportunity in Phase Change, we again find scaffolding appears to have a positive impact on transfer to Free Fall. As Table 20 shows, 85.3% of the students who received data collection scaffolds designed controlled experiments in this first practice attempt compared to 64.7% of the students in the no data collection scaffolding condition, $\chi^2(1) = 6.87$, $p = .009$. Significantly more Students in the DC condition (87.0%) tested their stated hypotheses than students in the NDC condition (64.4%), $\chi^2(1) = 6.29$, $p = .012$. In

terms of mastery, 94.4% of these students in the DC condition mastered designing controlled experiments by the end of the Free Fall activities compared to 68.6% of the NDC students. This difference was significant, $\chi^2(1) = 13.54$, $p < .001$. The same was true for mastering the testing stated hypothesis skill in Free Fall; 93.5% of the DC students did so compared to 57.8% of the students in the NDC condition, $\chi^2(1) = 15.82$, $p < .001$.

In sum, our results suggest that scaffolding not only helped students acquire skills during their learning in Phase Change, but also enabled them to transfer skill to a second physical science topic, Free Fall. Given the high percentage rates of skill demonstration and mastery in each learning condition (DC and NDC) across domains, and the finding that scaffolding appears to improve transfer across domains, it appears these inquiry skills have a degree of domain generality (cf. van Joolingen, de Jong, & Dimitrakopoulout, 2007). We discuss the implications of these findings in the following section.

**Table 20.** Crosstabulations of practice condition, and whether students who did not originally demonstrate skill in their first practice attempt eventually demonstrated skill in their first attempt at collecting data in the Free Fall activities, two weeks after the intervention, $n = 119$ for designing controlled experiments, and $n = 91$ for testing stated hypotheses.

| | Designed Controlled Experiments? | | Tested Stated Hypotheses? | |
|---|---|---|---|---|
| | No | Yes | No | Yes |
| No Data Collection Scaffolding | 18 | 33 | 16 | 29 |
| Scaffolding | 10 | 58 | 6 | 40 |
| | $\chi^2(1) = 6.87$** | | $\chi^2(1) = 6.29$* | |

*p < .05; **p < .01

**Table 21.** Crosstabulations of practice condition, and whether students who did not originally demonstrate skill in their first Phase Change practice attempt eventually mastered the skills ($L_{final} > .95$) in the Free Fall activities, $n = 119$ for designing controlled experiments, and $n = 91$ for testing stated hypotheses.

| | Designing Controlled Experiments | | Testing Stated Hypotheses | |
|---|---|---|---|---|
| | Not Mastered | Mastered | Not Mastered | Mastered |
| No Data Collection Scaffolding | 16 | 35 | 19 | 26 |
| Scaffolding | 4 | 64 | 3 | 43 |
| | $\chi^2(1) = 13.54$*** | | $\chi^2(1) = 15.82$*** | |

***p < .001

## 6.4 Discussion

As previously stated, science educators and reseachers agree that cultivating inquiry skills is critical for students to become scientifically literate (e.g. NRC, 1996, 2000, 2011; Kuhn, 2005a) and to be well-poised for the job demands of the 21st century (Clarke-Midura el al., 2011). As such, it is of importance to understand how best to foster acquisition and transfer of such skills to novel tasks, particularly since students typically have difficulty with inquiry (Schauble, 1990; de Jong & van Joolingen, 1998; Brown & Campione, 1994). In our approach, we extended our inquiry environment, Inq-ITS, a computer-based environment that can automatically assess students' scientific inquiry skills (Sao Pedro et al., 2012a, 2013a,b; Gobert et al., 2012), to incorporate automated, real-time scaffolding. In particular, we explored whether scaffolding would help students acquire and transfer two data collection skills, designing controlled experiments and testing stated hypotheses.

To do so, we randomized which students received data collection scaffolds in one set of Physical Science activities for Phase Change, and determined if they could demonstrate skill in inquiry activities for a second topic, Free Fall. We evaluated 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 by leveraging Bayesian Knowledge Tracing (Corbett & Anderson, 1995). This enabled us to aggregate information across all practice activities, and account for the potential of students' learning. Overall, we found that our scaffolding approach appears to help students acquire these skills in the same domain in which they were learned (Phase Change), and also to transfer them to the second science topic (Free Fall). In other words, more students in the scaffolding condition showed that they

could design controlled experiments and test their hypotheses than students in the no-scaffolding condition when collecting data with a simulation once these scaffolds are removed.

This work makes three contributions to the literature on inquiry learning. First, these findings are particularly promising as an approach to simultaneously assess and support inquiry skill development in a scalable way because the approach is entirely computer-based. Thus, this system has the potential to be implemented readily in a classroom setting, or as virtual homework, and can provide individualized support to students who need it. Furthermore, this is the first system to our knowledge that evaluates students' inquiry processes (the actions they take while collect data) and uses that information to provide immediate feedback as students experiment.

Second, we showed that a "middle ground" between direct instruction and discovery learning (Kirschner et al., 2006; Hmelo-Silver et al., 2007) has the potential to enable acquisition and transfer of these skills, in line with Vygotsky's notion of scaffolding (Vygotsky, 1978). We do recognize, though, it is possible that the implied transfer of skill may be due to the structural similarities of the activities (Thorndike & Woodworth, 1901), or because both topics were in the Physical Science domain. Thus, we interpret our findings as showing near transfer of skill (cf. Cree & Macaulay, 2000). It will be beneficial to conduct a similar study using activities from dissimilar domains with different activity structures, like Life and Earth Science (Gobert et al., 2012), to tease apart these possible effects and determine if scaffolding enables broader transfer. Even more broadly, a study could be conducted to determine if practice and scaffolding within this environment improves (and predicts) performance on science tasks external to the environment (e.g. Klahr & Nigam, 2004).

Third, our findings also contribute to prior findings of the domain generality of inquiry skills (Glaser et al., 1991; Kuhn et al., 1992; Harrison & Schunn, 2004; Kuhn & Pease, 2008), particularly for skills related to data collection (Klahr & Nigam, 2004; Sao Pedro et al., 2012b). This is evidenced by students in the scaffolding condition being able to transfer their skills to the second science topic.

Because our approach is scalable, we were able to analyze a larger nuber of students than prior work (e.g. Kuhn & Pease, 2008). Furthermore, to our knowledge, research about the effectiveness of providing real-time, explicit scaffolding of inquiry on the transfer of skills between *specific content domains* has not been conducted in computer-based inquiry environments.

There are some limitations to this study. First, we note that our study did not randomize the order in which science topics were presented; students always conducted inquiry in Phase Change first. This design was chosen because our primary goal was to first establish the efficacy of our scaffolding approach, not to assess the domain generality of inquiry skills. We note, though, that are scaffolding approach is general and can be readily integrated within the Free Fall activities. As such, it is of interest to run a study to determine the bidirectionality of transfer (cf. Kuhn et al., 1992).

Second, we did not fully address whether this environment could be used to teach the data collection skills to students who do not know these skills (cf. Siler et al., 2010). In general, we envision our learning environment to be an assessment platform that provides students just-in-time help, not as a pure instructional tool. In other words, we expect this tool to be used as an environment to hone inquiry skills that provides scaffolds as needed during practice (cf. Anderson et al., 1995; Heffernan et al., 2006; VanLehn, 2004), after students are exposed to these inquiry topics in their regular curriculum (Gobert et al., 2012). The analyses presented seek to address learning by separating out students who initially did not demonstrate data collection skills to better capture the effects of scaffolding. From these analyses, scaffolding appears to have positive impact on acquisition and transfer of skill. However, this analytical approach is an approximate measure for pretesting students (which unfortunately due to student availability, was not possible) and using those data to gauge learning. Given our findings about the domain generality of these two activities, we could use activities from a third Physical Science topic, e.g. density (Sao Pedro et al., 2012b; Gobert et al., 2012), to ascertain if our system could be used as a learning tool for these inquiry skills.

Third, the evidence of acquisition and transfer is rooted primarily in procedural demonstration of the two data collection skills, and does not tap changes in conceptual / metastrategic knowledge of when and why one should apply them (cf. Kuhn, 2005a,b; Zohar & David, 2008). One possible way to address this is to have students explain why they chose to design the experiments they did and code the open responses for evidence of such understanding (e.g. Kuhn & Pease, 2008). This approach, however, would be difficult to scale. Another possibility is to triangulate students' performance in the "analyze data" task in which students make inference about the data they collected (Gobert et al., 2012) with their performance in the "experiment" task. If students are able to successfully warrant their interpretations relative to their hypotheses by identifying which data enabled them to make inferences, this would be evidence of conceptual understanding of the data collection skills.

For future work, we will address if there were differential impacts of certain scaffolds on skill acquisition and transfer. In other words, certain scaffolds may have been more useful than others in helping students understand how to design controlled experiments or test their stated hypotheses. Our logfiles, though, will enable us to address this question because they contain all interactions between the student and the pedagogical agent. On a related note, we did not tease apart if certain scaffolds were more effective for certain students than others. In particular, the scaffolding approach is proactive in that as soon as the system detects students are off-track, scaffolding is activated. It also presents advice to students in the same order each time progressing from more general to more specific help. This is one of many potential "scaffolding policies" (Chi et al., 2010) that may better help students acquire and transfer these skills.

The approach taken in this chapter was to determine the effects of scaffolding, transfer and domain generality of inquiry skills by comparing students' performance and mastery with chi-squared analyses. In the next chapter, we use an alternate approach, extending the classic Bayesian Knowledge Tracing model (Corbett & Anderson, 1995) to, in part, address these same questions. The aim will not

only to determine if the results found here can be borne out by interpreting the BKT models (cf. Baker & Yacef, 2009), but also if these data can improve the reliability in predicting student performance within our learning environment.

# 7 Extending the Bayesian Knowledge Tracing Framework to Incorporate Scaffolding and Changing of Science Topics

Many extensions to the classic Bayesian Knowledge Tracing (BKT) model (Corbett & Anderson, 1995) have been developed to improve performance at predicting skill within intelligent tutoring systems, and to increase the interpretability of the model. For example, extensions have been made to account for individual student differences (Pardos & Heffernan, 2010; Baker et al., 2010b), to incorporate item difficulty (Pardos & Heffernan, 2011), to address learning activities requiring multiple skills (Koedinger, Pavlik, Stamper, Nixon, & Ritter, 2010), and even to incorporate the effects of automated support given by the system (Beck, Chang, Mostow, & Corbett, 2008; Jonsson et al., 2005; Yudelson et al., 2008). Extensions have also been added to increase model interpretability and to provide insight about tutor effectiveness. For example, Beck et al. (2008) incorporated scaffolding into BKT to determine if automated support improved students' learning and performance. However, taking into account the differences in tutor contexts, the nature of the activities or problems in which skills are applied, has only been studied in a limited fashion (Yudelson et al., (2008) is one of the few examples). Context is important to consider, because skills learned or practiced in one context may not transfer to new contexts (Thorndike & Woodworth, 1901; Singley & Anderson, 1989), which in turn, could reduce predictive performance if the model is to be used across contexts. Explicitly considering context within knowledge modeling may also increase model interpretability and potentially reveal whether some skills are more generalizable, and thus transferrable.

