DESIGN, DEVELOPMENT, AND EVALUATION OF SCAFFOLDS FOR DATA INTERPRETATION PRACTICES DURING INQUIRY

by

Raha Moussavi-Aghdam

A Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Learning Sciences and Technologies

May 2018

APPROVED:

Dr. Janice Gobert, Advisor
Professor, Rutgers University Graduate School of Education

Dr. Ryan Baker
Associate Professor, University of Pennsylvania Graduate School of Education

Dr. Ivon Arroyo
Assistant Professor, Worcester Polytechnic Institute

Dr. Michael Sao Pedro
CTO & Co-Founder, Apprendis LLC

Abstract

Developing explanations is a key inquiry practice in national science standards (NGSS Lead States, 2013) and essential for learning science content (McNeill & Krajcik, 2011) and is conceptualized as consisting of three aspects: claims, evidence, and reasoning (Toulmin, 1958). However, students often have difficulty with these tasks (McNeill & Krajcik, 2011; Schunn & Anderson, 1999). Prior work by our group (Sao Pedro et al., 2014) has shown that auto-scaffolding in Inq-ITS (Inquiry Intelligent Tutoring System; Gobert et al., 2013) can help students acquire inquiry skills and transfer them to a new science topic. These data provide a rationale for the work presented, namely, designing, developing, and evaluating a real-time scaffolding approach for the development of the inquiry practices specifically for data interpretation and warranting claims, which, to us, underlie the explanation practices necessary for communicating science findings. Unpacking these practices can help us better understand, assess, and, in turn, scaffold them. Specifically, this work addresses the: (1) design of scaffolds for data interpretation practices; (2) efficacy of scaffolds for supporting these practices using a modified Bayesian Knowledge Tracing framework that captures the complexities of science inquiry, and (3) transfer of these practices within one science topic to another. Results from this work show that the developed scaffolds were effective in aiding students' acquisition and transfer of the assessed practices. As such, this research builds on prior work on the nature of explanation (McNeill & Krajcik, 2011) as well as prior work on the assessment and scaffolding of science inquiry skills (Gobert et al, 2013; Sao Pedro et al., 2014).

Acknowledgements

I would like to thank all the people who helped and supported me throughout this endeavor.

I would first like to thank my advisor, Janice Gobert. My experience working with you as an undergraduate and your recommendation for me to return to WPI for my graduate work are the main reasons I started this entire journey. Thank you for your advocacy, support, expertise, and over the past years.

I would also like to thank my dissertation committee – Ryan Baker, Ivon Arroyo, and Mike Sao Pedro – for all their help in navigating this project and this dissertation. You provided thoughtful expertise and feedback and helped push my thinking and expand my knowledge and skills.

Thanks also to the whole Inq-ITS team (past and present) for all your good humor, friendship, and help. You helped make this process fun! As an additional input of good humor and fun - thank you to my friends for your emotional support and encouraging words.

Last, but not least (!), I want to thank my family – my dad (Massoud), mom (Farkhondeh), and my sisters (Mehrnoush & Ava). You never stopped believing in me and supporting me through the natural ups and downs I went through. Your unconditional love and support has been a major part of my life and played such a big role in my journey to becoming a Dr. Thank you!!

Table of Contents

Tables

Figures

Figure 23: School 1 - Student Performance in the Free Fall Microworld on the Warranting Claims practice/skill (green = Rex condition, blue = No Rex condition)

Figure 24: School 2, Grade 8 - Student Performance in the Density Microworld (green = Rex condition, blue = No Rex condition)

1 – Introduction

The work presented here presents a real-time scaffolding approach for the development of the inquiry skills/practices (used here interchangeably) for data interpretation and warranting claims, which, to us, underlie the explanation practices necessary for communicating science findings. This research builds on prior work on the nature of explanation (McNeill & Krajcik, 2011), which is important due to the fact that developing explanations is a key inquiry practice in national science standards (NGSS Lead States, 2013) and essential for learning science content (McNeill & Krajcik, 2011). However, the current work and literature on explanation has gaps in assessing and supporting in real-time the development of the underlying practices of explanation, which is the focus of this work.

Real-time assessment and scaffolding support is necessary because students often have difficulty with aspects of explanation, defined by Toulmin (1958) as consisting of: claims, evidence, and reasoning; others concur (McNeill & Krajcik, 2011; Gotwals & Songer, 2009; Kang, Thompson, & Windschitl, 2014; Berland & Reiser, 2009). Previous studies have found that students have difficulty using appropriate and sufficient data and providing reasoning for their claims (McNeill & Krajcik, 2011); linking their data to their claims (Schunn & Anderson, 1999); changing ideas about causality (Kuhn, Schauble, & Garcia-Mila, 1992); relating outcomes of experiments to theories being tested (Schunn & Anderson, 1999); and relying on theoretical arguments rather than on experimental evidence (Schunn & Anderson, 1999; Kuhn, 1991). These are just a few examples of the documented findings on students' difficulty with explanation, which will be more fully discussed in a later chapter, but these examples make it clear that many of students' difficulties in explanation have to do with the components of claim (McNeill & Krajcik, 2001; Kuhn, Schauble, & Garcia-Mila, 1992) and evidence (Schunn & Anderson, 1999;

Kuhn, 1991); two components that underlie what we refer to as data interpretation and warranting.

By unpacking and concretizing the underlying skills/practices (i.e. subskills) of data interpretation and warranting claims within the frame of explanation, a hole in the current literature, we can better understand the procedural difficulties students make in order to assess students' mastery. It will also allow us to give targeted help that allows students to master these skills/practices. This is especially important since analyzing data and warranting claims are conceptualized as sub-components of the explanation processes and students must be able to master these subcomponents of explanation before they can successfully do explanation as a whole.

Scaffolding, in particular, has been shown to successfully support students on the skills/practices of data interpretation and warranting and lead to student mastery. Specifically, using automatic, real-time scaffolding can help facilitate the acquisition of these subskills by preventing students from engaging in ineffective behaviors (Buckley et al., 2006; Gobert & Schunn, 2007); offering help even if they may not be aware that they need help (Aleven & Koedinger, 2000; Aleven et al., 2004); improving and supporting learning (Schauble, 1990; de Jong, 2006); personalizing the learning process by adapting to students' behaviors by both providing multiple levels of automatic scaffolds and allowing students to request for further help or clarification, as needed; and providing guidance that can be scaled to many learners. Additionally, prior work by our group (Sao Pedro et al., 2014) has shown that auto-scaffolding in Inq-ITS (Inquiry Intelligent Tutoring System; Gobert et al., 2013) can help students acquire two data collection inquiry practices/skills and transfer them to a new science topic.

However, one of the main drawbacks to the scaffolding presented in a number of learning environments (e.g. Belland, 2010; Kim & Hannifin, 2011; Gijlers & de Jong, 2009) is that they are provided in the form of worksheets or in some other similar form that is neither scalable nor fine-grained with regard to the skills/practices it addresses. The work here builds on this prior work and addresses these drawbacks in scaffolding approaches by using a computer-based scaffolding approach aimed at inquiry subskills so that the needed support can be determined and utilized automatically in real-time, which should better support students in their inquiry.

In summary, the work presented here builds on prior work and addresses the gaps in the literature on explanation subskills as well as scaffolding approaches for supporting these subskills. In doing so, this work presents a real-time scaffolding approach for the development of the inquiry practices for data interpretation and warranting claims, which, to us, underlie explanation. More specifically, this research builds on our prior work on the assessment and scaffolding of science inquiry practices/skills in an intelligent tutoring system (Gobert et al, 2013; Sao Pedro et al., 2014) by extending it to the inquiry skills/practices of data interpretation and warranting with the goal of assessing and scaffolding these skills/practices in real-time in the system. In doing so, we make progress on our goal to have scaffolds for the full complement of inquiry practices outlined by the NGSS.

Specifically, this work addresses the: (1) operationalization of the underlying subskills of data interpretation and warranting claims, (2) design of scaffolds for data interpretation and warranting practices/skills; (3) efficacy of scaffolds for supporting these practices/skills, and (4) transfer of these practices within domains.

For Study 1, the skills/practices of data interpretation and warranting were operationalized into subskills, aligning with previous research and literature on student difficulties in this area.

These (elaborated on in study 1) include difficulty with: creating a claim that relates to the hypothesis, selecting appropriate data, creating a claim that reflects the data, and identifying if the claim supports the hypothesis. Scaffolds were then designed based on these difficulties. These scaffolds were pilot-tested one-on-one with middle school students to iteratively refine the scaffolds. Study 2 comprised of an assessment of students' performance on data interpretation and warranting with and without scaffolding. Study 3 focuses on assessing the transfer of data interpretation and warranting practices/skills to a new activity within a domain.

2 – Explanation and Data Interpretation

As previously mentioned, this work builds on the nature of explanation, an important practice in inquiry and a core part of national science standards. The meaning of scientific explanation is under conceptualized in the literature on science inquiry as well as in the Next Generation Science Standards (NGSS), specifically with regards to the difference between explanation and argumentation (Kang, Thompson, & Windschitl, 2014), and the language to describe it is not always consistent (McNeill & Krajcik, 2011). Because of this, in this work, we do not claim to differentiate between explanation and argumentation and instead use explanation to refer to the practices of creating a claim, using evidence, and reasoning.

Two of the inquiry practices/skills involved in the process of explanation, creating a claim and using evidence, are conceptualized by us as data interpretation and warranting subskills and considered to be integral to state standards and students' understanding of science content and science literacy. Appropriately, data interpretation and warranting is one of the four stages of inquiry assessed in Inq-ITS and is the focus of this work.

This chapter will discuss the benefits students experience when practicing explanation and will briefly present our approach for operationalizing the data interpretation and warranting claims subskills contained within explanation, which will be more fully described in Study 1.

2.1 Why is Explanation Important?

Not only is the practice of developing and critiquing scientific explanations essential for scientists, it is also important and highly beneficial for students learning science inquiry (Bricker & Bell, 2008; McNeill & Krajcik, 2011; National Research Council, 2011). Science educators and researchers recognize the importance of engaging students in epistemic practices such as explanation (Duschl, Schweingruber, & Shouse, 2007; Bricker & Bell, 2008) and these practices are increasingly viewed as a leading educational goal for science education (Bell, 2004; Duschl, Schweingruber, & Shouse, 2007; Bricker & Bell, 2008). McNeill and Krajcik (2011) concretize the benefits of engaging with explanation into five points; specifically, explanation can help students: (1) understand science concepts, (2) develop 21st century skills, (3) use evidence to support claims, (4) reason logically, and (5) understand the nature of science.

The educational benefits that can result from engagement in explanation include improvements in content learning (Chinn & Clark, 2013). In fact, the development of explanations is seen as critical for the construction of knowledge, specifically to producing, evaluating, and advancing scientific knowledge (McNeill & Krajcik, 2007; Bricker & Bell, 2008). When writing explanations to answer a question, students have to apply scientific ideas and use appropriate evidence. This allows students to make sense of their data and enriches their understanding of the scientific concepts by helping them make connections and reason about a phenomenon (McNeill & Krajcik, 2011). Similarly, when students argue to learn, they engage in

explanation for the purpose of mastering the content about which they are explaining/arguing (Chinn & Clark, 2013). Therefore, one of the benefits of explanation is to have students deeply learn content (Chinn & Clark, 2013).