In this section, we explore the impacts of incorporating scaffolding and the change of tutor context within the BKT framework. We apply these models to track students' data collection inquiry skills (cf. NRC, 1996) as they experiment within interactive, scaffolded simulations for two Physical Science topics (Gobert et al. 2012). We extend BKT in two ways to address learning. First, like Beck et

al., (2008), Johnson et al., (2005) and Yudelson et al. (2008), we incorporate scaffolding as an observable and add model parameters to account for its potential impacts on learning. Second, we add parameters and observables to account for when students change the science topic in which they practice inquiry.

These extensions are motivated by our prior work (Sao Pedro et al., 2012b; Sao Pedro et al., 2013b; Section 5) in constructing BKT models to track skills within an *unscaffolded* version of simulation-based activities about Phase Change. Though the models could predict students' performance, we noticed they had very low learning rate parameters. Since then, we added scaffolding to these activities that automatically provides feedback to students when they engage in unproductive data collection. By incorporating scaffolding into our BKT models, we aim to improve prediction and to determine the degree to which scaffolding impacts skill acquisition. In other words, to paraphrase Beck et al. (2008), we want to know "Does our help help?" Explicitly modeling this improvement may enhance the learning environment's ability to predict performance. In particular, if the scaffolding we provided is effective, we expect that learning rate should increase when students receive help by the system (cf. Beck et al., 2008).

Similarly, the tutor context, e.g. the science topic in which students are practicing and demonstrating their inquiry skill, may also play an important role, especially given evidence that inquiry skills may be tied to the context in which they were learned (van Joolingen, de Jong, & Dimitrakopoulout, 2007). Though evidence exists inquiry skills can be domain general (e.g. Sao Pedro et al., 2012b; Glaser et al., 1991; Kuhn et al., 1992), it is still an open question if inquiry skills are domain general, and if there are factors that enable transfer across contexts (cf. Thorndike & Woodworth, 1901; Singley & Anderson, 1989). Thus, from the viewpoint of predicting student performance, changing domains may impact how well a BKT model can predict future performance. By

explicitly modeling tutor context, we may be able to discern from the model parameters the degree to which inquiry skills are domain general.

## 7.1 Participants and Procedure

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

These data were collected as part of a study to determine the impacts of automated scaffolding on acquisition and transfer of data collection skills across science topics (discussed in detail in Chapter 6). In this study, students were assigned 5 Phase Change inquiry activities, and two weeks later, 5 Free Fall activities. Students were allotted approximately two class periods per science topic to complete the activities. Due to time constraints, some students did not finish all the activities in each science topic. Recall that in each activity, students formulated hypotheses, collected data and analyzed data. In the first 4 Phase Change activities, all students had scaffolding available as they formulated hypotheses. However, some students were randomly chosen to have data collection scaffolds available, whereas others did not. In the scaffolding condition, Rex (see Section 6.1) provided feedback to the students when they were evaluated as not demonstrating good data collection behavior. Students who were in the no-scaffolding condition received no feedback on their data collection. In the "analyze data" inquiry task, no students received scaffolding.

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

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

## 7.2 Evaluating the Demonstration of Data Collection Skill

We used automated methods for evaluating data collection skills (Sao Pedro et al., 2012a, 2013a,b). This evaluation was used both to trigger scaffolding, and to provide observables of student performance for building Bayesian Knowledge Tracing models. Specifically, we aim to assess two data collection process skills, designing controlled experiments and testing stated hypotheses (Gobert et al., 2012). These are demonstrated as students collect data using the simulation in the "experiment" stage of inquiry. Briefly, students design controlled experiments when they generate data that make it possible to determine what the effects of independent variables (factors) are on outcomes. They test stated hypotheses when they generate data that can support or refute an explicitly stated hypothesis. Since these are process skills, students are assessed based on the actions they take while collecting data.

We evaluate whether students demonstrate these skills by combining predictions made by data mined detectors (Sao Pedro et al., 2012a, 2013a,b) with knowledge-engineered rules to handle specific edge cases. The process, described in detail in Section 6.1.3, enabled successful evaluation of students' data collection process skills. In the next section, we describe the data distilled from students' usage of the Phase Change and Free Fall activities. These data are used to develop and test the BKT extensions.

## 7.3 Dataset for Building and Evaluating Extended BKT Models

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

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

## 7.4 Extending BKT to Incorporate Scaffolding and Tutor Context

We amalgamated students' performances across activities within a Bayesian Knowledge-Tracing framework (Corbett & Anderson, 1995). BKT is a two-state Hidden Markov Model that estimates the probability a student possesses latent skill ($L_n$) after *n* observable practice opportunities[7] ($Prac_n$). In our domain, latent skill is defined as knowing how to perform the data collection skills, and a practice opportunity is an evaluation of whether skill was demonstrated during data collection in an inquiry activity. A practice opportunity begins when students enter the "experiment" task in an inquiry activity. An opportunity ends when a student switches from the "experiment" task to the 'analyze data" task (see Chapter 3). As mentioned, the detectors / knowledge engineered rules evaluate students' actions, and these evaluations act as the observables. A student is evaluated as not having demonstrated skill ($Prac_n = 0$) if one of two cases occurs. The first is if they are evaluated as not demonstrating a skill when they signal completion of data collection (e.g. attempt to switch to the "analyze data" task). The second is if, while collecting data, the system believes the student does not know either skill and provides scaffolding.

[7] Practice opportunities as written in this section are equivalent to clips, as discussed in Section 5.2.

This approach to address scaffolding's impact on student correctness is similar to others (e.g. Pardos & Heffernan, 2011).

The classic BKT model (Corbett & Anderson, 1995) is characterized by four parameters, *G*, *S*, $L_0$, and *T*. The Guess parameter (*G*) is the probability the student will demonstrate the skill despite not knowing it. Conversely, the Slip parameter (*S*) is the probability the student will not demonstrate the skill even though they know it. $L_0$ is the initial probability of knowing the skill before any practice. Finally, *T* is the probability of learning the skill between practice attempts. From these values, the likelihood of knowing a skill P($L_n$) is computed as follows:

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

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

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

This classic BKT model (Corbett & Anderson, 1995) carries a few assumptions. First, the model assumes that a students' latent knowledge of a skill is binary; either the student knows the skill or does not. The model also assumes one set of parameters per skill and that the parameters are the same for all students. Finally, the classic model assumes that students do not forget a skill once they know it.

Relevant to this work, the classic model does not take into account whether students received any scaffolding from the learning environment and does not account for the tutor context in which skills are demonstrated (Yudelson, Medvedeva, & Crowley, 2008) (e.g. the science topic). The same skill in different contexts would either be treated as two separate skills, or as having no differences between contexts. In other words, there would be no challenge in transferring the skill to the new tutor context,

an assumption known to be questionable (Singley & Anderson, 1989; Bransford & Schwartz, 1999). Below, we describe our approach to incorporate both factors.

### 7.4.1 Taking Scaffolding into Account

We introduce scaffolding into BKT as an observable, $Scaffolded_n$ = {True, False}, because it can directly be seen if our pedagogical agent provided help to students as they collected data. A similar approach was taken by Beck et al. (2008) to develop the Bayesian Evaluation and Assessment model. In their domain, reading, this scaffolding observable was true if a student received help just before reading a word (each word was treated as a skill). The observable was linked to all four BKT model parameters, meaning that scaffolding could have an impact on initial knowledge ($L_0$), guess ($G$), slip ($S$) and whether or not students learn between practice opportunities ($T$). As a result, their BKT model contained 8 parameters to account for scaffolding.

We instead chose to condition *only* the learning rate ($T$), for three reasons. First, the increase in the number of parameters could result in overfitting, especially since the classic BKT model is already known to be overparametrized (Beck & Chang, 2007). Second, though the additional parameters may facilitate model interpretation, it is unclear whether conditioning all the classic BKT parameters on scaffolding improves predictive performance. In particular, Beck et al. (2008) found no increase in predictive performance when accounting for scaffolding. Finally, the immediate effects of scaffolding on performance may not be relevant because we only look at first practice opportunities (thus looking at overall effects of scaffolding), and because there is a time delay between data collection performance attempts. In particular, students attend to a different inquiry task, analyzing data, after their data collection (see Chapter 3 for more details).

In our extension, conditioning learning on whether students receive scaffolding yields two learning rate parameters, *T_scaffolded* and *T_unscaffolded*. Thus, this model tries to account for the

differential impacts scaffolding may have on whether or not students learn a skill (e.g. the latent variable knowledge transitions from "doesn't know" to "know" after practicing). Mathematically, the original equation for computing P($L_n$) is conditionalized to account for the observable as follows:

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

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

### 7.4.2 Taking Tutor Context into Account

We also developed BKT extensions to the take into account the science topic in which students demonstrate their inquiry skills. Recall that students first practiced inquiry in Phase Change activities (possibly scaffolded or unscaffolded) and then practiced inquiry in unscaffolded Free Fall activities, a different science topic. Modeling the change in science topic is of importance since the degree to which inquiry skills are "domain general" is unclear (van Joolingen et al., 2007). Even though students may learn and demonstrate inquiry skills in one science topic, they may not understand those same skills are applicable to other topics. Thus, tutor context may play an important role in predicting performance (Yudelson, Medvedeva, & Crowley, 2008). We hypothesize that incorporating the change of science topic in the BKT framework may improve models' predictive performance.

We incorporate changing of science topics in two ways. First, we hypothesized that there may be a differential effect in learning between topics. For example, practice in Phase Change may prepare students to learn (and subsequently demonstrate) skills in Free Fall, called "preparation for future learning" (Bransford & Schwartz, 1999). To model differential learning between topics, we again break out the learning rate ($T$), this time for each topic: *T_PhCh*, *T_FF*. A new observable is also added for the current science topic, $Topic_n$ = {PhaseChange, FreeFall}. The result is a "BKT learn rate topic" model with a modification to the P($L_n$) equation similar to the "scaffolded BKT model" described previously.

Our second model for incorporating the change of science topics posits that students may not understand that the skills are applicable in both domains. In other words, students may not understand that these skills are domain general (van Joolingen et al., 2007). We model this notion by adding in a linear degradation factor, $k \in (0,1)$, to potentially offset the likelihood students know the skill P($L_n$) when the science topic switches. If $k = 1$ this implies there is no effect on students' knowledge when the topic switches. When $k = 0$, students will be presumed to not know the skill when the topic switches. We also add an observable $Topic\_Switch_n$ = {True, False} to address when the science topic changes from Phase Change to Free Fall (just before the student's first opportunity to practice in Free Fall). The corresponding P($L_n$) modification for the "BKT skill degradation model" is:

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

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

Note that the degradation parameter $k$ is different than modeling "forgetting" in the BKT framework (cf. Corbett & Anderson, 1995; Yudelson et al., 2008) in two ways. First, we note that the factor is applied to both conditional expressions in the P($L_n$) equation, not just P($L_{n-1}|Prac_n$) as done when modeling forgetting. Second, in these earlier approaches forgetting is modeled at each practice opportunity, whereas our factor is applied at a single point, when the science topic switches.

### 7.4.3 Combining Models

The above models introduce three new potential observables to the BKT framework relevant to our learning environment: $Scaffolded_n$ = {True, False}, $Topic_n$ = {PhaseChange, FreeFall}, and $Topic\_Switch_n$ = {True, False}. The models above individually incorporate the observables by conditioning the learning rate parameter, $T$, on them, or by adding a multiplicative reduction factor, $k$, to the computation of P($L_n$). As part of this work, we also combined the extensions described above into

larger models. The most complicated model incorporated all observables and contained seven parameters: ($L_0$, $G$, $S$, $T\_Scaff\_PhCh$, $T\_Unscaff\_PhCh$, $T\_Unscaff\_FF$, $k$). We next describe our process for fitting these models.

### 7.4.4 Model Fitting

As in Baker et al., (2010b) and Sao Pedro et al., (2013b), we use brute force search to find the best fitting parameters (see Section 5.4 for more details). This method has been found to produce comparable or better model parameters than other methods (Pardos et al., 2012). In this approach, all potential parameter combinations in the search space are tried at a grain-size of 0.01. The best parameter set yields the lowest sum of squares residual (SSR) between the likelihood that the student would demonstrate skill, P($Show\_Skill_n$), and the actual data. This likelihood is computed as follows (Corbett & Anderson, 1995):

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

Once this set has been found, another brute force search around those parameters is run at a grain-size of 0.001 to find a tighter fit. We bound $G$ to be less than 0.3 and $S$ to be less than 0.1 (cf. Pardos et al., 2012); all other parameters can be assigned values in (0.0, 1.0).

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

## 7.5 Results

We determine if extending the classic BKT model to include scaffolding and changing of science topics will 1) improve predictions of future student performance in our learning environment, and 2) yield insights about the effectiveness of our scaffolding approach, and the transferability of the inquiry skills. To address predictive performance, we determined if the new models' predictions of skill demonstration P($Show\_Skill_n$), aggregated from evidence over times {1…$n$-1}, can predict actual student performance at time $n$ better than the classic BKT model. We train and test our models' performance by conducting six-fold student-level cross-validation, stratifying by both learning condition (having scaffolding available in Phase Change or not) and class section. Cross-validating in this way helps ensure that each fold equally represents learning conditions, and students from each class section/school. This increases assurance that models can be applied to new students.

Model goodness was determined using A' (Hanley & McNeil, 1982). This is an appropriate metric to use when the predicted value is binary (either students demonstrated skill in $Prac_n$ or they did not), and the predictors for each model are real-valued, e.g. P($Show\_skill_n$). As a reminder, a model with A' of 0.5 predicts at chance level and a model with A' of 1.0 predicts perfectly.

We compute two variants on A' for student performance data as follows. First, we report overall A' values of each model collapsing over students as we did in Section 5.6.1. Second, we compute the A' values of each model per student (Baker et al., 2010b), and report the average per-student A'. These approaches have different strengths and weaknesses (cf. Sao Pedro et al., 2013b; Baker et al., 2012; Pardos et al. 2012). Collapsing over students is straightforward and enables comparison of models' broad consistency in predicting skill demonstration. In other words, this approach can show, in general, whether or not high likelihoods of demonstration of skill predicted by the model correspond with actual demonstration of skill. In addition, collapsing can be used when there is not enough within variance for each student to produce a meaningful per student A' (cf. Sao Pedro et al., 2013b). Collapsing over

students, however, provides weaker estimates of predicting an individual student's learning and performance than the A' per student metric (Baker et al., 2010b; Pardos et al., 2012). Collapsing may also yield estimates that are biased towards students who practiced more with the system since they contribute more data (Pardos et al., 2012).

When computing A' per student, we only used students who had variation in their evaluations. In other words, a student was not considered if they were evaluated correct on all practice opportunities or incorrect on all practice opportunities. This was necessary because A' is undefined unless there is at least one 'positive', and at least one 'negative' evaluation for a student (Hanley & McNeil, 1982).

Finally, we ascertain whether any of the BKT model variants significantly outperforms the classic BKT model by comparing models' A' values. We use slightly different approaches depending on the A' computed. To compare models via the collapsed A', we first compute standard errors for A' for each model (Hanley & McNeil, 1982)[8] and then perform a *Z*-test to see if the models' predictive power is significantly different (Fogarty, Baker, & Hudson, 2005). To compare models via A' per student, the approach changes somewhat to account for multiple student observations which violate independence assumptions. First, individual *Z* score differences between models' A' are computed for each student using the process just described. Then, the *Z* scores are aggregated using Stouffer's method[9] (cf. Pardos et al., 2012; Baker et al., 2008a,b) to determine if there are significant differences.

[8] $Z = \frac{A'_1 - A'_2}{\sqrt{SE(A'_1)^2 + SE(A'_2)^2}}$

[9] Stouffer's $Z = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}$, where *n* is the number of students.

**Table 22.** BKT model variant performance predicting whether students will demonstrate skill in their next practice attempt in the learning environment. The A' values were computed under six-fold student-level cross-validation Overall, the best model for both skills is the one in which the learning rate is conditioned on whether or not the student received scaffolding during Phase Change (*T_Scaffolded*).

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

*p < .05; **p < .01; ***p < .001, Stouffer's Z difference of A' between Variant and Classic Model (cf. Pardos et al., 2012; Baker et al., 2008a,b).

### 7.5.1 Models' Overall Predictive Capability

As shown in Table 22, all of the models show strong consistency, meaning that high estimates of skill demonstration are associated with actual demonstration of skill. This is evidenced by collapsed A' values ranging from .817 to .837 for the designing controlled experiments skill, and collapsed A' values ranging from .840 to .853 for the testing stated hypotheses skill. Recall that these high collapsed A' values do not reflect the models' ability to predict individual student trajectories (Pardos et al., 2012), because they factor out the student term. In terms of consistency, none of the BKT variants was better statistically at predicting future performance than the Classic BKT model for either skill ($Z <= 1.72$, $p > .08$ for each comparison against the Classic model). In terms of consistency, the model with the highest A' = .837 for predicting future performance of the designing controlled experiments skill 1) conditioned the learning rate on whether the student received scaffolding (T_Scaffolded extension), and 2) incorporated skill degradation when switching between science topics (kLn_TopicSwitch extension).

The model with the highest A' = .853 for predicting future performance of testing stated hypotheses was the full model that incorporated all three extensions to the Classic BKT model.

In terms of predicting individual student performance, some of the models performed reasonably well. As a baseline, the Classic BKT model for designing controlled experiments had a per-student average A' = .635, which is significantly better than chance A' = .5 (Stouffer's $Z = 13.94$, $p < .001$). For testing stated hypotheses, the Classic BKT model had a per-student average A' = .613, again significantly better than chance prediction (Stouffer's $Z = 10.65$, $p < .001$). Though the Classic BKT models predict better than chance, their A' values are somewhat low.

When incorporating some of the BKT variants, the per-student average A' increased. In particular, BKT variants that leveraged conditioning on scaffolding (T_Scaffolded model) performed significantly better than the Classic BKT model (Table 1). For example, the best BKT model variant for both skills incorporated only scaffolding. The per-student average A' of this model for designing controlled experiments was .685, a significant jump over the Classic BKT model (Stouffer's $Z = 3.78$, $p < .001$). The per-student average A' for testing stated hypotheses was .656, and again, outperformed the Classic BKT model (Stouffer's $Z = 2.75$, $p < .01$). These A' values are on par with the extended BKT models developed in Beck et al. (2008) that incorporated scaffolding.

### 7.5.2 Extended BKT Model Interpretation

Like Beck et al. (2008), we interpreted the models' parameters to understand what they reveal about the impacts of scaffolding and the learning and transfer of scientific inquiry skills between Physical Science topics. Since the full models with 7 parameters had A' performance on par with the other best performing models, we chose to interpret their parameters. The parameter averages and standard

deviations for each skill model across all six folds are presented in Table 23. We focus on interpreting the new parameters we added to the model.

In Phase Change, the learning rate when students were scaffolded is much higher than the learning rate without scaffolding, *T_Scaff_PhCh* = .638 vs. *T_UnScaff_PhCh* = .190 for designing controlled experiments, and *T_Scaff_PhCh* = .823 vs. *T_UnScaff_PhCh* = .158 for testing stated hypotheses. These values indicate that scaffolding students' inquiry appears to have a positive effect on whether students learn the skills (Beck et al., 2008).

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

Finally, the skill degradation parameter *k*, which captures the degree of skill transfer between science topic (0 is no transfer, 1 is full transfer), was high for both skills. For designing controlled experiments, *k* = .973 and for testing stated hypotheses, *k* = .961. Given our earlier argument about many students demonstrating skill in FreeFall, these high values suggest that skill transfers from Phase Change to Free Fall within our learning environment, and that, more broadly, these inquiry skills may be domain general. We elaborate on this finding in more detail in the next section.

**Table 23.** Means and standard deviations of the parameter values for full BKT model variant, across all six folds.

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

## 7.6 Discussion and Conclusions on Extended BKT Models

In the classic Bayesian Knowledge Tracing framework (Corbett & Anderson, 1995), scaffolding and the tutor context, the nature of the activities in which skills are applied, are not taken into account when predicting students' future performance. Similar to others' prior work (Beck et al., 2008; Jonsson et al., 2005; Yudelson et al., 2008) we explored here whether extending the BKT framework to incorporate these factors improves prediction of students' skill demonstration. This work was conducted to predict students' acquisition of two data collection inquiry skills, designing controlled experiments and testing stated hypotheses (cf. Sao Pedro et al., 2012a, 2013a,b), in performance-based inquiry tasks across two Physical Science topics, Phase Change and Free Fall. Specifically, we added three extensions to the BKT model: 1) conditioning the learning rate on whether or not students were scaffolded; 2) conditioning the learning rate depending on the topic in which students practiced inquiry (Phase Change or Free Fall); and 3) adding a degradation parameter to potentially lower the likelihood of a student knowing a skill when the science topic changed. Overall, we found that BKT can track development of both skills, in accordance with our prior work (Sao Pedro et al., 2013b; Section 5), and that our extensions led to improvements in prediction and model interpretability.

In comparing our BKT extension that incorporates scaffolding, our approach is closest to the one taken in Beck et al. (2008). Our model assumes that scaffolding will *only* impact learning, whereas Beck

et al. (2008) capture that scaffolding may differentially impact learning *and* immediate performance. Our modeling choice was motivated in part by parsimony given that BKT is already overparametrized (Beck & Chang, 2007), a possibility hypothesized in Beck et al. (2008), and by the delay between performance attempts of the skills in our learning environment. Unlike their work, we found that taking scaffolding into account significantly improved the ability to predict individual student learning and performance over the classic BKT model, possibly due to increased parsimony. We also teased apart the effects of scaffolding on our models' predictive abilities overall (collapsing over students) and on predicting individual student performance.

When interpreting the parameters of the extended model, we found that scaffolding appears to have a positive impact on learning, as in Beck et al. (2008). We do note, though, that we did not tease out the differential impacts of specific scaffolds in our multi-level scaffolding approach. It is possible that specific scaffolds trigger different degrees of learning. One possible way to incorporate this is to condition learning rate on the different kinds of scaffolds, not just whether or not students received scaffolding in general.