Explanation also helps students develop important 21st century skills, such as science literacy, by requiring them to communicate with others, engage in problem solving, and do self-monitoring (McNeill & Krajcik, 2011). Another benefit of explanation is that it requires students to use the evidence they have collected to support claims, which students seldom do in their science classrooms, and with which they can have difficulty (McNeill & Krajcik, 2011). These difficulties will be further addressed in a later section, as they are critical to the work presented here.

The development of explanations is also critical for fostering student reasoning (Schauble, Glaser, Duschl, Schulze, & John, 1995) and explanation is thought to be an essential kind of informal reasoning that is central to solving problems, making judgments and decisions, and formulating ideas and beliefs (Kuhn, 1991). As such, many science educators agree that student-constructed explanations can be used as evidence of deep learning since the process of generating an explanation requires students to engage in reasoning from data (Kang, Thompson, & Windschitl, 2014). Furthermore, when students engage in explanation, it not only improves their ability to reason, but also allows them to become more proficient at analytical thinking (McNeill & Krajcik, 2011).

While the presented literature makes the benefits of engaging with explanation clear, what is unclear is the understanding of the different subcomponents within explanation. In order to effectively promote student understanding of and mastery in explanation, it is important to identify and separate out these subcomponents of explanation, which also allows us to provide a

more nuanced approach to their assessment and scaffolding as that is the aim of the work presented here.

2.2 Data Interpretation Inquiry Subskills as a Part of Explanation

The construction of explanation is a fundamental feature of science inquiry and it allows students to engage in the epistemic practices of inquiry (Kang, Thompson, & Windschitl, 2014). The importance of developing, using, and critiquing scientific explanation, evidence, and argument are prevalent throughout the standards (McNeill & Krajcik, 2011). In fact, of the five essential features of classroom inquiry listed in NRC's Inquiry and the National Education Standards (2011), three are related to the development of explanation; these are: (1) formulating explanations from evidence, (2) giving priority to evidence, and (3) communicating and justifying explanations (McNeill & Krajcik, 2011).

These categories align with Toulmin's (1958) model of argumentation, which many researchers (McNeill & Krajcik, 2011; Gotwals & Songer, 2009; Kang, Thompson, & Windschitl, 2014; Berland & Reiser, 2009) use to identify the aspects of explanation (both in writing and spoken discourse), defined as the use of claims, evidence, and reasoning. As such, explanation is a process of making claims and providing justifications for the claims using evidence (Carr, 1999; Voss, Lawrence, & Engle, 1991; Lajoie et al., 2001). These two underlying practices of explanation are framed here as being akin to data interpretation and warranting practices/skills.

It is important to operationalize the data interpretation and warranting subskills that are within the overall practice of explanation because students must be able to address the subcomponents of explanation – such as analyzing data and warranting claims – before they can

successfully do explanation (Berland & Hammer, 2012). If students are having specific, concrete problems with analyzing their data, creating a claim, or warranting their claim, then they will not be able to successfully engage in explanation as a practice/skill expected by NGSS (2013). Therefore, unpacking and concretizing the subskills of data interpretation and warranting claims within the frame of explanation gives us the nuance to examine these complex practices and allows us to look specifically at the aspects with which students are having difficulty.

2.3 Difficulty with Data Interpretation Practices/Skills

For this work, a literature review was conducted to examine the specific difficulties students have with data interpretation and warranting as subskills of explanation. This review led to the categorization of four types of "buggy" data interpretation and warranting behaviors, listed below.

1. Claim IV/DV does not match Hypothesis IV/DV
2. The claim does not reflect the data selected
3. The data selected are not properly controlled
4. Claim is incorrectly said to support/not support the hypothesis

These "buggy" behaviors then led to the generation of the data interpretation and warranting subskills, which will be further discussed in Study 1.

2.4 Domain-Generality and Transfer of Data Interpretation and Warranting

Domain-general knowledge refers to a set of skills, concepts, and strategies that are necessary for a person to adapt to cross-curricular problem environments (McNeill & Krajcik, 2009; Greiff et al., 2014). This encompasses general cognitive skills or problem-solving abilities and involves

exploring and understanding, formulation, planning and executing, and monitoring and reflecting (Greiff et al., 2014). In science, domain-general expertise is tied to reasoning or inquiry practices that can be used across scientific domains, such as designing experiments, controlling variables, constructing arguments, and using evidence to support claims (Schunn & Anderson, 1999; Zimmerman, 2000; Kuhn et al., 1992; McNeill & Krajcik, 2009). Conversely, domain-specific knowledge is more related to the concepts and strategies developed within specific content domains.

This is mirrored by the dimensions of learning in NGSS where domain-specific knowledge is represented by disciplinary core ideas that are grouped in four domains – the physical sciences, the life sciences, the earth and space sciences, and engineering, technology, and applications of science – and domain-general knowledge is represented by practices, which are inquiry practices. As such, while domain-specific strategies are important and have explanatory power for solving specific problems that are domain-bound, domain-general problem-solving skills are helpful in problem situations across domains and are also helpful when it comes to working with new problem situations (Greiff et al., 2014).

When it comes to categorizing specific skills as domain-general or domain-specific, there is some debate as to how separate or distinct the categories of domain-generality and domain-specificity are (Greiff et al., 2014; McNeill & Krajcik, 2009; Schunn & Anderson, 1999). Many researchers believe that there is a continuum from domain-general knowledge to domain-specific knowledge, rather than a dichotomy (Kuhn et al., 1992; Greiff et al., 2014; McNeill & Krajcik, 2009; Schunn & Anderson, 1999). As such, despite the importance of context and domain familiarity (Harrison & Schunn, 2004; Lazonder et al., 2010; Bassok & Holyoak, 1989; Kuhn, 1989), the acquisition of content within domains does not wholly explain inquiry skill

development and transfer (Kuhn et al., 1992; Harrison & Schunn, 2004). Instead, there is thought to be a mutual influence of both domain knowledge and inquiry/reasoning practices/skills (Lazonder et al., 2010).

Furthermore, prior research has made a case that a variety of types of knowledge – domain-general, domain-specific, and in-between – is important for successfully engaging in reasoning or inquiry tasks (McNeill & Krajcik, 2009; Duschl et al., 2007; Gotwals & Songer, 2006; Schunn & Anderson, 1999; Zimmerman, 2000) and that the practices/skills involved in inquiry have aspects of both domain-generality and specificity (Kuhn et al., 1992). For example, successfully doing inquiry requires one to both know how to use evidence as well as understand the content (McNeill & Krajcik, 2009).

3 – Scientific Inquiry

3.1 Scientific Inquiry in Practice

Science education should help students participate in a science-infused world (McGinn & Roth, 1999) and this can be done through the use of inquiry practices in the classroom (McNeill & Krajcik, 2011). When scientific inquiry is used in the classroom, students have experiences through which they can anchor their understanding of scientific phenomena and are therefore better able to interact with the content they learn (McNeill & Krajcik, 2011). Therefore, the goals for inquiry-based science teaching include learning to inquire in order to construct scientific knowledge (Yore et al., 2007; de Jong & van Joolingen, 1998; Klahr & Dunbar, 1988). This consists of incorporating key scientific inquiry practices such as: asking/formulating questions, developing hypotheses, designing experiments, analyzing data, drawing conclusions, and constructing explanations (AAAS, 1993; National Research Council, 2011; de Jong & van

Joolingen, 1998; Klahr & Dunbar, 1988). If students are guided in the classroom towards the modeling of these practices that are inherent in the inquiry process of expert scientists, then they are more likely to benefit from inquiry (de Jong & van Joolingen, 1998; Klahr & Dunbar, 1988).

However, despite the fact that cultivating these inquiry practices/skills is increasingly seen as important to students' science education (NGSS Lead States, 2013), these practices/skills are complex and students have difficulties with them (Gobert et al., 2012; Gobert et al., 2013; Schauble, Klopfer, & Raghavan, 1991; Hsu, Lai, & Hsu, 2015; McNeill & Krajcik, 2011). While prior research has shown that young children are capable of conducting inquiry, there are many subtasks and ideas – such as detecting complex patterns, distinguishing between causal and non-causal relationships, using models to represent problems, etc. – that students must navigate as they continue to develop expertise in inquiry (Windschitl, 2000). These more complex inquiry tasks may be difficult for students to grasp as they may have a fragmented understanding of the purposes and mechanisms of inquiry and may hold confounding ideas regarding the process of scientific inquiry (Schauble, Klopfer, & Raghavan, 1991). As such, it is important that students have access to the appropriate learning materials and tools that will support them as they continue to develop and acquire their inquiry practices/skills.

Furthermore, despite the emphasis on inquiry and inquiry assessments in the classroom (NRC, 2011), inquiry teaching is difficult for teachers to successfully implement and it is difficult for them to assess their students' mastery of inquiry skills/practices (de Jong et al., 2005; Anderson, 2002). This is due to the inherent complexity in conducting inquiry, as it requires students to apply and coordinate multiple skills (both cognitive and metacognitive) simultaneously. This, in turn, complicates assessment of inquiry, making it challenging for

teachers to create assessment tasks for these practices/skills and/or understand how their students are struggling (Gobert et al. 2013; Fadel, Honey, & Pasnick, 2007; McNeill & Krajcik, 2011).

Because of this, in many classrooms, science teaching and learning often defaults to rote vocabulary, facts, and formulas and more often than not, inquiry is assessed using paper and pencil items such as short answer tests (Alonzo & Aschbacher, 2004). Even in many studies (Kuhn, 1991; Schunn & Anderson, 1999; McNeill & Krajcik, 2011), researchers have used either verbal protocols or written responses to measure these practices/skills, which requires coding and is both labor-intensive and requires the use of inquiry progressions and rubrics (Gotwals & Songer, 2009; McNeill & Krajcik, 2011). However, important to this work, it is unclear whether these types of written assessments can adequately identify inquiry practices/skills (Black, 1999; Pellegrino et al., 2001) and, furthermore, they have not lead to scalable assessment or scalable support for students.

More recently, assessing and scaffolding students' inquiry practices/skills via performance-based assessments has become possible with the technological capabilities of computers and, in particular, the application of educational data mined algorithms (cf. Sao Pedro et al., 2010) to students' data including log file data generated as students work. The environment used in this research, Inq-ITS (Inquiry Intelligent Tutoring System) uses both knowledge-engineered and educational data-mined models to automatically evaluate the inquiry work products students create and inquiry processes students follow. In the system, we operationalize each of the three over-arching inquiry practices/skills – hypothesizing, data collection, and data interpretation and warranting – into subskills and we then use those subskills as discrete measures of inquiry skills (Gobert et al., 2012); this approach will be further discussed in Chapter 4. Conceptualizing and supporting the components of the explanation

framework in an automated and fine-grained way with appropriate subskills helps us unpack and target known difficulties documented by previous research in a scalable way (Gotwals & Songer, 2009; McNeill & Krajcik, 2011; Schunn & Anderson, 1999).

3.2 Scientific Inquiry in Frameworks

Nationally, the Next Generation Science Standards (NGSS), which have either been adopted or adapted by 38 states across the country and the District of Columbia, provide information on how students best learn scientific concepts, as well as set the standards for what students of every grade from kindergarten to 12th grade should learn in their science curriculum (NRC, 2011). These standards provide a three-dimensional view of what it means to be proficient in science. Each of the three dimensions is as follows: Dimension 1 – Practices, Dimension 2 – Crosscutting Concepts, and Dimension 3 – Disciplinary Core Ideas (National Research Council, 2011).

The first dimension of practices encompasses scientific processes or inquiry practices/skills that should be integrated into the science curriculum. Not only does the NGSS framework expect that students be able to use scientific inquiry techniques, but it also expects that these techniques be used for the investigation of natural phenomena and for solving meaningful problems. The framework lists eight essential practices (National Research Council, 2011):

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking

6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

As noted, this work focuses on the Data Interpretation stage of inquiry in the Inq-ITS environment. This stage encompasses the practices/skills of analyzing and interpreting data, constructing a (explanation or) claim, and supporting that claim through the use of collected evidence. All of these practices/skills, aside from being a natural part of inquiry and, as we conceptualize them, important to the practices of argumentation/explanation, are also stressed by the NGSS Frameworks as essential inquiry practices/skills that students are expected to be able to understand and be competent at. More specifically, two of the eight practices for K-12 science classrooms set out by NGSS are "analyzing and interpreting data" and "engaging in argument from evidence." These practices map most directly to the skills/practices of interpreting data and warranting claims, which are practiced and assessed in Inq-ITS.

4 – Inq-ITS

The environment used in this work is Inq-ITS, Inquiry Intelligent Tutoring System. Addressing the need for science inquiry support, Inq-ITS is a computer based tutoring and assessment system for science inquiry practices/skills. Inq-ITS consists of different interactive microworlds, or virtual labs, for different topics in Physical, Life, and Earth Science. Inq-ITS aims to automatically assess and scaffold middle school students' inquiry as they experiment with interactive simulations. As such, Inq-ITS encompasses the entire inquiry process. Within each microworld in Inq-ITS, students progress through the inquiry stages by articulating a testable

hypothesis, “experimenting” by collecting data, analyzing their data, and communicate their findings. This process is shown in Figure 1 below and subsequently described in detail.

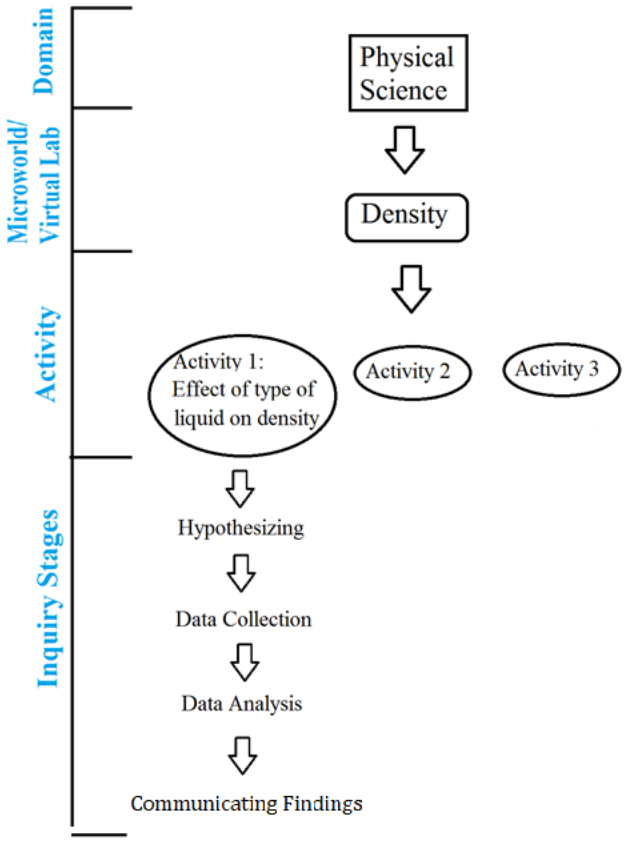


Figure 1: Inquiry in Inq-ITS

Within an activity, students first attempt to construct a hypothesis they can test by using the hypothesis widget. Students choose among a list of relevant independent variables and dependent variables (for example, in the density microworld, students would choose amongst amount of liquid, density of the liquid, shape of the container, and type of liquid) from the drop-down widgets when creating a hypothesis. Students are expected to create a testable hypothesis

with an independent variable they can change and a dependent variable that can be measured, even if it is scientifically inaccurate.

After hypothesizing, students collect data by designing and running experiments with the simulation. As part of the simulation, students collect trials by choosing which independent variable they want to change. After collecting data, students then proceed to data interpretation. This stage of inquiry, which is the focus of this work, is described in more detail below and illustrated in Figure 2.

During the data interpretation stage, students are reminded at the top of the page of their goal for the experiment as well as their hypothesis. They then use the data they collected during experimentation, which is presented for them in a table at the bottom of the page, to create a claim. In their claim, students must state an appropriate IV and an appropriate DV as well as the relationship they discovered between those two variables. For example, in the Density virtual lab pictured in Figure 2, students choose amongst amount of liquid, density of the liquid, shape of the container, and type of liquid for their IV and their DV from the drop-down widgets when creating a hypothesis. This way, a student could make a claim such as: “When I changed the [amount of liquid], the [density of the liquid] did not change.” Students can also make scientifically incorrect claims in multiple ways, such as by stating an incorrect relationship between the variables (e.g. “When I changed the [amount of liquid], the [density of the liquid] changed”) or confounding the IV with the DV (e.g. “When I changed the [density of liquid], the [amount of the liquid] did not change”). As part of their claim, students also have to state whether or not their claim supports their hypothesis through a drop-down widget, which allows them to choose between “supports” and “does not support”.

During their data interpretation, students also have to use the boxes next to the list of trials they collected to select which trials they believe support their claim. Because the data table contains all the trials they collected in the previous stage, students have to choose which trials are most relevant. For example, if a student were to write the claim "When I changed the [amount of liquid], the [density of the liquid] did not change," they would then have to choose only the trials in their data table in which they have changed only the amount of liquid and held all other IVs constant in order to complete their inquiry completely.

The reminders of the goal and hypothesis, the pulldown widgets, and the evidence table all serve as implicit scaffolding during this data interpretation stage. The reminder of the goal and hypothesis allow students to remember the task at hand and what they had originally set out to investigate through their experimentation. The pulldown widgets help students by providing a sentence starter for their claim, thereby offloading some of the work it might require to remember how to write a claim as a sentence, but still focusing students on the parts of their claim that we want to assess, i.e. their ability to identify an appropriate IV and DV, state which variable is being changed, and state the relationship between the variables. We are also able to offload some of the work for the students by presenting them with their collected data already in a data table. This bypasses the chance that students may have difficulty organizing their data or creating a data table and focuses them on the task of selecting the most relevant data to warrant their claim.

However, this implicit scaffolding does not help the student at all when it comes to the inquiry skills/practice of correctly interpreting their data. To provide that type of support, more explicit scaffolding is needed that calls the students' attention to the specific subskills they are demonstrating. Inq-ITS is designed to automatically deliver explicit scaffolds to students while

they work via a pedagogical agent named Rex. By providing just-in-time scaffolding with short explanations, we can catch those students who may be having difficulty with inquiry (Schauble, 1990; de Jong, 2006) and prevent them from exhibiting any unproductive behaviors (Buckley, Gobert, & Horwitz, 2006), thereby improving and supporting learning. Prior to this work, the hypothesizing and data collection stages of the microworlds contained scaffolds that could be automatically delivered. The creation of this type of scaffolding for data interpretation and warranting was one of the goals of this work and is further described in subsequent chapters.

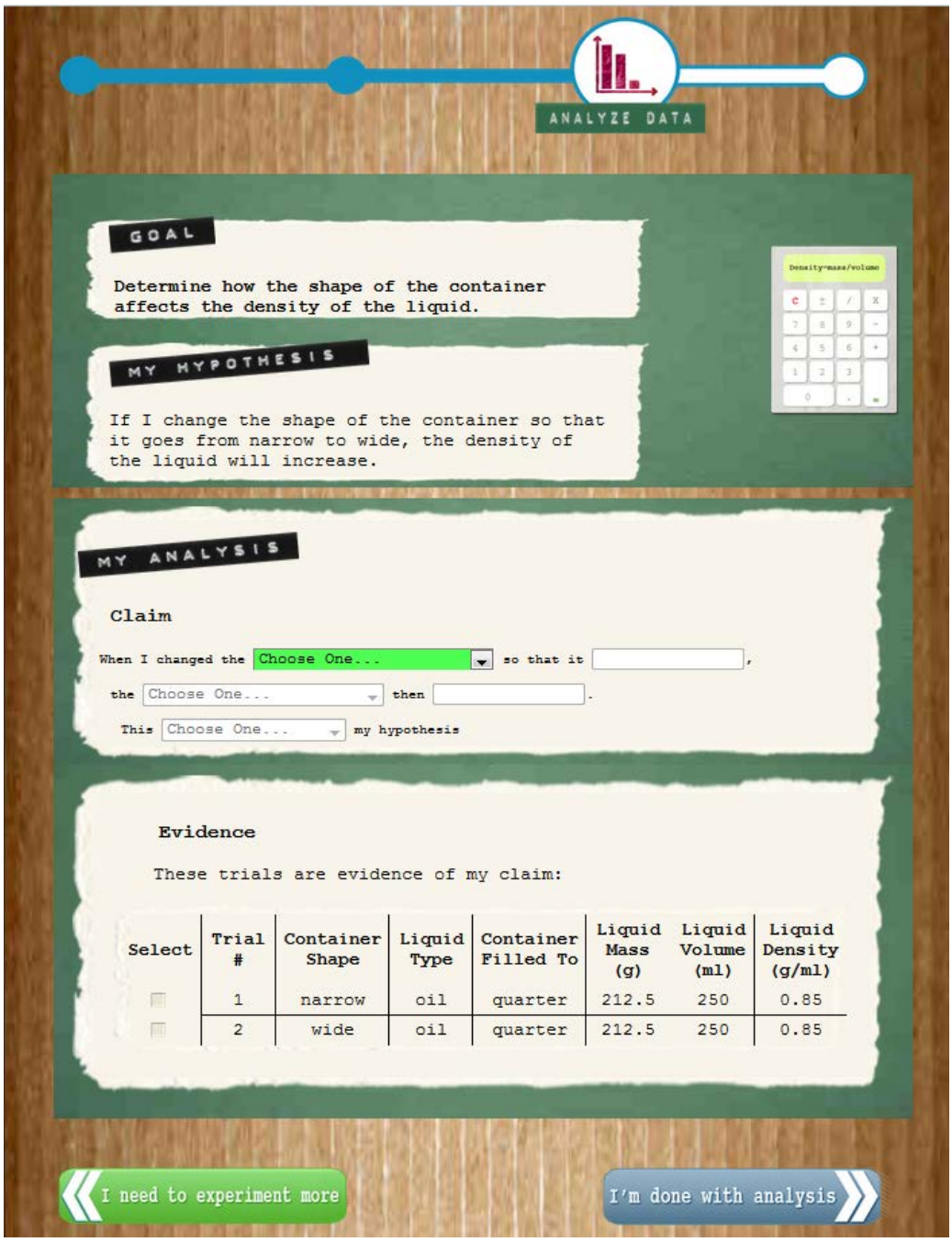


Figure 2: Analyze Data stage of the Density Microworld

Another key aspect of the Inq-ITS system is its capacity to provide an assessment of students' inquiry practices/skills. The system uses both knowledge-engineered and educational data-mined models to automatically evaluate the inquiry work products students create (e.g.

claims generated using the widget) and processes students follow (e.g. how the student interacts with the simulation when collecting data) in the microworlds to measure students' skill at inquiry. In the system, we operationalize each of the four overarching inquiry practices/skills assessed in Inq-ITS – hypothesizing, data collection, data interpretation and warranting, and communicating findings – into subskills. In total, Inq-ITS evaluates student performance on 21 different subskills within these categories of practices/skills.