We also incorporated parameters to account for the possible effects of demonstrating inquiry skill within different science topics (Phase Change and Free Fall). This modeling was inspired by the empirical question of whether inquiry skills are tied to a domain in which they are learned (van Joolingen et al., 2007), or if they are domain general and can transfer across contexts (Thorndike & Woodworth, 1901; Singley & Anderson, 1989). Though incorporating these parameters did not increase the predictive performance of our models, they do provide possible insights to inquiry learning. In particular, the model parameters suggest that the data collection skills of interest may be domain general and thus transfer across science topics. This supports earlier findings (e.g. Glaser et al., 1991; Kuhn et al., 1992; Sao Pedro et al., 2012b). There are limits to how certain we can be about this interpretation, though. First, in our study design, we only randomized whether students received scaffolding in Phase

Change, and then measured transfer to Free Fall. A stronger approach to increase parameter interpretability would be to also randomize the science topic order. Second, it is possible that the implied transfer of skill may be due to the structural similarities of the activities (Thorndike & Woodworth, 1901) across Physical Science tasks. In the future, it will be beneficial to conduct a similar study using activities from dissimilar domains, like Life and Earth Science (Gobert et al., 2012), with different activity structures to tease apart these possible effects.

This study offers three contributions. First, to our knowledge, this work is the first application of BKT to track the development of inquiry process skills across science topics. This work strengthens our earlier findings in using BKT for a single group of students and single topic (Sao Pedro et al., 2013b), because we cross-validated our models with students from multiple schools who engaged in two science topics. Second, we extended BKT by incorporating scaffolding. Though this extension is similar to others' (Beck et al., 2008; Yudelson et al., 2008), it enabled a "discovery with models" analysis (cf. Baker & Yacef, 2009) that shed light on the potential relationships between performance in the environment, scaffolding, and transfer of inquiry skills (van Joolingen et al., 2007). Furthermore, conditioning the BKT learning rate on whether students received scaffolding significantly improved prediction of individual students' trajectories over the classic model. Finally, we incorporated tutor context directly in the BKT model, unlike Yudelson et al. (2008), who addressed this by selecting subsets of training and testing data to target specific contexts. As mentioned, context played an important role in leveraging BKT to discern the domain generality of the data collection skills.

We note that this work focuses primarily on validation and interpretation of skill *within our learning environment.* In our prior work (Sao Pedro et al., 2013b; Section 5), we also showed that BKT models not only had this internal reliability, but were also moderately predictive of other transfer measures of inquiry. In the future, we will determine if our model extensions can also improve external

validation, thus realizing the full potential of using our learning environment to estimate and track authentic inquiry skills.

# 8 Summary of Findings, Implications, and Future Work

Science educators, reseachers, and K-12 frameworks (NRC, 1996, 2000, 2011) agree that rich integration of content knowledge, critical thinking, and inquiry skills are all necessary for students to become scientifically literate (e.g. Kuhn, 2005a), and to thrive in a knowledge-based economy (Clarke-Midura et al., 2011). The National Research Council (1996, 2000, 2011) further calls for rigorous knowledge about learning processes to inform better assessments of science, which will in turn, lead to deeper student understanding. Despite these widespread calls to transform science education, rote vocabulary, facts, and formulas are still emphasized and assessed. One of several reasons for this is that assessing inquiry in a scalable, reliable way is challenging (Fadel et al, 2007). Typically, if inquiry is assessed it is done so using multiple-choice items (cf. Alonzo & Aschbacher, 2004), even though there is wide acknowledgement that these do not and cannot adequately measure complex science process skills (Black, 1999; Pellegrino et al., 2001; Quellmalz et al., 2007; Clarke-Midura et al., 2011; Gobert et al, 2012). Such summative assessments may also occur weeks after instruction, too late for timely intervention by teachers.

To overcome these challenges, organizations such as NAEP, PISA, the National Educational Technology Plan, and the National Research Council (NRC, 2011), and researchers (e.g. Buckley et al., 2006, 2010; de Jong, et al., 2010; Clarke-Midura et al., 2011; Quellmalz et al., 2012; Rowe & Lester, 2010; Ketelhut et al., 2010; Gobert et al., 2012) have begun exploring the potential of using computer-based environments as a scalable approach to assess inquiry. These environments present their own assessment challenges, though, and as such are not yet used for assessment to a substantial degree (Quellmalz et al., 2009). One key barrier is making sense of the complex and voluminous amount of log data that are generated as students conduct inquiry. Since activities include many, non-trivial, inter-related tasks that must be triangulated to measure skill (Williamson et al., 2006), traditional psychometric approaches do not readily apply (Mislevy et al., 2012). Thus, new techniques are required

to both measure and elicit students' thinking about inquiry which can in turn be used to provide contextualized and effective feedback to both students and teachers.

The overarching goal of this dissertation was to develop automated techniques that directly address these issues of measurement, reliability, scalability and feedback when assessing students' inquiry within computer-based environments. In particular, we extended the Inq-ITS system (Gobert et al., 2012), an inquiry environment in which students experiment using simulations / microworlds, to automatically evaluate, track and scaffold two skills related to data collection (cf. NRC, 1996), designing controlled experiments and testing stated hypotheses.

These skills are process skills (cf. Rupp et al., 2010) that are demonstrated as students design and run experiments with the simulation. As such, students' interactions with the system form the basis for skill assessment. The automated approach to evaluate (auto-score) and track inquiry based on student interaction data heavily relied on techniques developed by the Educational Data Mining field (cf. Romero & Ventura, 2007; Baker & Yacef, 2009). The automated approach for scaffolding was based prior work from the Intelligent Tutoring Systems field (cf. Anderson et al., 1990, 1995; Koedinger & Corbett, 2006). Finally, the evaluation and scaffolding systems were integrated within Inq-ITS Physical Science activities and used to test the effectiveness of scaffolding inquiry, a question of interest to the Learning Sciences. As such, this work represents an intersection of these three fields.

Below, we summarize the overall findings of this dissertation and present a broad discussion of the implications of our technique and findings. More specific discussions, contributions and limitations for each topic appear at the end of each chapter.

- *Leveraging data mining techniques as the basis for evaluating data collection skills (Chapter 4):* Overall, we found that our data-mining based detectors could be used to reliably assess the two data

collection skills of interest, with some limitations[10]. Through a rigorous, iterative design approach we found that improving the construct validity of our detectors by judiciously selecting the features which comprise them also improves their predictive capabilities. Furthermore, by increasing construct validity, the same detectors could be used for multiple purposes: evaluating skill and driving scaffolding (Section 4.2). Finally, we developed alternative approaches for estimating the generalizability (cf. Baker et al., 2008c) of our data-mined models by testing them in a different context in which they were built (Section 4.3). We found that under student-level validation, in which we ensure each student is equally represented in the test set, these models could predict well for a different science topic (Free Fall) than they were build (Phase Change). Our approach also revealed some weaknesses by exploring a different kind of generalizability, ensuring that different kinds of experimentation patterns were equally represented in the test set.

- *Amalgamating students' performance to estimate latent skill (Chapters 5 and 0):* We employed a second EDM technique, Bayesian Knowledge Tracing (Corbett & Anderson, 1995), to aggregate students' performances across time, and estimate latent skill. These models were built, in part, using the evaluations of student performance made by the detectors. We found that they could predict whether students would engage in good data collection on their next opportunity to collect data within the environment, indicating that they have internal validity (Sections 5.6.1, and 7.5.1). We also determined if the latent estimates had additional meaning by testing their external validity. When doing so, we found that they were loosely correlated with multiple-choice style inquiry measurements, and performance-based assessments that focus on skills in isolation (e.g. Sao Pedro et al., 2009, 2010b). Most relevant to this dissertation, these models were leveraged to compare and provide insight to students' acquisition and transfer of inquiry skills, discussed next.

[10] The specific limitations and their implications are discussed in Sections 4.4 and 6.1.3.

- *Modeling and analyzing the complex relationship between scaffolding, and the acquisition and transfer of skill across two physical science domains (Chapters 6 and 7):* Though there are other computer-based learning environments that structure and assess inquiry (see Section 2.3 for a list of systems), this work is the first to detect when students are being haphazard in their inquiry, and intervene in real-time with scaffolding. By conducting a randomized controlled study in which some students received data collection scaffolds and others did not, we were able to determine the impacts of scaffolding on student performance. In particular, we found that practice in scaffolding in one Physical Science topic (Phase Change) yielded more students who could demonstrate those skills in the same domain. In addition, the scaffolding enabled students to transfer those skills to a second domain (Free Fall). These findings were further corroborated by extending the Bayesian Knowledge Tracing model to incorporate scaffolding and changing of science topics. By inspecting these extended models more closely we also saw that scaffolding appeared to have a positive effect on students' performance. Furthermore, incorporating scaffolding also enabled us to better track individual student performance within the learning environment (Section 7.6).

The approach we took to assess and scaffold inquiry has several implications. We discuss these in the following sections.

## 8.1 Implications of our Data Mining-based Approach to Assess Inquiry Skills

At the heart of this work was making sense of and interpreting students' low-level actions within the interface, the complex log data mentioned earlier. This interpretation was necessary to evaluate whether

or not students demonstrate skill as they collect data with the simulations. To handle the complexity, our approach was to employ Educational Data Mining (cf. Romero & Ventura, 2007; Baker & Yacef, 2009) as a means to (1) construct detectors to evaluate students' experimentation, and (2) track students' developing skills across data collection tasks. This approach required text replay tagging (Baker et al., 2006) as a first step, a step that is novel to the systematic study of inquiry learning (Montalvo et al., 2010; Sao Pedro et al., 2010a, 2013b). Text replay tagging, a form of protocol analysis (Ericsson & Simon, 1980, 1984), leveraged human judgment to identify whether students' log files demonstrated inquiry skill. Data mining enabled us to leverage human's codes to build and validate automated "detectors" of each skill that can replicate human judgment. These detectors evaluated students' interactions with the microworld and enabled us to distinguish students who designed controlled experiments and tested their hypotheses from those who did not. Data mining also enabled us to aggregate together each students' evaluations to build a profile of their proficiency across tasks. As mentioned earlier, this was executed using the Bayesian Knowledge Tracing (Corbett & Anderson, 1995) framework. This approach, combining text replay tagging and data mining, we believe, has several benefits that impact the measurement, reliability, scalability and feedback when assessing inquiry skills in computer-based environments.

From a measurement standpoint, this approach has shown promise in evaluating and tracking skill demonstration where there is variation and noise in what students do in a learning environment. This variation is expected because students can explore the simulation in any way they choose, and there is no single right or wrong way to demonstrate the data collection skills of interest (Shute, Glaser & Raghavan, 1989; Glaser et al., 1991). Our technique leverages students' interaction data and human judgment of those interactions to develop canonical models of what it means to demonstrate each skill. In other words, each model evaluates if a student is demonstrating skills taking into account variability in their experimentation processes. Thus, this work represents an advancement over prior approaches

because they do not account for this variability (e.g. Dean Jr. & Kuhn, 2006; Kuhn & Pease, 2008; McElhaney & Linn, 2008, 2010; Gobert & Koedinger, 2011), which may lead to over or underestimation of skill demonstration. In addition, the approach we took to aggregate students' activities to produce estimates of latent skill was principaled and data-driven. Though the use of Bayesian Knowledge Tracing to track skills is not new, its application to tracking specific inquiry skills is new.

Reliability is directly addressed because the goodness of EDM-based detectors and the Bayesian Knowledge Tracing models is quantified. These goodness metrics were computed by testing models' predictive capabilities against data *not used to build the models* (cf. Efron & Gong, 1983; Witten & Frank, 2005). In terms of evaluating students' skills in a single activity, these techniques enabled us to measure our iterative improvements of the detectors (Section 4.2) and let us measure how they generalize to entirely new sets of students and activities from which they were built (Section 4.3). Arguably, the most important advantage of using data mining based-models over knowledge-engineered approaches used previously (e.g. McElhaney & Linn, 2008, 2010) is in this ability to validate their goodness more easily since test labels are available (they are necessary for classification). In terms of practical significance, however, it is still an open question whether our machine-learned models better represent our inquiry behaviors than their knowledge-engineered counterparts. Thus, a valuable future step will be to compare knowledge-engineered approaches to our machine-learned models in terms of which better predicts a new set of text-replay tagged data.