In the data interpretation stage of Inq-ITS, once the student has submitted their analysis, the system uses its knowledge-engineered rules to evaluate the student's claim components (IV, DV, relationship between the two, relationship to hypothesis) and evidence selected for warranting based on pre-determined criteria. This evaluation is binary in that specifies whether a student successfully met the criteria (receiving a score of 1 for that practice/skill) or not (receiving a score of 0). All parts of a student's analysis are evaluated at once based on the aforementioned criteria. The specific subskills evaluated during the data interpretation stage as well as the criteria for each evaluation will be further discussed in Study 1.

5 – Scaffolding

While inquiry activities have long been used in classrooms, until recently, hands-on activities were used without scaffolding to help students with these complex practices/skills (Kirschner, Sweller, & Clark, 2006). However, since students practice inquiry with varying degrees and types of difficulties, instructional environments should accommodate these (Windschitl, 2000). These difficulties students face when conducting inquiry can be negotiated with the help of scaffolding (Brown, Collins, & Duguid, 1989; Reiser, 2004; Puntambekar & Hubscher, 2005; Quintana, et al., 2004).

This chapter will discuss the learning theory behind and the definition of scaffolding and will discuss different educational learning environments that target the acquisition of data interpretation and warranting practices/skills via some sort of scaffolding. The literature and prior work on scaffolding presented in this chapter forms the basis for the data interpretation scaffolding designed in this work and further discussed in Study 1.

5.1 Learning Theory

This work draws from Vygotsky's original ideas about learning (Vygotsky, 1978) as well as previous research documenting students' difficulties with data interpretation and warranting. Vygotsky's notions of scaffolding are employed to create our scaffolds that directly address several of the known buggy behaviors pertaining to the interpretation of data and warranting of claims present during science inquiry.

Vygotsky believed that conceptual structures develop due to the interaction of the mind with language and that learning itself went beyond development. He presented the idea of the zone of proximal development, which indicates what students are capable of with help versus without help. Each student has his or her own zone of proximal development, that is, each student may need a different amount of help to successfully complete each task they are attempting. Vygotsky stated that learning occurs when students are within their zone of proximal development (ZPD) and the ZPD is a necessary condition for learning to occur.

Essentially, scaffolding is support that provides students the help they need to operate within their ZPD and successfully learn. Scaffolding – as defined by Wood, Bruner, and Ross (1976) – is a just-in-time support that is provided by a teacher or a parent and that allows students to participate in and learn problem solving. Scaffolding is an effective way to support

students during problem solving because it can help to simplify processes (Reiser, 2004) and address complex processes and knowledge (Puntambekar & Hubscher, 2005).

Over the years, scaffolding has evolved from support that is provided by a teacher or a parent to support that can be provided automatically in real-time by a pedagogical agent within a virtual environment. Having scaffolds that are automatically delivered helps to support students in their inquiry processes (Schauble, 1990; de Jong, 2006) by preventing students from engaging in ineffective behaviors (Gobert & Schunn, 2007; Buckley, Gobert, & Horwitz, 2006) and offering help even if students may not be aware that they need help (Aleven & Koedinger, 2000; Aleven, McLaren, Roll, & Koedinger, 2004).

5.2 Why Scaffold Data Interpretation and Warranting

When it comes to constructing evidence-based explanations – a highly complex task that poses multidimensional challenges in understanding the task itself, planning a response, and producing representations of one's thinking (Kang, Thompson, & Windschitl, 2014) – it is important for students to receive the appropriate amount of support that will aid them in the acquisition and transfer of these practices/skills in a scalable way (Lajoie et al., 2001; Loh et al., 1997; Xun & Land, 2004). This support helps students as they coordinate inquiry process and manage complex information, which can be a problem for many students (Loh et al., 1997). Therefore, providing effective scaffolding is necessary, not optional, when trying to support students in meeting 21st century standards (Kang, Thompson, & Windschitl, 2014). When students are having difficulty with inquiry, scaffolds can help them achieve the success they could not on their own (Kang, Thompson, & Windschitl, 2014; McNeill & Krajcik, 2011).

Furthermore, because different learners encounter different kinds of challenges in the process of constructing evidence-based explanations, it is important there are levels of scaffolds based on individual student responses (Kang, Thompson, & Windschitl, 2014). Scaffolds that are developed to address specific aspects of scientific inquiry on a fine-grained level can help students receive the help they need and target the exact subskill with which they are having difficulty. Correspondingly, it is important for assessment and scaffolding to be interwoven and not treated as separate tasks (Kang, Thompson, & Windschitl, 2014) as it allows us to assess how a student is progressing and where they are having difficulty and then target those specific difficulties with scaffolding support. Furthermore, scaffolding approaches that react in real-time within a computer environment can provide scalable guidance and support the development of inquiry practices/skills by automatically detecting problems with inquiry (Sao Pedro et al., 2013; 2014), contributing to a deeper understanding of the content and inquiry processes by learners.

5.3 How to Scaffold Data Interpretation and Warranting

When it comes to scaffolding to support data interpretation and warranting practices/skills in particular, scaffolds have been found to (Land & Zembal-Saul, 2003; McNeill & Krajcik, 2011; Kirschner, Sweller, & Clark, 2006; Kang, Thompson, & Windschitl, 2014): (1) facilitate the ongoing articulation, reflection, and revision of explanations through the externalization of ideas, findings, or interpretations; (2) support explanation building and hypothesis generation by stimulating the learners to be more precise in their explanations, to provide explanations, and to connect evidence with their claims; and (3) provide opportunities for learners to organize, reflect upon, and revise their project artifacts.

There is a wide range of educational learning environments that have similar broad goals of enriching science instruction and inquiry and these environments contain two main categories of software support: (1) self-contained, content-embedded investigation environments with reflective supports built in and (2) content-neutral communicative structures designed to promote reflection (Loh et al., 1997).

Regardless of how the support is delivered, structure-oriented scaffolding around students' claim, evidence, and reasoning is the most popular form of support given for scaffolding students' explanations (Kang, Thompson, Windschitl, 2014). This type of scaffolding focuses on providing the structure of explanations (claim, evidence, and reasoning), as set forth by Toulmin, and providing students the opportunity to construct their explanations by following this template (Kang, Thompson, Windschitl, 2014). Scaffolding around students' explanation also tends to combine both conceptual and epistemic scaffolds, with conceptual scaffolds taking the form of a series of sentence frames or prompts that guide students to framing what happened during their inquiry and why (Kang, Thompson, Windschitl, 2014; Xun & Land, 2004; Kammerer, Meier, & Stahl, 2016). These prompts can support students by activating prior knowledge as well as helping them monitor and evaluate problem-solving processes by guiding their attention to specific aspects of the process and allowing them to reflect on them (Xun & Land, 2004; Kammerer, Meier, & Stahl, 2016).

This section will examine a number of different learning environments that focus on explanation, argumentation, and the data interpretation and warranting practices/skills therein with an eye towards the specific type and scalability of scaffolding support they offer. Specifically, this section will look at the following learning environments that address data interpretation and warranting inquiry practices/skills: ConnectionLog (Belland, 2010), WISE

(Linn, Clark, & Slotta, 2003; Kim & Hannafin, 2013), Galapagos Finches (Tabak et al., 2001; Kyza, 2009), KIE & Sensemaker (Bell, 2000), Belvedere (Paolucci et al., 1996), Progress Portfolio (Land & Zembal-Saul, 2003), and Seismic/Eruption (Hsu et al., 2014). A brief description of each of these learning environments is presented below.

The ConnectionLog (Belland, 2010) is a computer-based argumentation scaffold system for middle school students' that helps them construct evidence-based arguments while working in problem-based learning (PBL) units. It is a web-based system in which students are guided through the following stages: define problem, determine needed information, find needed information, organize information, develop claim, and link evidence to claim (Belland, 2010). A screenshot of the ConnectionLog can be seen in the figure below:

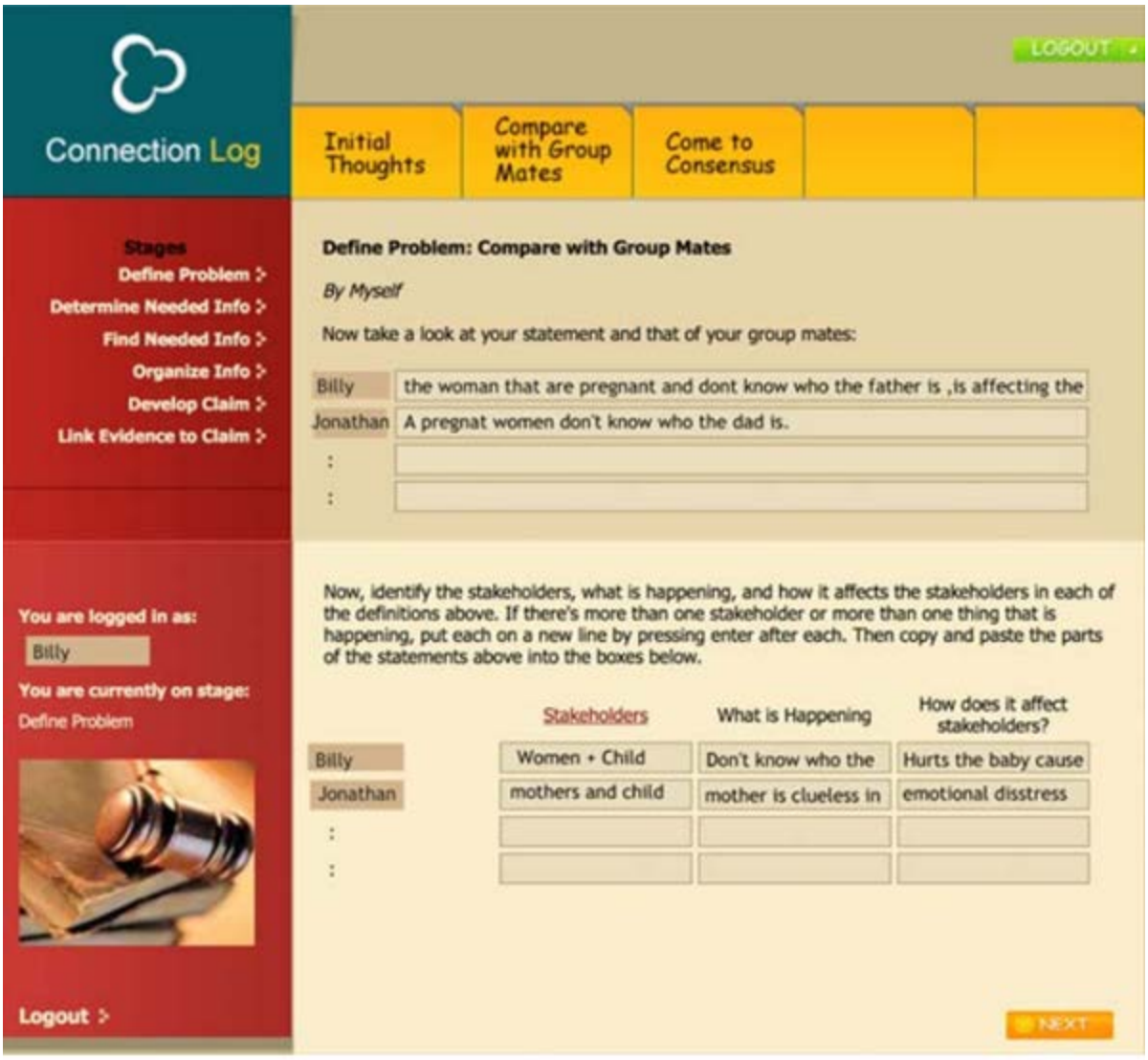


Figure 3: Screenshot of the ConnectionLog (Belland, 2010)

KIE is a technology-based learning environment used in science classrooms and the accompanying Sensemaker tools are argument representation tools that promote science learning during a debate activity (Bell, 2000). KIE and Sensemaker are typically used with middle school students and are formed around a collaborative debate project, such as: "How far does light go?" (Bell, 2000). KIE is a precursor to the next environment, WISE.

WISE (Linn, Clark, & Slotta, 2003) is an online suite of inquiry tools that allows teachers to develop units that center on a problem and that provide students with relevant materials related to the problem. In WISE, students are guided through an inquiry map and provided with evidence pages, hints, and prompts to help them monitor and reflect on their progress (Linn, Clark, & Slotta, 2003). In the study done by Kim & Hannafin (2013), a unit on wolves in WISE was used as part of a 6th grade science class. Screen captures from the WISE unit on wolves can be seen in the figure below:

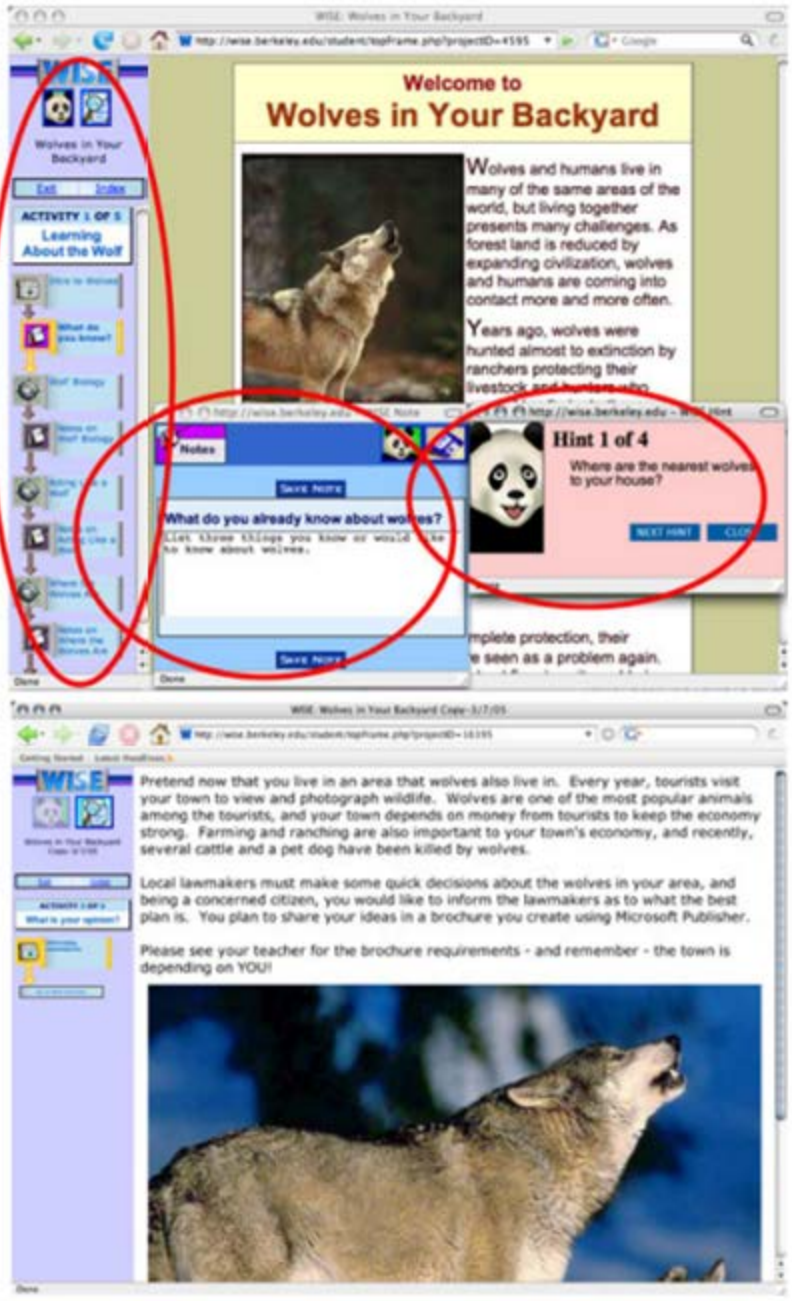


Figure 4: WISE (Linn, Clark, & Slotta, 2003; Kim & Hannafin, 2013)

In Galapagos Finches (Tabak et al., 2001), students conduct a virtual investigation in which they examine the reasons leading to the differential survival of the finch population on an

island in the Galapagos. The environment helps students engage in evolutionary inquiry in a manner similar to that of evolutionary biologists (Kyza, 2009).

The Belvedere environment is designed to provide representational and coaching support to students engaged in collaborative learning (Paolucci et al., 1996). It supports students while they conduct discussions of and solve ill-structured problems in science (Paolucci et al., 1996). Specifically, students can explore these problems by producing and manipulating graphical representations of arguments called “inquiry diagrams” (Paolucci et al., 1996). An example of such a diagram can be seen in the figure below:

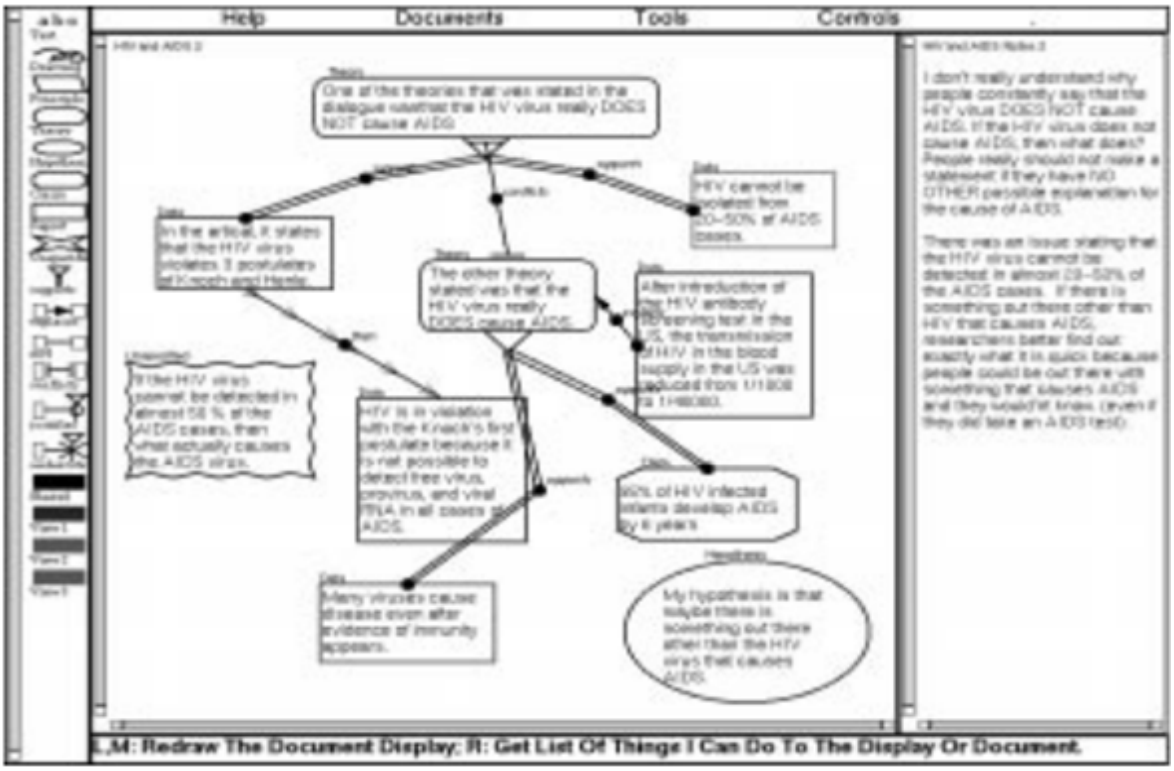


Figure 5: Belvedere Inquiry Diagram (Paolucci et al., 1996)

Progress Portfolio is a content-neutral software program that can be integrated with any learning environment to support students during inquiry (Land & Zembal-Saul, 2003). It contains

a number of "experiment pages" that prompt students to articulate experimental procedures, findings, and claims as they work through various experiment pages (Land & Zembal-Saul, 2003). An example of one of these experiment pages can be seen in the figure below:

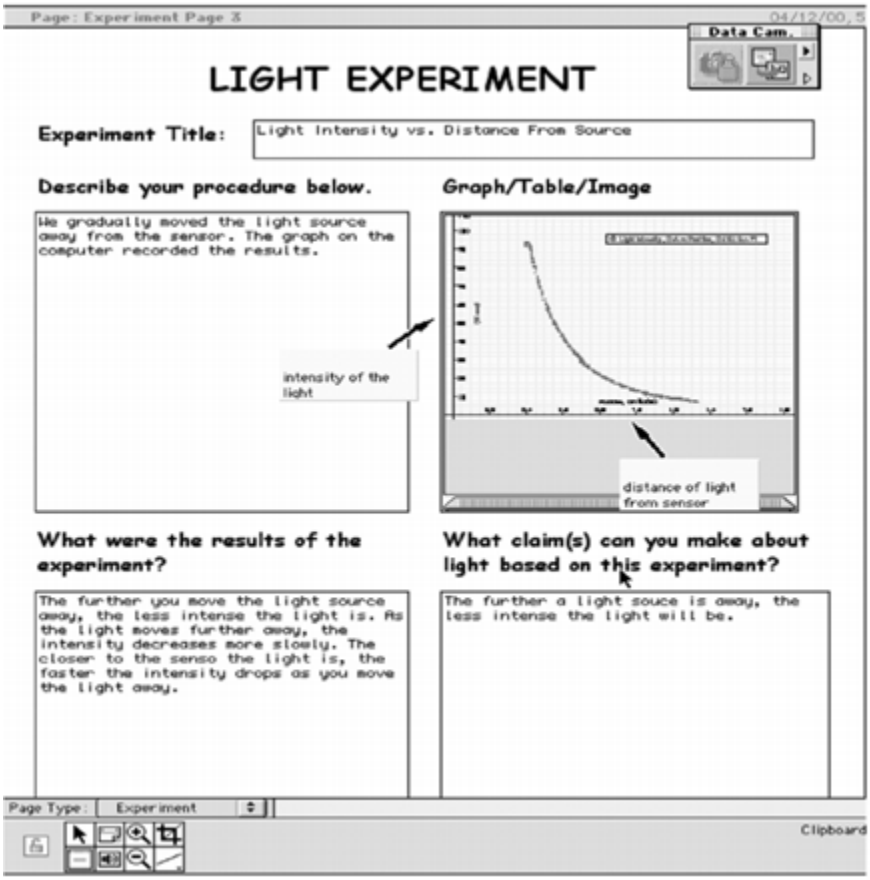


Figure 6: Example Experiment Page from Progress Portfolio (Land & Zembal-Saul, 2003)

Seismic/Eruption is a content-specific computer software package that contains inquiry lessons on plate tectonics (Hsu et al., 2014). In it, students are led through the inquiry process of posing hypotheses, collecting and analyzing data, interpreting and concluding, and, finally, reflecting (Hsu et al., 2014).

As described, all of these learning environments have been used to address science inquiry and, more specifically, have been used to address explanation, argumentation, and/or the underlying data interpretation and warranting practices/skills in some form. However, important to this work, of these seven learning environments, none of them assesses the addressed data interpretation and warranting inquiry practices/skills in real time. While they all are online learning environments, none of them have built-in assessments that can monitor students' progress in data interpretation and warranting in real-time. This then also affects the scaffolding present in these environments – since they do not have real-time automatic assessments, they cannot have real-time automatic scaffolds that are directly tied to students' difficulties.

In fact, none of the seven learning environments contains scaffolding prompts that are delivered automatically in real-time. ConnectionLog (Belland, 2010), Galapagos Finches (Tabak et al., 2001; Kyza, 2009), Progress Portfolio (Land & Zembal-Saul, 2003), and Seismic/Eruption (Hsu et al., 2014) all have implicit scaffolds that are built into the environment. These "hard" scaffolds are similar to a paper worksheet that has instruction and spaces for students to fill in, for example, their claim, evidence, and reasoning (similar to the claim-evidence-reasoning prompts given by McNeill & Krajcik, 2011). The only difference here being that instead of being presented on paper, these scaffolds are built into online environments. For example, in Connection Log, the hard scaffolds are organized in six stages, each divided into two to four steps, which guide students in following the process of building evidence-based arguments (Belland, 2010). Figure 3 shows how this process is organized for students as they work on defining the problem they are going to address in their argument. Similarly, in Progress Portfolio students work on an "explanation page" (seen in Figure 6) in which they are scaffolded around the claim-evidence-reasoning principles through the use of static prompts: (1) describe your

explanation for what happens to light below, (2) three prices of evidence, (3) explain how the evidence supports your explanation (Land & Zembal-Saul, 2003).

KIE (Bell, 2000) and Belvedere (Paolucci et al., 1996) differ in that their scaffolding prompts are not “hard” scaffolds that are built into the environment, but rather are supports that students can request to be given on-demand. For example, the pedagogical agent in KIE, Mildred, can provide hints at the level of project activities, evidence, and scientific claims (Bell & Davis, 2000). In the context of a debate project on, “How far does light go?" an example of an evidence hint is, "Why is the person in white clothes easier to see? What is happening to the light?” (Bell & Davis, 2000). An example of claim hint is, "What would happen if there were a heat source in a dark room? Would someone wearing black get hotter than someone wearing white?" (Bell & Davis, 2000). Similarly, in Belvedere there is an automated agent that gives advice on demand on how an argument in this environment can be extended or revised based on its syntactic structure or the consistency of its information. Since the actual text that a student inputs into their argument is not interpreted or assessed, this advice is phrased as either a suggestion or a question based solely on the assessed structural patterns the students have created (i.e. the boxes and arrows) (Paolucci et al., 1996). Examples of these scaffolds are: "This looks like a circular argument. Is there a statement in this group of statements that doesn't depend on accepting the rest of them?" and "Could the empirical data that supports one theory also support the other?" (Paolucci et al., 1996). In this way, these scaffolds are more automated than the hard scaffolds in the previous environments, but they still depend on the students to recognize when they need help and then ask for it, rather than automatically giving help when a student exhibits difficulties.

Another important aspect of scaffolding is the presence of levels within the prompts that serve to personalize the support and target specific difficulties students are having by giving them exactly the amount of help they need. However, of these seven learning environments, only four – WISE (Linn, Clark, & Slotta, 2003; Kim & Hannafin, 2011), KIE & Sensemaker (Bell, 2000), Belvedere (Paolucci et al., 1996), and Seismic/Eruption (Hsu et al., 2014) – offer different types of scaffolding prompts or levels within their scaffolding prompts. Furthermore, these levels of scaffolding are either not tied to the actual scientific content in the student's explanation (but rather representative of the structure of the explanation) (e.g. Paolucci et al., 1996) or are built into the system and not adaptive to the student's work. For example, in KIE & Sensemaker (Bell, 2000), students can request activity hints, evidence hints, or claim hints from the pedagogical agent. When they do so, because there is no assessment of student work, all students who request an activity hint receive the same generic hint such as, "When you're reading the arguments, pay close attention to what the scientists are saying. How are they using the evidence to support their ideas?" (Bell, 2000). This happens regardless of whether or not that hint actually targets the help the student needs because there is no way of differentiating between different student work and needs due to the lack of assessment.

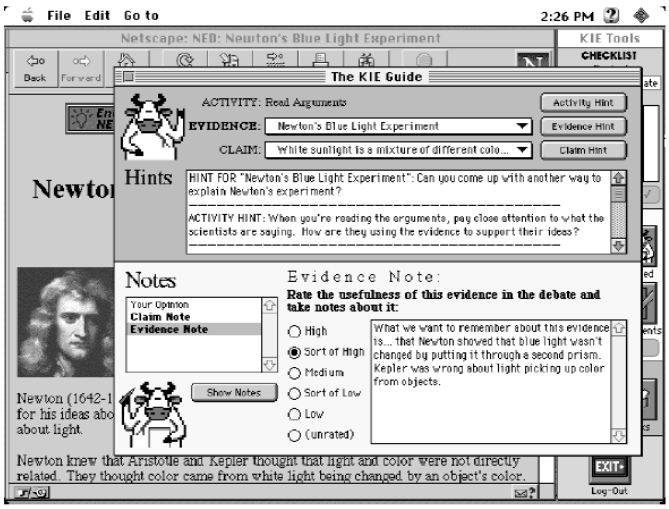


Figure 7: Example Hint in KIE

The table below illustrates the commonalities and differences between this set of learning environments and the scaffolding they contain, as described above.

Table 5: Learning Environments with Scaffolding

Environment	Addresses science inquiry	Addresses data interpretation inquiry skills/subskills	Assesses data interpretation inquiry skills in real-time	Scaffolding prompts given automatically in real-time	Different types of scaffolding prompts that adapt to students' difficulties
ConnectionLog (Belland, 2010)	✓	✓	✗	✗	✗
WISE - Chemical Reactions (Rafferty et al., 2013)	✗	✗	✗	✓	✓
WISE - Wolves (Kim & Hannfin, 2011)	✓	✓	✗	✗	✗
Galapagos Finches (Kyza, 2009)	✓	✓	✗	✗	✗
ScratchPad (Gijlers & de Jong, 2009)	✓	✗	✗	✗	✗
KIE & Sensemaker (Bell, 2000)	✓	✓	✗	✗	✓
Belvedere (Paolucci et al., 1996)	✓	✓	✗	✗	✓
Progress Portfolio (Land & Zembal-Saul, 2003)	✓	✓	✗	✗	✗
Seismic/Eruption (Hsu et al., 2014)	✓	✓	✗	✗	✓

5.4 Scaffolding Data Interpretation in Inq-ITS

One of the main drawbacks to the scaffolding presented by the learning environments discussed in the previous section is that many of them are provided in the form of text-based worksheets, or in some other similar form that is either not scalable or fine-grained. Additionally, these approaches typically require a student to know when they need help; however, students may not have the metacognitive skills needed to recognize this (Aleven & Koedinger, 2000; Aleven et al., 2004).

The work proposed here will address these drawbacks by using a computer-based scaffolding approach aimed at inquiry subskills. The scaffolding for this work is integrated into the Inq-ITS system in such a way that ties it to the automatic and ongoing evaluations of the inquiry subskills. Thus, these scaffolds are triggered not by a fixed schedule decided by a designer or by students when they believe they need help, but by the actual inquiry products the students produce such that students must successfully demonstrate that they no longer need the scaffolds and can perform the inquiry practice/skill correctly before they stop receiving the scaffolds.

Specifically, the scaffolds for this work have the following components:

1. Address data interpretation and warranting inquiry skills and subskills
2. Address data interpretation and warranting skills in real-time
3. Scaffolding through the use of prompts
4. Scaffolding prompts given automatically in real-time
5. Different types/levels of scaffolding prompts that adapt to students' difficulties

By incorporating all six of these elements into the data interpretation scaffolds in Inq-ITS, this work addresses a gap in previous work in scaffolding explanation/argumentation. Therefore, the end product of a set of automatic, real-time scaffolds that address data interpretation and warranting as a science inquiry practice/skill with different levels that adapt to students' difficulties and are tied to their assessment is a major contribution of this work.

6 - Study 1

The goals for study 1 were to: (1) design data interpretation scaffolds that adapt to specific difficulties and trigger in real-time and (2) develop the wording of the scaffolds so that they are understandable to the students using them. There were two research questions for this study:

1. What are the subskills in data interpretation and warranting that need to be assessed and scaffolded?
2. What are common difficulties students have when interpreting data and warranting claims?