From a tracking standpoint, we address reliability in two ways. First, reliability is addressed using cross-validation (cf. Efron & Gong, 1983) to test if the classic BKT models, and the extensions we proposed (Chapter 7), could predict students' performance on an inquiry task given their prior evaluations. Second, our approach has students engage in multiple attempts to collect data, enabling us to get more stable estimates of students' skills, as opposed to a single point estimate (cf. Shavelson et al.,

1999). Of importance for high-stakes testing will be to determine the minimum number and nature (e.g. how many different topics) of activities that are needed to produce stable and valid estimates of students' inquiry skills, given that the time to administer such tests is limited.

Scalability is addressed in three ways. First, our approach to capturing and assessing students' inquiry is entirely computer-based and runs over the web with no installation necessary (Gobert et al., 2012). Thus, students who have internet access and a computer can use our materials to practice inquiry. Second, our automated approach enables rigorous, automatic scoring of students' inquiry, which is an improvement over prior methods that used hand-scored open responses and/or reports (Kuhn et al., 1992; Kuhn & Pease, 2008). Finally, our approach to testing the reliability of our models using unseen data also impacts the scalability of the approach. In particular, the generalizability tests we conducted (Section 4.3) indicates how well these models can scale to evaluate skills in different simulation-based contexts (Gobert et al., under review; Sao Pedro et al., 2013a) and different students (e.g. Sao Pedro et al., 2013a).

Finally, our approach addresses the timeliness of feedback. As students work, their inquiry is evaluated in real-time using the detectors. Thus, teachers can receive immediate notification about how each individual student, and their class as a whole is progressing on individual inquiry skills. We will discuss the implications of this in more detail in the next section. The detectors were also used in part to scaffold students. In other words, as students work, the system is capable of providing immediate feedback as they work. This feedback has been shown to help students acquire and transfer the data collection skills of interest, as shown in Chapters 6 and 0. As such, this represents a significant advance over assessments that take days (or even months) to provide feedback.

## 8.2 Implications for Formative and Summative Assessments of Inquiry

Taken together, the aforementioned advances in measurement, reliability, scalability, and feedback all impact formative and summative assessment of inquiry. Central to this system is the use of computer-based simulations and inquiry support tools (e.g. automatically populating data tables, hypothesizing tools, etc.) as a means to elicit students' inquiry processes and work products (cf. Rupp et al., 2010). Simulations, generally speaking, have many affordances for learning about (cf. Gobert, 2005) and assessing inquiry (NRC, 2011; Quellmalz et al., 2009, 2012; Gobert et al., 2012). But, perhaps most relevant to this work, they provide two key benefits. First, they enable students to design experiments and run trials (replications) more quickly than experimenting using physical apparatuses. As such, in a normal class period, students can spend more time engaging in inquiry and not have to wait for phenomena to occur. Second, because we evaluate students on their data collection skills based on their actions, teachers now have data available to them on *how their students conduct inquiry*, not just the final work products they create. This information, in turn, can provide teachers important information about students' progress and drives the automated support given to students who are struggling.

The evaluations given by the detectors and the estimates of latent knowledge given by the BKT models can be leveraged to both summatively and formatively assess students' inquiry. In terms of summative assessment, our materials could be used as a means to assess students' data collection skills by turning off scaffolding for all students. Furthermore, the techniques we described for validating assessment models of inquiry process skills could enable more widespread use of these materials for high-stakes testing since they directly address measurement and reliability as discussed previously (cf. Clarke-Midura et al., 2011). Though this work is promising, more data and validation of these data are needed to determine the degree to which our assessment models can scale to larger numbers of students from wider demographics, and to more science topics.

In terms of formative assessment, these data can be used to generate assessment reports for teachers as students work. Currently, the detectors have been deployed in the Inq-ITS system (as was necessary to implement scaffolding), and reports have been implemented to aggregate student performance in various ways. For example, teachers can view whether individual students are having difficulty with data collection skills or can view the performance of their class as a whole. Thus, as students work, teachers can make pedagogical decisions about whether to provide assistance to the class as a whole, or help individual students. Currently, students' overall performance in these reports is computed using an average-based approach. Though useful, in the future, we aim to integrate the Bayesian Knowledge Tracing models to provide a more accurate representation of student latent knowledge. Furthermore, the system can detect when students are off-track, and is capable of providing individualized support as students work. This is particularly important since receiving one-on-one help in a typical classroom setting can be difficult, and furthermore, students themselves may not recognize when they are having difficulty (Aleven et al., 2004). The system as a whole thus blends learning activities and assessment activities, because students practice and receive help as they work, and teachers/parents can keep track of their progress. As such, this work represents an embodiment of a learner-centered environment (Bransford, Brown & Cocking, 2000; Quellmalz & Pellegrino, 2009; Quellmalz et al, 2012; Siemens, 2012) capable of providing detailed information about students' inquiry proces skills and sub-skills[11].

An important area not addressed in this work was researching how best to integrate this technology into classrooms and existing science curricula (cf. Heffernan & Koedinger, 2012). The environment and approach we developed is flexible; for example, the materials could be used as pretests, stand-alone assessments of inquiry, inquiry-based learning activities to aid content understanding, or for more general inquiry practice. The materials could be used in class or as "smart homework" (Mendicino,

[11] Though the focus of this work is specifically on data collection skills, our research group assesses and provides automated feedback on other skills related to hypothesizing, analyzing data and warranting claims with data (Gobert et al., 2012).

Heffernan & Razzaq, 2009; Singh et al., 2011). Generally speaking, we envision that students use our inquiry environment having first studied a science topic and having already been exposed to aspects of inquiry in their regular science class. Thus, the activities double as inquiry practice and content reinforcement, particularly since the activities are aligned to Massachusetts' content standards (Gobert et al., 2012). However, more research is needed to address the "best-practices" for use of this technology and where it best fits in a curriculum.

## 8.3 General Implications for Assessment of Ill-Defined Skills

Our approach for assessing inquiry is a realization of a more general approach for assessing complex skills within open-ended learning environments. To concretize this idea, the discussion is framed in terms of Evidence-Centered Design, a principled framework for designing assessments (Messick, 1994; Mislevy et al., 2001; Mislevy et al., 2012; Pellegrino et al., 2001), as done in Gobert et al. (2012). At a terse level, this framework suggests that assessment design should entail (in part) identification of skills to be measured in a *student model*, design of learning activities in a *task model* to elicit demonstration those skills, and an *evidence model* to specify how student responses in the tasks indicate proficiency at the skills. We focus here on the student and evidence models developed in this work.

As described in Gobert et al. (2012), we first took broadly defined inquiry skills listed in the NSES (NRC, 1996) and identified key subskills that could be measured and tracked to create the student model for Inq-ITS. This work focused on two subskills related to the NSES strand "designing and conducting experiments", designing controlled experiments and testing stated hypotheses. Then, data mining techniques were used to construct detectors to evaluate students' performance. BKT models were used to aggregate performance as evidence of those skills. Most notable in this work is the use of data mining to create detectors (models) that infer whether students demonstrate skills as part of our

evidence model. As mentioned earlier, this approach was chosen to address the issue that the skills themselves are ill-defined, because there is no single "right" or "wrong" way to demonstrate them. As such, our work supports using data mining in the evidence model as a viable approach for evaluating ill-defined skills in open-ended learning environments (e.g. Stevens et al., 2004; Rowe & Lester, 2010; Shores et al., 2011; Sil et al., 2012; Kerr & Chung, 2012; Baker & Clarke-Midura, in press).

It is important to note that constructing data-mined models, particularly the detectors, was a complex endeavor. The process entailed 1) operationalizing the skills to discern what to look for in students' log files to indicate skill demonstration, 2) collecting student usage data, 3) hand-coding that data and achieving inter-rater reliability, 4) understanding prior work to indentify indicators of inquiry skills, and 5) developing, refining and testing the predictive power of those models. Thus, there are many considerations and "moving parts", all of which impact the interpretability, performance, and generalizability of such models. Relevant to this dissertation, we showed how judicious feature selection can improve models' predictive power and reuse (Sao Pedro et al., 2012a), and how stratification can reveal model limitations (Sao Pedro et al., 2013a). There are additional areas to consider which may affect model performance, generalizability, and interpretability. One example is to understand how variability of students' actions in the training data impacts the robustness of the detectors. For example, we hypothesized in Section 4.4 that the detectors' performance may have suffered because there were fewer clips with exactly two simulation runs, underrepresenting a kind of variability in student actions. A second example is to estimate bounds on how many training instances or students are required to produce models that predict and generalize well. Our detectors were constructed starting from 148 students' data (Section 4.1.1); it is possible we could construct generalizable detectors using less data. Researching factors such as these will further address the potential of leveraging data mining for evaluating ill-defined skills.

## 8.4 Additional Future Work

Aside from the possibilities already mentioned to further address the applicability of these approaches for formative and summative assessment of inquiry, this work limited its scope to focus on two specific inquiry process skills within Physical Science simulations. The simulations we used had two to four manipulable variables that took on *nominal* values, all of which had causal effects on the outcomes. When simulations have more variables that have complicated interactions and continuous values students can manipulate, like Inq-ITS Life Science simulations (Gobert et al., 2012), the ways in which students demonstrate the skills of interest may be different. In a similar vein, the data collection skills were demonstrated in an environment that provided a fair amount of implicit scaffolding of the overall inquiry process. In more open-ended inquiry explorations, the nature of students' data collection may also change. All of these factors could impact the degree to which skills transfer (van Joolingen et al., 2007), and the degree to which the models can measure skills in these situations (cf. Baker et al., 2008c).

In addition, as outlined in Zimmerman (2007) and Veermans (2003, p.45), there are several additional data collection skills that come into play when simulations (or more generally inquiry contexts) are more complex. For example, when a variable is continuous and a relationship between that variable and an outcome is suspected, try extreme values of the variable to determine any limits on the relationship (Schunn & Anderson, 1999). Similarly, trying incremental values of a continuous variable can identify limitations and possible relationships (Schunn & Anderson, 1999). A related skill is to address how students' data collection changes if and when they revise their hypotheses based on prior data collections in order to find relationships (van Joolingen & de Jong, 1993). Tracking skills such as these will be important as new simulations with continuous variables that have complex interactions between them are used.

There are other considerations as well that are of importance when measuring and scaffolding inquiry. We note that our work focuses more on procedural aspects of inquiry, not the conceptual or

epistemological aspects, which are also of great importance. In other words, the "how" and "why" one engages in inquiry is just as important when assessing and teaching inquiry as can one "do" inquiry. There are other factors as well that may help students acquire and retain inquiry skills. For example, prior work has shown that self-explanation plays a role in acquiring skills and knowledge (e.g. Chi et al., 1989; 1994), even for data collection skills (Sao Pedro et al., 2010b, 2011). Such reflection activities may could be incorporated in environments to promote learning.

Finally, there are many other inquiry skills and science practices such as analyzing data, warranting claims, understanding of models, and making predictions (NSES, 1996; NRC, 2011) that were not addressed in this dissertation. Inq-ITS addresses some of these already, namely hypothesizing, analyzing data and warranting claims (Gobert et al., 2012). Peformance on these other inquiry skills can also be triangulated with performance on data collection skills to possibly improve assessment and can better understand developmental interdependenices between skills. In summary, incorporating measurement of more inquiry skills across science topics, determining the applicability of our models to these new scenarios, and addressing the conceptual and epistomolgical aspects of inquiry will pave the way for more encompassing, scalable, computer-based formative assessments of inquiry.

# 9 Bibliography


Aleven, V., & Koedinger, K. (2000). Limitations of Student Control: Do Students Know When They Need Help? In G. Gauthier, C. Frasson, & K. VanLehn (Ed.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000* (pp. 292-303). Berlin: Springer-Verlag.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward Tutoring Help Seeking. In J. C. Lester, R. M. Vicari, & F. Paraguaçu (Ed.), *Proceedings of the Seventh International Conference on Intelligent Tutoring Systems, ITS 2004* (pp. 227-239). Berlin: Springer-Verlag.

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research, 73*(3), 277-320.

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.

Amershi, S., & Conati, C. (2009). Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining, 1*(1), 71-81.

Anderson, J. (1993). *Rules of the Mind.* Hillsdale, NJ: Erlbaum.

Anderson, J., Boyle, C., Corbett, A., & Lewis, M. (1990). Cognitive Modeling and Intelligent Tutoring. *Artificial Intelligence, 42*, 7-49.

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

Aulls, M. (2002). The Contributions of Co-Occurring Forms of Classroom Discourse and academic Activities to Curriculum Events and Instruction. *Journal of Educational Psychology, 94(3)*, 520-538.

Baker, R. S., Mitrovic, A., & Mathews, M. (2010a). 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., & Clarke-Midura, J. (in press). Predicting Successful Inquiry Learning in a Virtual Performance Assessment. *To appear in Proceedings of the 21st International Conference on User Modeling, Adaptation and Personalization.* Rome, Italy.

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., & Aleven, V. (2008a). Improving Contextual Models of Guessing and Slipping with a Truncated Training Set. In R. S. Baker, T. Barnes, & J. E. Beck (Ed.), *Proceedings of the 1st International Conference on Educational Data Mining, EDM 2008*, (pp. 67-76). Montreal, Quebec, Canada.

Baker, R., Corbett, A., & Aleven, V. (2008b). More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge-Tracing. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Ed.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems, ITS 2008. LNCS 5091*, pp. 406-415. Montreal, Quebec, Canada: Springer-Verlag.

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., Corbett, A., Gowda, S., Wagner, A., MacLaren, B., Kauffman, L., et al. (2010b). Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. 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. 52-63. Big Island of Hawaii, HI: Springer-Verlag.

Baker, R., Corbett, A., Koedinger, K., & Wagner, A. (2004). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". In E. Dykstra-Erickson, & M. Tscheligi (Ed.), *Proceedings of ACM CHI 2004: Computer-Human Interaction* (pp. 383-390). Vienna, Austria: ACM Press.

Baker, R., Corbett, A., Roll, I., & Koedinger, K. (2008c). Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction, 18*(3), 287-314.

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.

Barrow, L. (2006). A Brief History of Inquiry: From Dewey to Standards. *Journal of Science Teacher Education, 17*, 265-278.

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

Beck, J. (2005). Engagement Tracing: Using Response Times to Model Student Disengagement. In C. K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Ed.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education, AIED 2005* (pp. 88-95). Amsterdam, Netherlands: IOS Press.

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.

Beck, J., Chang, K., Mostow, J., & Corbett, A. (2008). Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Ed.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, *LNCS 5091*, pp. 383-394. Montreal, QC.

Ben-David, A. (2008). About the Relationship between ROC Curves and Cohen's Kappa. *Engineering Applications of Artificial Intelligence, 21*, 874-882.

Bereiter, C., & Scardamalia, M. (1993). *Surpassing Ourselves: An Inquiry into the Nature and Implications of Expertise.* La Salle, IL: Open Court.

Bernardini, A., & Conati, C. (2010). Discovering and Recognizing Student Interaction Patterns in Exploratory Learning Environments. In V. Aleven, J. Kay, & J. Mostow (Ed.), *Proceedings of the 10th International Conference of Intelligent Tutoring Systems, ITS 2010, Part 1* (pp. 125-134). Berlin Heidelberg: Springer-Verlag.

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

Blikstein, P. (2011). Using Learning Analytics to Assess Students' Behavior in Open-Ended Programming Tasks. In P. Long, G. Siemens, G. Conole, & D. Gasevic (Ed.), *Proceedings of the 1st International Conference on Learning Analytics and Knowledge (LAK2011)*, (pp. 110-116). Banff, Alberta, CA.

Bloom, B. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher, 13*(6), 4-16.

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.

Bransford, J., Brown, A., & Cocking, R. (2000). *How People Learn: Brain, Mind, Experience, and School.* Washington, DC: National Academy Press.

Brown, A., & Campione, J. (1994). Guided discovery in a community of learners. In K. M. (Ed.), *Classroom lessons: integrating cognitive theory and classroom practice.* Cambridge, Massachusetts: MIT Press.

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.

Cetintas, S., Si, L., Xin, Y., & Hord, C. (2010). Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques. *IEEE Transactions on Learning Technologies, 3*(3), 228-236.

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.

Chi, M., VanLehn, K., & Litman, D. (2010). Do Micro-Level Tutorial Decisions Matter: Applying Reinforcement Learning to Induce Pedagogical Tutorial Tactics. In V. Aleven, J. Kay, & J. Mostow (Ed.), *ITS 2010, Part I, LNCS 6094* (pp. 224-234). Berlin Heidelberg: Springer-Verlag.

Clarke-Midura, J., Dede, C., & Norton, J. (2011). *The Road Ahead for State Assessments.* Policy Analysis for California Education and Rennie Center for Educational Research & Policy, Cambridge, MA.

Cocea, M., & Weibelzahl, S. (2009). Log File Analysis for Disengagement Detection in e-Learning Environments. *User Modeling and User-Adapted Interaction, 19*, 341-385.

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.

Cree, V., & Macaulay, C. (2000). *Transfer of Learning in Professional and Vocational Education.* London: Routledge Psychology Press.

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.

de Jong, T., van Joolingen, W., Giemza, A., Girault, I., Hoppe, U., Kindermann, J., et al. (2010). Learning by creating and exchanging objects: The SCY experience. *British Journal of Educational Technology, 41*(6), 909-921.

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

Dehn, D., & van Mulken, S. (2000). The Impact of Animated Interface Agents: A Review of Empirical Research. *International Journal of Human-Computer Studies, 52*(1), 1-22.

Desmarais, M., & Baker, R. (2012). A Review of Recent Advances in Learner and Skill Modeling in Intelligent Learning Environments. *User Modeling and User-Adapted Interaction, 22*(1-2), 9-38.

Dunbar, K. (1993). Concept Discovery in a Scientific Domain. *Cognitive Science, 17*, 397-434.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8.* Washington, D.C.: National Academies Press.

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

Eltinge, E., & Roberts, C. (1993). Linguistic Content Analysis: A Method to Measure Science Inquiry in Textbooks. *Journal of Research Science Teaching, 30*, 65-83.

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.

Fadel, C., Honey, M., & Pasnick, S. (2007). Assessment in the Age of Innovation. *Education Week, Volume 26 (38)*, 34-40.

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.

Fogarty, J., Baker, R., & Hudson, S. (2005). Case Studies in the Use of ROC Curve Analysis for Sensor-Based Estimates in Human Computer Interaction. *Proceedings of Graphics Inerface, GI 2005* (pp. 129-136). Victoria, British Columbia: Canadian Human-Computer Communications Society.

Frederiksen, J., & White, B. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition and Instruction, 16(1)*, 3-118.

Gertner, A., & VanLehn, K. (2000). Andes: A Coached Problem Solving Environment for Physics. In G. Gauthier, C. Frasson, & K. VanLehn (Ed.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000. LNCS 1839*, pp. 133-142. Montreal, Quebec, Canada: Springer-Verlag.

Ghazarian, A., & Noorhosseini, S. M. (2010). Automatic Detection of Users' Skill Levels Using High-Frequency User Interface Events. *User Modeling and User-Adapted Interaction, 20*(2), 109-146.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1991). Scientific Reasoning Across Different Domains. In E. DeCorte, M. Linn, H. Mandl, & L. Verschaffel (Eds.), *Computer-based Learning Environments and Problem-Solving* (pp. 345-371). Heidelberg, Germany: Springer-Verlag.

Gobert, J. (2005). Leveraging Technology and Cognitive Theory on Visualization to Promote Students' Science Learning and Literacy. In J. Gilbert, *Visualization in Science Education* (pp. 73-90). Dordrecht, The Netherlands: Springer-Verlag.

Gobert, J., & Buckley, B. (2000). Introduction to Model-Based Teaching and Learning in Science Education. *International Journal of Science Education, 22*(9), 891-894.

Gobert, J., & Koedinger, K. (2011). *Using Model-Tracing to Conduct Performance Assessment of Students' Inquiry Skills within a Microworld.* Paper presented at the Society for Research on Educational Effectiveness (SREE).

Gobert, J., & Schunn, C. (2007). Supporting Inquiry Learning: A Comparative Look at What Matters. *A symposium presented at the Annual Meeting of the American Educational Research Association. Chicago, IL, April 9-13.*

Gobert, J., Buckley, B., Levy, S., & Wilensky, U. (2007, April 12). Teasing Apart Domain-Specific and Domain-General Inquiry Skills: Co-evolution, Bootstrapping, or Separate Paths? *Paper presented at the annual meeting of the American Educational Research Association.* Chicago, IL: Retrieved April 18, 2011, from the AERA Online Paper Repository.

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. (under review). From Log Files to Assessment Metrics for Science Inquiry using Educational Data Mining. *Journal of the Learning Sciences*.

Gong, Y., Beck, J., & Heffernan, N. (2010). Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting Procedures. In V. Aleven, J. Kay, & J. Mostow (Ed.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, (pp. 35-44). Pittsburgh, PA.

Gotwals, A., & Songer, N. (2006). Measuring Students' Scientific Content and Inquiry Reasoning. In S. Barab, K. Hay, & D. Hickey (Ed.), *Proceedings of the 7th International Conference of the Learning Sciences, ICLS 2006* (pp. 196-202). Bloomington, IN: Lawrence Erlbuam Associates.

Graesser, A., Chipman, P., Haynes, B., & Olney, A. (2005). AutoTutor: An Intelligent Tutoring System with Mixed-Initiative Dialogue. *IEEE Transactions on Education, 48*(4), 612-618.

Gu, T., Wu, Z., Tao, X., Pung, H. K., & Lu, J. (2009). epSICAR: An Emerging Patterns based approach to sequential, interleaved and Concurrent Activity Recognition. *Proceedings of the 2009 IEEE*

*International Conference on Pervasive Computing and Communications, PERCOM '09* (pp. 1-9). Galveston, TX: IEEE Computer Society.

Guzdial, M. (1995). Software-Realized Scaffolding to Facilitate Programming for Science Learning. *Interactive Learning Environments, 4*(1), 1-44.

Hair, J., Anderson, R., Tatham, R., & Black, W. (1998). *Multivariate Data Analysis (5th Edition).* Upper Saddle River, NJ: Prentice Hall.