In order to address these research questions and determine what difficulties students have when interpreting data and warranting claims and what subskills underlie these difficulties, it was necessary to conduct a literature review. This review led to the categorization of four types of procedural "buggy" data interpretation and warranting behaviors that students can exhibit:

1. Claim IV/DV does not match Hypothesis IV/DV
2. The claim does not reflect the data selected
3. The data selected are not properly controlled
4. Claim is incorrectly said to support/not support the hypothesis

These categories of difficulty, which will be further described below, led to the initial data interpretation scaffolds used for study 1.

6.1 Difficulty 1 – Claim IV/DV does not match Hypothesis IV/DV

Prior research has shown that students may create a claim that does not answer the question posed (McNeill & Krajcik, 2011) or does not relate the outcomes of the experiments to the theories being tested in the hypothesis (Schunn & Anderson, 1999; Klahr & Dunbar, 1988; Chinn & Brewer, 1993).

This difficulty is important to address because a student's expertise in inquiry practices can be seen by how well the student coordinates their intents by creating a claim based on the variables they intended to investigate (Kuhn, Schauble, & Garcia-Mila, 1992). In order to stay focused on the overall experimentation goal (and the variables addressed therein) (de Jong, 2006), it is essential that students build a claim around the same independent variable and dependent variable pair. Causal arguments should not only attempt to explain the data, but, ultimately, should serve to support or refute prior claims or hypotheses (Windschitl, Thompson, & Braaten, 2008), which can only be done when the independent and dependent variables in the claim match those in the hypothesis.

Within the Inq-ITS environment, this difficulty can present itself when students have a claim where the IV or the DV are not appropriately chosen or do not match those specified in the hypothesis. For example, this difficulty can be seen if a student were to create a hypothesis that states, "If I increase the amount of heat, then the boiling point of water increases" and the create a claim that states, "When I increased the amount of ice, the melting point of ice stayed the same." In this example, while the student has created a scientifically accurate claim, their claim

describes a relationship between the amount of ice and the melting point of ice, which is not the relationship they showed intention of exploring in their hypothesis. By creating this claim, this student cannot support or refute their hypothesis, which is an important piece of causal arguments (Windschitl, Thompson, & Braaten, 2008).

6.2 Difficulty 2 – The claim does not reflect the data selected

Prior research has shown that students lack understanding to provide evidence-based arguments (Fischer et al., 2014; Cerbin, 1998). They often state conclusions that are inconsistent with their data (Kanari & Millar, 2004); impose patterns on their data based on their prior knowledge or expectations (Kanari & Millar, 2004); and/or have difficulty in making a valid inference and reconciling previous conceptions with their collected data, falling back on prior knowledge (Schauble, 1990).

This difficulty is important to address because when scientists are answering questions, one of the integral processes they undergo is making sense of the data. Formulating explanations and claims from evidence as well as communicating those explanations/claims are two of the five essential features of classroom inquiry in national benchmarks (National Science Education Standards, 2000). Furthermore, analyzing and interpreting data, constructing explanations, and engaging in argument from evidence are three of the eight science and engineering practices emphasized in the Next Generation Science Standards (NGSS Lead States, 2013). Similarly, Toulmin's model of argumentation, which has been used by many science educators and researchers (McNeill & Krajcik, 2011; Berland & Reiser, 2009), prominently features the proper use of a claim.

Within the Inq-ITS environment, this difficulty can be seen when a student's claim does not match the relationship exhibited between the IV and DV in the data selected for warranting, i.e., the student constructs a claim that is not considered scientifically accurate in describing the phenomenon. For example, a student is exhibiting this difficulty if after collecting data that shows that the amount of ice does not change the melting point of ice, the student creates a claim that states, "When I increased the amount of ice, the melting point of ice increased" and uses their scientifically correct data that shows this to be untrue to warrant their claim.

6.3 Difficulty 3 –The data selected for warranting are not controlled.

An important part of science is its use of data as evidence to understand natural phenomena (McNeill & Krajcik, 2011). As such, warranting, or justifying, an explanation is one of the five essential features of classroom inquiry per NRC's National Science Education Standards (2011).

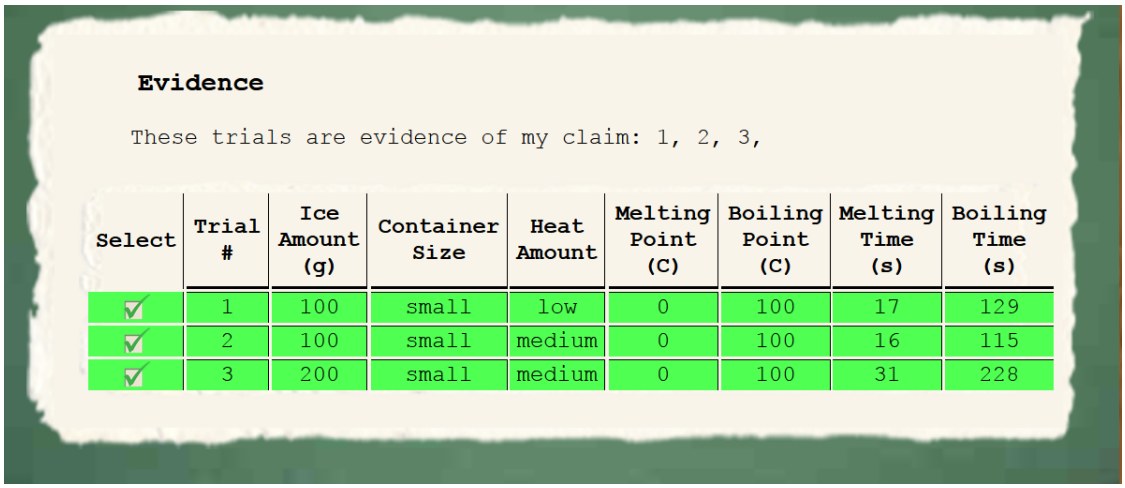
When constructing an explanation, students must engage in warranting in order to justify their explanation with sufficient evidence (Kang, Thompson, & Windschitl, 2014; Cho & Jonassen, 2002). Accordingly, an essential part of scientific explanation is providing a justification that both accounts for the patterns in the data as a way to explain complex phenomena, as well as links the claim to the evidence collected (Cho & Jonassen, 2002; Kang, Thompson, & Windschitl, 2014; McNeill & Krajcik, 2011; Sandoval & Reiser, 2004). As such, the use of evidence is often utilized to examine the quality of explanations (Kang, Thompson, & Windschitl, 2014).

One of the characteristics used to rate the quality of evidence is appropriateness, i.e., whether or not the evidence is scientifically relevant, and one way this can be assessed is by checking if the control for variables strategies was used. Determining if there are appropriate

data that can be used to support a claim is a critical aspect of creating explanations and can help build students' scientific literacy (McNeill & Krajcik, 2011).

While using evidence is important for convincingly communicating results, in many science classrooms students are not tasked to make use of the evidence they collect (McNeill & Krajcik, 2011). As such, students often provide little to no justification for their claims and when they do provide evidence for their claims, they often use inappropriate data (by drawing on data that do not support their claim) (Cerbin, 1998; McNeill & Krajcik, 2011). Furthermore, students may rely on their own opinions or experiences (McNeill & Krajcik, 2011), make no mention of specific evidence (Chinn et al., 2008), or generally state that an entire data table is evidence (McNeill & Krajcik, 2011; Chinn et al., 2008).

Within the Inq-ITS environment, this category of difficulty can manifest itself when a student demonstrates difficulty with selecting data that are properly controlled when warranting their claim. For example, given a claim that states, "When I increased the amount of ice, the melting point of ice stayed the same", a student may choose trials in which more than one variable changes to warrant this claim, as shown in the evidence table below where both the variables of amount of ice and amount of heat are changing.



Evidence

These trials are evidence of my claim: 1, 2, 3,

Select	Trial #	Ice Amount (g)	Container Size	Heat Amount	Melting Point (C)	Boiling Point (C)	Melting Time (s)	Boiling Time (s)
☑	1	100	small	low	0	100	17	129
☑	2	100	small	medium	0	100	16	115
☑	3	200	small	medium	0	100	31	228

Figure 8: Selected Evidence for Warranting

6.4 Difficulty 4 – Claim is incorrectly said to support/not support the hypothesis

Prior research has shown that students may not relate the outcomes of their experiment back to the theories being tested (from their hypothesis) (Schunn & Anderson, 1999) or may be hesitant to state that their claim does not support their hypothesis because they do not want to admit that their hypothesis was "wrong". Much of prior research surrounding this difficulty focuses on the claim itself (and the reasoning and evidence for it) and not on the link between the claim and the hypothesis. While students may be directed to compare the data they have collected to their original hypothesis in order to revise their hypothesis and guide further experimentation, they do not need to definitively state a claim and relate it back to their hypothesis by saying whether or not it supports or refutes it (Klahr & Dunbar, 1988).

Within the Inq-ITS environment, this difficulty can be seen when a student cannot correctly state whether or not their claim supports their hypothesis. For example, a student may hypothesize that "If I increase the amount of ice, the melting point of ice increases" and then correctly find in their claim that, "When I increased the amount of ice, the melting point of ice stayed the same," but state that their claim supports their hypothesis when in fact it does not.

6.5 Difficulty due to confirmation bias

Confirmation bias occurs when students make a scientifically incorrect (but testable) hypothesis, collect controlled data that refutes their hypothesis, but revert to pre-existing, scientifically incorrect beliefs while interpreting the data (Nickerson, 1998). Prior research has shown that confirmation bias affects the whole inquiry process (Klahr & Dunbar, 1988), but that specifically when students are interpreting their data, confirmation bias may cause students to revert to their pre-existing beliefs and draw a conclusion in line with their prediction despite the presence of

new, conflicting data and ideas, which they tend to reject or distort (Chinn & Brewer, 1993; Millar & Lubben, 1996). As such, they do not abandon the belief that a causal effect exists (Kuhn, Schauble, & Garcia-Mila, 1992), which can be an impediment from making important conceptual changes and learning content-related concepts (Chinn & Malhotra, 2002).

Addressing this difficulty is important because in order for students to create a scientifically accurate claim, they must be able to successfully coordinate theory and evidence and therefore must be able to reflect on their own theory and realize that it may be wrong as well as that another theory may disconfirm their theory (Kuhn, Schauble, & Garcia-Mila, 1992).

6.6 Summarizing categories of difficulty

As this literature on student difficulties with data interpretation and warranting demonstrates, while there are many benefits to engaging with data interpretation, these tasks can be very complex for students to accomplish. Furthermore, the variety and prevalence of these difficulties demonstrates the need for immediate, real-time assessment and scaffolding, otherwise learning may be impeded. This literature was therefore used to concretize four main categories of difficulty students have with data interpretation and warranting claims and it is upon these categories that our data interpretation and warranting evaluations and scaffolds are founded. Again, these four categories of difficulty are as follows:

1. Claim IV/DV does not match hypothesis IV/DV
2. Claim does not reflect the data selected
3. Data selected for warranting are not properly controlled
4. Claim is incorrectly said to support/not support the hypothesis

As previously mentioned, a key aspect of Inq-ITS is its capacity to provide an assessment of students' inquiry practices. During students' data interpretation and warranting, the subskills assessed in Inq-ITS are based on prior experience with students' inquiry as well as the four main categories of difficulty mentioned above. These subskills are evaluated using the work products students create through knowledge-engineered rules that specify if the subskill has been demonstrated. Each of the subskills, along with the criteria for their assessment, can be found in Table 2. All of these subskills are evaluated as either demonstrated or not (0 vs. 1 scoring) (Sao Pedro et al., 2013).