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.

Heffernan, N., & Koedinger, K. (2012). Integrating Assessment Within Instruction: A Look Forward. *Paper presented at the Invitational Research Symposium on Technology Enhanced Assessment*.

Heffernan, N., Turner, T., Lourenco, A., Macasek, M., Nuzzo-Jones, G., & Koedinger, K. (2006). The ASSISTment builder: Towards an analysis of cost effectiveness of ITS creation. *Proceedings of the 19th International FLAIRS Conference*, (pp. 515-520). Melbourne Beach, Florida, USA.

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.

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

Kanari, Z., & Millar, R. (2004). Reasoning from Data: How Students Collect and Interpret Data in Science Investigations. *Journal of Research in Science Teaching, 41*(7), 748-769.

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.

Kerr, D., & Chung, G. (2012). Identifying Key Features of Student Performance in Educational Video Games and Simulations through Cluster Analysis. *Journal of Educational Data Mining, 4*(1), 144-182.

Ketelhut, D., Nelson, B., Schifter, C., & Kim, Y. (2010). Using Immersive Virtual Environments to Assess Science Content Understanding: The Impact of Context. In D. Kinshuk, J. Sampson, P. I. Spector, & R. Vasiu (Ed.), *Proceedings of the Iadis International Conference on Cognition and Exploratory Learning in the Digital Age*, (pp. 227-230). Timisoara, Romania.

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.

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

Koedinger, K., Suthers, D., & Forbus, K. (1999). Component-Based Construction of a Science Learning Space. *International Journal of Artificial Intelligence in Education (IJAIED), 10*, 292-313.

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., Black, J., Keselman, A., & Kaplan, D. (2000). The Development of Cognitive Skills to Support Inquiry Learning. *Cognition and Instruction, 18*(4), 495-523.

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

Kumar, R., & Rose, C. (2011). Architecture for Building Conversational Agents that Support Collaborative Learning. *IEEE Transactions on Learning, 4*(1), 21-34.

Lester, J., Towns, S., & Callaway, C. (1997). Cosmo: A Life-like Animated Pedagogical Agent with Deictic Believability. *Working Notes of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent*, (pp. 61-69). Nagoya, Japan.

Lester, J., Towns, S., & Fitzgerald, P. (1999). Achieving Affective Impact: Visual Emotive Communication in Lifelike Pedagogical Agents. *International Journal of Artificial Intelligence in Education, 10*(3-4), 278-291.

Levy, S., & Wilensky, U. (2006, April 11). Emerging Knowledge through an Emergent Perspective: High-school Students' Inquiry, Exploration and Learning in the Connected Chemistry Curriculum. *Presented at the annual meeting of the American Educational Research Association.* San Francisco, CA: Retrieved April 18, 2011 from the AERA Online Paper Repository.

Linn, M., Davis, E., & Bell, P. (2004). *Internet Environments for Science Education.* Mahwah, NJ: Erlbaum.

Linn, R. (1994). Performance Assessment: Policy Promises and Technical Measurement Standards. *Educational Researcher, 23*(9), 4-14.

Lynch, C., Ashley, K., Pinkwart, N., & Aleven, V. (2009). Concepts, Structures, and Goals: Redefining Ill-Definedness. *International Journal of Artificial Intelligence in Education, 19*, 253-266.

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

Massachusetts Department of Education. (2008). *Massachusetts Department of Elementary and Secondary Education (ESE)*. Retrieved 10 2008, from http://profiles.doe.mass.edu

Mayer, R. (2004). Should There Be a Three Strikes Rule Against Pure Discovery? The Case for Guided Methods of Instruction. *American Psychologist, 59*(1), 14-19.

Mayer, R., Dow, G., & Mayer, S. (2003). Multimedia Learning in an Interactive Self-Explaining Environment: What Works in the Design of Agent-Based Microworlds? *Journal of Educational Psychology, 95*(4), 806-812.

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.

Mendicino, M., Razzaq, L., & Heffernan, N. (2009). Comparison of Traditional Homework with Computer Supported Homework. *Journal of Research on Technology in Education, 41*(3), 331-359.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13-23.

Minstrell, J., & van Zee, E. (2000). *Inquiring into Inquiry Learning and Teaching in Science.* Washington, DC: American Association for Advancement of Science.

Mislevy, R., Behrens, J., DiCerbo, K., & Levy, R. (2012). Design and Discovery in Educational Assessment: Evidence Centered Design, Psychometrics, and Data Mining. *Journal of Educational Data Mining, 4*, 11-48.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Mitrovic, A. (2003). An Intelligent SQL Tutor on the Web. *International Journal of Artificial Intelligence in Education, 13*(2-4), 173-197.

Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a Constraint-Based Tutor for a Database Language. *International Journal of Artificial Intelligence in Education, 14*(3-4), 235-278.

Mitrovic, A., & Suraweera, P. (2000). Evaluating an Animated Pedagogical Agent. *Intelligent Tutoring Systems, 1839*, 73-28.

Mitrovic, A., Koedinger, K., & Martin, B. (2003). A Comparative Analysis of Cognitive Tutoring and Constraint-Based Modeling. In P. Brusilovsky, A. Corbett, & F. de Rosis (Ed.), *Proceedings of the 9th International Conference on User Modeling*, (pp. 313-322). Pittsburgh, PA.

Mitrovic, A., Mayo, M., Suraweera, P., & Martin, B. (2001). Constraint-Based Tutors: A Success Story. In L. Monostori, J. Vancza, & M. Ali (Ed.), *Proceedings of the 14th International Conference on Industrial and Engineering Application of Artificial Intelligence and Expert Systems:*

*Engineering of Intelligent Systems, IEA/AIE-2001. LNCS 2070*, pp. 931-940. Budapest, Hungary: Springer-Verlag.

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.

Moreno, R. (2005). Multimedia Learning with Animated Pedagogical Agents. In R. Mayer, *The Cambridge Handbook of Multimedia Learning* (pp. 507-524). New York: Cambridge University Press.

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. (2006). *Systems for State Science Assessment.* (M. Wilson, & M. Bertenthal, Eds.) Washington, DC: The National Academies Press.

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

Njoo, M., & de Jong, T. (1993). Exploratory Learning with a Computer Simulations for Control Theory: Learning Processes and Instructional Support. *Journal of Research in Science Teaching, 30*, 821-844.

Oh, I.-S., Lee, J.-S., & Moon, B.-R. (2004, November). Hybrid Genetic Algorithms for Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(11), 1424-1437.

Ohlsson, S., & Mitrovic, A. (2006). Constraint-based Knowledge Representation for Individualized Instruction. *Computer Science and Information Systems, 3*, 1-22.

Papert, S. (1980). Computer-based Microworlds as Incubators for Powerful Ideas. In R. Taylor, *The Computer in the School: Tutor, Tool, Tutee* (pp. 203-201). New York, NY: Teacher's College Press.

Papert, S. (1993). *Mindstorms: Children, Comptuers, and Powerful Ideas, 2nd Edition.* New York: Basic Books.

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

Pardos, Z., & Heffernan, N. (2011). KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. *Proceedings of the 19th International Conference on User Modeling, Adaptation and Personalization*, (pp. 243-254). Girona, Spain.

Pardos, Z., Baker, R., Gowda, S., & Heffernan, N. (2011). The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software. *SIGKDD Explorations, 13*(2), 37-44.

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

Pardos, Z., Heffernan, N., Anderson, B., & Heffernan, C. (2010). Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. In C. Romero, S. Ventura, S. R. Viola, M. Pechenizkiy, & R. S. Baker, *Handbook of Educational Data Mining* (pp. 417-426). Boca Raton, FL: CRC Press.

Pavlik, P., Cen, H., & Koedinger, J. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Ed.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009* (pp. 531-540). Brighton, UK: IOS Press.

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

Penner, D. (2000/2001). Cognition, Computers, and Synthetic Science: Building Knowledge and Meaning through Modeling. *Review of Research in Education, 25*, 1-35.

President's Council of Advisors on Science and Technology (PCAST). (2010). *Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Math (STEM) for America's Future.* Washington, DC: Office of Science and Technology Policy, Executive Office of the President.

Pudil, P., Novovicova, J., & Kittler, J. (1994). Floating Search Methods in Feature Selection. *Pattern Recognition Letters, 15*(11), 1119-1125.

Quellmalz, E., Kreikemeier, P., DeBarger, A., & Haertel, G. (2007). A Study of the Alignment of NAEP, TIMSS, and New Standards Science Assessments with the Inquiry Abilities in the National Science Education Standards. Chicago, IL: Paper presented at the Annual Meeting of the American Educational Research Association.

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

Quellmalz, E., Timms, M., Silberglitt, M., & Buckley, B. (2012). Science Assessments for All: Integrating Science Simulations into Balanced State Science Assessment Systems. *Journal of Research in Science Teaching, 49*(3), 363-393.

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

Quinn, J., & Alessi, S. (1994). The Effects of Simulation Complexity and Hypothesis Generation Strategy on Learning. *Journal of Research on Computing in Education, 27*, 75-91.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., et al. (2004). A Scaffolding Design Framework for Software to Support Science Inquiry. *The Journal of the Learning Sciences, 13(3)*, 337-386.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N., Koedinger, K. R., Junker, B., et al. (2005). The Assistment Project: Blending Assessment and Assisting. In G. M. C.K. Looi, *Proceedings of the 12th International Conference on Artificial Intelligence In Education* (pp. 555-562). Amsterdam: ISO Press.

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

Reiser, B. (2004). Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. *The Journal of the Learning Sciences, 13*(3), 273-304.

Reye, J. (2004). Student Modeling Based on Belief Networks. *International Journal of Artificial Intelligence in Education, 14*(1), 1-33.

Rickel, J., & Johnson, W. (1999). Animated Agents for Procedural Training in Virtual reality: Perception, Cognition, and Motor Control. *Applied Artificial Intelligence, 13*, 343-382.

Rieber, L., Tzeng, S., & Tribble, K. (2004). Discovery Learning, Representation, and Explanation within a Computer-Based Simulation: Finding the Right Mix. *Learning and Instruction, 14*, 307-323.

Ritter, S., Harris, T., Nixon, T., Dickinson, D., Murray, R., & Towle, B. (2009). Reducing the Knowledge-Tracing Space. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Ed.), *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009*, (pp. 151-160). Cordoba, Spain.

Roll, I., Aleven, V., & Koedinger, K. (2010). The Invention Lab: Using a Hybrid of Model Tracing and Constraint-Based Modeling to Offer Intelligent Support in Inquiry Environments. In V. Aleven, J. Kay, & J. Mostow (Ed.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems, ITS 2010* (pp. 115-124). Pittsburgh, PA: Springer-Verlag.

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.

Rowe, J., & Lester, J. (2010). Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments. In C. G. Youngblood, & V. Bulitko (Ed.), *Proceedings of the 6th Annual AI and Interactive Digital Entertainment Conference, AIIDE 2010* (pp. 57-62). Palo Alto, CA: AAAI Press.

Ruiz-Primo, M., & Shavelson, R. (1996). Rhetoric and reality in science performance assessment. *Journal of Research in Science Teaching, 33(10)*, 1045-1063.

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. (2010a). 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. (2010b). 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. (2012a). Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. In J. Masthoff, B. Mobasher, M. Desmarais, & R. Nkambou (Ed.), *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP 2012)*, (pp. 249-260). Montreal, QC, Canada.

Sao Pedro, M., Baker, R., & Gobert, J. (2013a). 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.* Leuven, Belgium.

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (2013b). 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., Gobert, J., & Baker, R. (2012b). Assessing the Learning and Transfer of Data Collection Inquiry Skills Using Educational Data Mining on Students' Log Files. *Paper presented at The Annual Meeting of the American Educational Research Association.* , (pp. Retrieved April 15, 2012, from the AERA Online Paper Repository). Vancouver, BC, Canada.

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. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31-57.

Schauble, L. (1996). The Development of Scientific Reasoning in Knowledge-Rich Contexts. *Developmental Psychology, 32*(1), 102-119.

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, Inc.

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

Settlage, J., & Southerland, S. (2007). *Teaching Science to Every Child: Using Culture as a Starting Point.* New York, NY: Routledge Taylor & Francis Group.

Shavelson, R., Wiley, E. W., & Ruiz-Primo, M. (1999). Note On Sources of Sampling Variability in Science Performance Assessments. *Journal of Educational Measurement, 36(1)*, 61-71.

Shores, L., Rowe, J., & Lester, J. (2011). Early Prediction of Cognitive Tool Use in Narrative-Centered Learning Environments. *Proceedings of the 15th International Conference on Artificial Intelligence in Education, AIED 2011*, (p. to appear). Auckland, New Zealand.

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.

Siemens, G. (2012). Learning Analytics: Envisioning a Research Discipline and a Domain of Practice. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK2012)* (pp. 4-8). Banff, Alberta, CA: ACM.

Sil, A., Shelton, A., Ketelhut, D., & Yates, A. (2012). Automatic Grading of Scientific Inquiry. *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, (pp. 22-32). Montreal, QC.

Siler, S., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010). Predictors of Transfer of Experimental Design Skills in Elementary and Middle School Children. In V. Aleven, J. Kay, & J. Mostow (Ed.), *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems, ITS 2010. Part II, LNCS 6095*, pp. 198-208. Pittsburgh, PA: Springer.

Singh, R., Saleem, M., Pradhan, P., Heffernan, C., Heffernan, N., Razzaq, L., et al. (2011). Improving K-12 Homework with Computers. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Ed.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, (pp. 328-336). Auckland, NZ.

Singley, M., & Anderson, J. (1989). *The Transfer of Cognitive Skill.* Cambridge, MA: Harvard University Press.

Slotta, J., & Linn, M. (2009). *WISE Science: Web-Based Inquiry in the Classroom.* New York, NY: Teachers College Press.

Stecher, B., & Klein, S. (1997, Spring). The Cost of Science Performance Assessments in Large-Scale Testing Programs. *Educational Evaluation and Policy Analysis, 19*(1), 1-14.

Stevens, R., Soller, A., Cooper, M., & Sprang, M. (2004). Modeling the Development of Problem Solving Skills in Chemistry with a Web-Based Tutor. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Ed.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems, ITS 2004. LNCS 3220*, pp. 580-591. Maceio, Alagoas, Brazil: Springer.

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

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

Timms, M., Clements, D., Gobert, J., Ketelhut, D., Lester, J., Reese, D., et al. (2012). *New Measurement Paradigms.* Vancouver, BC: Community for Advancing Discovery Research in Education.

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

U.S. Department of Education. (2010). *National Educational Technology Plan 2010.* Washington, DC: U.S. Department of Education.

van Joolingen, W. (1999). Cognitive Tools for Discovery Learning. *International Journal of Artificial Intelligence in Education, 10*, 385-397.

van Joolingen, W. R., & de Jong, T. (2003). SimQuest, Authoring Educational Simulations. In T. Murray, S. Blessing, & S. Ainsworth, *Authoring Tools for Advanced Technology Learning Environments: Toward Cost-effective Adaptive, Interactive, and Intelligent Educational Software* (pp. 1-31). Dordrecht: Kluwer.

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.

VanLehn, K. (2006). The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227-265.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., et al. (2005). The Andes physics tutoring system: Lessons Learned. *International Journal of Artificial Intelligence and Education, 15(3)*, 1-47.

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

Voss, J. (2006). Toulmin's Model and the Solving of Ill-Structured Problems. In D. Hitchcock, & B. Verheij, *Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation* (pp. 303-311). Berlin: Springer.

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

Walker, J., Sproull, L., & Subramani, R. (1994). Using a Human Face in an Interface. *CHI '94 Human Factors in Computing Systems*, (pp. 85-91). Boston, MA.

Walonoski, J., & Heffernan, N. (2006). Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In M. Ikeda, K. Ashlay, & T.-W. Chan (Ed.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS 2006. LNCS 4053*, pp. 382-391. Johngli, Taiwan: Springer-Verlag.

Weinbrenner, S., Engler, J., Wichmann, A., & Hoppe, U. (2010). Monitoring and Analysing Students' Systematic Behaviour - The SCY Pedagogical Agent Framework. In M. Wolpers, P. Kirschner, M. Scheffel, S. Lindstaedt, & D. V. (Ed.), *Proceedings of the Fifth European Conference on Technology Enhanced Learning, EC-TEL 2010. LNCS 6383*, pp. 602-607. Barcelona, Spain: Springer.

White, B. (1993). ThinkerTools: Causal Models, Conceptual Change, and Science Education. *Cognition and Instruction, 10*(1), 1-100.

White, B., & Frederiksen, J. (1998). Inquiry, Modeling and Metacognition: Making Science Accesible to All Students. *Cognition and Instruction, 16(1)*, 3-118.

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

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition.* San Francisco: Morgan Kaufmann.

Wixon, M., Baker, R., Gobert, J., Ocumpaugh, J., & Bachmann, M. (2012). WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. In J. Masthoff, B. Mobasher, M. Desmarais, & R. Nkambou (Ed.), *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization (UMAP 2012)*, (pp. 286-296).

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, (pp. 856-863). Washington DC.

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

Zohar, A., & David, A. B. (2008). Explicit Teaching of Meta-Strategic Knowledge in Authentic Classroom Situations. *Metacognition Learning, 3*, 59-82.

# Appendix A: Scaffolds and Help Options

| Constraint | Triggered Scaffold Message | Help Button | Help Response |
|---|---|---|---|
| Tried to finish collecting data with only a single trial collected | Level 1:<br>I don't think you can test your hypothesis with only one trial. Try collecting more data. | Why? | You can't tell if the **[IV]** affects the **[DV]**, because you only tried one value for the **[IV]**. |
| | | What should I do? | Try to **design a controlled experiment to test your hypothesis**. I'll jump in and try to help if I think you need it. |
| | Level 2:<br>With only one trial, you can't tell if the **[IV]** affects the **[DV]**, because you only tried one value for the **[IV]**. | What should I do? | Try to **design a controlled experiment to test your hypothesis**. I'll jump in and try to help if I think you need it. |
| | Level 3:<br>**Run some more trials** and try to **design a controlled experiment to test your hypothesis**. I'll jump in and try to help if I think you need it. | | |
| During data collection, never changes any variables by 3rd trial | Level 1:<br>Try changing the variables before running your next trial. Then, I'll help you test your hypothesis. | | |
| | Level 2:<br>You still haven't changed any variables. I'll try and help you more after you change some variables. | | |
| 3rd time the same trial has been run | Level 1:<br>Did you know you already ran a trial using these variables?<br><br>The **table** has all the results of your trials, so you don't need to run this trial again. | | |
| | Level 2:<br>You already ran this trial. The results of it are in the table.<br><br>You don't need to run it again. | | |

| Constraint | Triggered Scaffold Message | Help Button | Help Response |
|---|---|---|---|
| Not designing controlled experiments and not testing stated hypotheses | Level 1:<br>I think the data you're collecting won't help you test your hypothesis because **you aren't designing a controlled experiment**. | How do I do that? | Design a controlled experiment by **changing only the variable you are testing** while **keeping all the other variables the same**. |
| | | Which variable am I trying to test? | It's in your hypothesis. It says you want to test if changing the **[IV]** affects the **[DV]**. |
| | | I need more help | Run trials where you:<br>(1) Change only the [IV], and<br>(2) Keep all the other variables the same. |
| | | Why do this? | Changing only the [IV] while keeping everything else the same lets you tell for sure if the [IV] affects the [DV]. |
| | Level 2:<br>Let me help you some more.<br><br>You said you wanted to test if changing the **[IV]** affects the **[DV]** in your hypothesis.<br><br>To do this, run pairs of trials where you:<br>(1) Change only the [IV], and<br>(2) Keep all the other variables the same. | Why do this? | Changing only the [IV] and keeping everything else the same lets you tell for sure if the [IV] affects the [DV]. |
| | Let me help you some more.<br><br>Just change the **[IV]** and run another trial. Don't change the other variables.<br><br>Doing this lets you tell for sure if changing the [IV] affects the [DV]. | | |

| Constraint | Triggered Scaffold Message | Help Button | Help Response |
|---|---|---|---|
| Designing controlled experiments, but not testing stated hypotheses | Level 1:<br>It looks like you did great at designing a controlled experiment, but let me remind you to **collect data to help you test your hypothesis**. | How do I do that? | Keep designing a controlled experiment, but make sure to **try different values of variable you're trying to test**. |
| | | Which variable is that? | Your hypothesis says you wanted to test if changing the **[IV]** affects the **[DV]**. |
| | | Why do this? | Changing the [IV] while keeping everything else the same lets you see how changing the [IV] affects the [DV]. |
| | Level 2:<br>Let me help again.<br><br>You said you wanted to test if changing the **[IV]** affects the **[DV]** in your hypothesis.<br><br>Keep designing controlled experiments, and **collect data for different values of the IV**. | Why do this? | Changing the [IV] while keeping everything else the same lets you see how changing the [IV] affects the [DV]. |
| | Level 3:<br>Let me help some more.<br><br>Just change the **[IV]** and run another trial. Don't change the other variables.<br><br>Doing this lets you tell for sure if changing the [IV] causes changes to the [DV]. | | |

| Constraint | Triggered Scaffold Message | Help Button | Help Response |
|---|---|---|---|
| Testing stated hypotheses, but not designing controlled experiments | Level 1:<br>I see you're collecting data about the [IV], but you can't test your hypothesis because you **aren't designing a controlled experiment**. | How do I do that? | Design a controlled experiment by **changing only the variable you are testing** while **keeping all the other variables the same**. |
| | | Which variable am I trying to test? | It's in your hypothesis. It says you want to test if changing the **[IV]** affects the **[DV]**. |
| | | I need more help | Run trials where you:<br>(1) Change only the **[IV]**, and<br>(2) Keep all the other variables the same. |
| | | Why do this? | Changing only the [IV] while keeping everything else the same lets you tell for sure if the [IV] affects the [DV]. |
| | Level 2:<br>Let me help again.<br><br>You said you wanted to test if changing the **[IV]** affects the **[DV]** in your hypothesis.<br><br>To do this, run pairs of trials and **change only the [IV]**. **Keep all the other variables the same** for the second trial. | Why do this? | Changing only the [IV] and keeping everything else the same lets you tell for sure if the [IV] affects the [DV]. |
| | Level 3:<br>Let me help you some more.<br><br>Just change the **[IV]** and run another trial. Don't change the other variables.<br><br>Doing this lets you tell for sure if changing the [IV] affects the [DV]. | | |