For example, for the subskill "Claim DV", the system evaluates whether or not the student has correctly chosen a scientifically accurate dependent variable in the appropriate widget for the claim. Within the context of the Density microworld (the one used in the present studies), the appropriate dependent variable is "density of the liquid". Therefore, if the student states "density of the liquid" as the dependent variable, they would be marked as correctly demonstrating the DV subskill. However, if the student chooses another variable, such as one of the independent variables like "type of liquid", as the dependent variable, then they would be scored as incorrectly demonstrating the DV sub-skill.

As such, the results from the analysis of student performance on the targeted subskills are provided as a fine-grained measure of correctness (i.e. the results are presented as being correct or incorrect rather than having levels of correctness), which allows us to tease apart separate sub-components under the broader practice/skill of analyzing data.

Table 2: Data Interpretation and Warranting Subskills and Criteria for Evaluation

Data Interpretation and Warranting Subskill	Criteria
Claim IV	Ensure the independent variable selected is a variable manipulable by the student (not measured). For example, choosing "amount of ice" as the independent variable would be scored as correct, whereas choosing "melting point" as the independent variable would be scored as incorrect.
Aligned IV	Ensure the independent variable in the claim is the same as the independent variable in the hypothesis, e.g., both independent variables are "amount of ice."
Claim DV	Ensure the dependent variable selected is a variable measured (not manipulated by the student). For example, choosing "melting point" as the independent variable would be scored as correct, whereas choosing "amount of ice" as the independent variable would be scored as incorrect.
Aligned DV	Ensure the dependent variable in the claim is the same as the dependent variable in the hypothesis, e.g., both dependent variables are "melting point."
Interpreting the IV/DV relationship	Ensure that the relationship between the independent and dependent variables stated in the student's claim is scientifically correct. For example, a correct claim would be "When I increased the amount of ice, the melting point of ice stayed the same", whereas an incorrect claim would be "When I increased the amount of ice, the melting point of ice increased."
Interpreting the hypothesis/claim relationship	Ensure that the claim correctly supports or refutes the hypothesis, as stated. For example, if the claim is the same as the hypothesis, it should be said to support the hypothesis. If the claim is different from the hypothesis, it should be said to refute the hypothesis.
Evidence	Ensure that more than one trial from the evidence table is chosen by the student to warrant their claim.
Controlled trials	Given the analysis, ensure that every trial selected by the student is controlled.

Target IV	Given the analysis, ensure that every trial selected by the student has the same independent variable as what is stated in the claim. For example, if the independent variable in the claim is "amount of ice," the selected trials used for warranting should all be about "amount of ice."
Warranting the IV/DV relationship	Ensure that the trials selected by the student reflect the relationship stated in the student's claim.
Warranting the hypothesis/claim relationship	Ensure that the trials selected by the student serve as evidence for determining if the claim is linked to the hypothesis.

The table below shows how each category of difficulty matches with the assessed data interpretation and warranting subskills discussed above:

Table 3: Data Interpretation and Warranting Subskills aligned to the Categories of Difficulty

Category of Difficulty	**Relevant Subskills**
Claim IV/DV does not match Hypothesis IV/DV	Aligned IV
	Aligned DV
Claim does not reflect the data selected	Interpreting the IV/DV relationship
	Warranting the IV/DV relationship
	Target IV
Data selected for warranting are not controlled	Controlled trials
Claim is incorrectly said to support/not support the Hypothesis	Interpreting the hypothesis/claim relationship
	Warranting the hypothesis/claim relationship

6.7 Scaffolding Frameworks/Guidelines

Seven main scaffolding guidelines for supporting science inquiry presented by Quintana et al. (2004) address the support of sense making and process management through several scaffolding mechanisms. These scaffolding mechanisms include (a) enlisting student interest, (b) controlling frustration, (c) providing feedback, (d) indicating important task/problem elements to consider, (e) modeling expert processes, and (f) questioning (Quintana, et al., 2004). The scaffolds developed through this work mainly address three of these mechanisms – indicating important elements to consider, modeling expert process, and questioning – through text-based prompts and hints. Along with the aforementioned scaffolding mechanisms, Quintana et al. (2004) present seven overall guidelines for creating scaffolds. Of the seven overall guidelines, five support the design of scaffolds in this work. These guidelines will be further discussed below.

One such guideline, scaffolding guideline 1, has to do with using representations and language that bridge learners' understanding (Quintana et al., 2004). Per this guideline, learning tools can serve as bridging scaffolds by supporting learners by using representations that they are familiar with and that also map onto expert practice (Quintana et al., 2004). This allows learners to make their own connections between problems and the concepts incorporated in them. Some strategies for implementing this guideline include providing visual conceptual organizers, using descriptions of complex concepts, and embedding expert guidance (Quintana et al., 2004). These tools can work to transform tasks in order to help learners with limited conceptual knowledge build more formal knowledge and successfully solve problems (Quintana et al., 2004). Within Inq-ITS, there are visual representations of the scientific experiments and concepts students are exploring as well as organizers, such as data tables that are automatically populated. These inquiry tools serve as implicit scaffolds within the learning environment. The explicit scaffolds

designed for this work draw from this guideline through embedded expert guidance and text-based prompts that explain complex concepts to students.

Another relevant guideline is scaffolding guideline 4: provide structure for complex tasks and functionality (Quintana et al., 2004). Some strategies for implementing this guideline include restricting a complex task by setting useful boundaries, describing a complex task, and constraining the space of activities by using functional nodes (Quintana et al., 2004).

Within Inq-ITS there are a number of inquiry tools that offer implicit scaffolding aligned with this guideline. For example, students' hypotheses and claims are constrained to a set list of relevant independent variables and dependent variables. This restricts the task somewhat so that students do not have to come up with their own variables to test; however, the pre-defined variables are chosen with specific student misconceptions and learning targets in mind, so they still allow students a complete opportunity to explore the concept at hand and conduct inquiry. Furthermore, the creation of a hypothesis and claim is restricted by the use of widgets that create a model for how the hypothesis and claim should look. Again, with these widgets, we are providing a structure for the tasks that allows all students, including ELL students who may struggle with the vocabulary, to easily create a correct hypothesis or claim. The explicit scaffolds designed for this work draw from this guideline by providing hints that describe complex tasks for students by decomposing them into smaller tasks. Essentially these scaffolds have different levels to address specific misconceptions or difficulties, giving students the exact, small piece of help they need to continue with their inquiry.

The main guideline that the explicit scaffolds designed through this work address is scaffolding guideline 5: embed expert guidance about scientific practices (Quintana et al., 2004). This guideline recommends providing access to expert knowledge about scientific practices such

as explaining, observing, and inferring (Quintana et al., 2004). This expert knowledge then helps learners understand both why they should embark on a particular task and how to steer their investigation. Strategies for implementing this guideline include embedding expert guidance that both clarifies the characteristics of scientific practices and indicates the rationales for them (Quintana et al., 2004). This is important because learners need to understand the reasons behind the different scientific practices they are performing (Collins, 1996; Quintana et al., 2004). The scaffolds designed for this work present prompts and hints to students from a pedagogical agent who is designed to be an expert in the learning environment. This agent provides different levels of help that give students the exact instruction they need to successfully continue in and complete their inquiry. Specifically, in the bottom-out hint, the hint students receive after having received all other hints for a difficulty, the agent presents information on not only what the scientific practice that they need to complete is, but also why they need to complete that practice. This taps into the idea of giving students help with the scientific practice as well as presenting the rationale behind it.

Two ways to create scaffolds that incorporate these guidelines and are effective for supporting ill-structured problem-solving processes is with prompts and through peer interaction (Xun & Land, 2004). The scaffolds in this work draw from the former and present support through prompts. Prompts can serve as effective scaffolds as they can facilitate the solution process by activating prior knowledge and mapping onto existing knowledge (Xun & Land, 2004). These prompts should be developed to elicit the desired cognitive and metacognitive activities and address student misconceptions and difficulties. Different types of prompts can be developed to focus on different problem-solving processes. These scaffolding prompts should also have levels that can be adjusted as students make progress or gain understanding. The

scaffolds designed in this work draw from these principles and utilize the use of prompts as well as levels within the prompts to create personalization that target students' specific difficulties. The scaffolding strategy used for this work is described in more detail in the next section.

6.8 Data Interpretation Scaffolding Strategy

The data interpretation scaffolds used here were designed to provide a framework for students' data interpretation and warranting, as well as address the four difficulties previously described and which are evaluated in Inq-ITS. These are:

1. Claim IV/DV does not match hypothesis IV/DV
2. The data selected for warranting are not properly controlled
3. The claim does not reflect the data selected
4. Claim is incorrectly said to support/not support the hypothesis

It is important to note that unlike on-demand help (Anderson et al., 1995) in which students ask for support, our pedagogical agent will provide scaffolding when our system detects, based on our assessment metrics, that the student needs help on a particular practice/skill. Although on-demand help can be effective, it requires students to have the metacognitive help-seeking skills to recognize when they actually need help (Aleven & Koedinger, 2000; Aleven et al., 2004); however, since students have difficulty engaging in inquiry without support (Schauble, 1990; de Jong, 2006) and in monitoring their progress (de Jong et al., 2005), we expect they will not necessarily be aware when they are haphazard in their inquiry (Buckley, Gobert, & Horwitz, 2006).

Since students may require scaffolding support for none, one, or many of these subskills underlying data interpretation and warranting, the scaffolds were designed to address these

difficulties in the order in which they were listed above, so that each step of data interpretation is completed before moving on to the next; in doing so, we have begun unpacking the necessary sub-components for a learning progression (McNeill & Krajcik, 2011) for data interpretation and warranting.

In order for scaffolds from a specific category to be triggered, one of the subskills associated with that category would have to be evaluated as incorrect. This allows for personalization to each student's learning, recognizing that different students may need different amounts of help to successfully hone different subskills.

For example, it is impossible for students to correctly select relevant trials for warranting if they have not specified an appropriate IV and DV in their claim. Therefore, difficulty with creating a claim with the correct IV and DV (i.e. category 1) is scaffolded first and no other difficulties are addressed until the subskill is demonstrated correctly. On the other hand, if a student only demonstrates difficulty with stating whether or not the claim supports the hypothesis, then the first three scaffolding categories are skipped and the student only receives the specific scaffolds that address category 4.

Our overall scaffolding strategy for the scaffolds was to provide students with more specifically targeted support (Corbett & Anderson, 1995; Koedinger & Corbett, 2006; Anderson et al., 1995) by first orienting students to the current task, then giving them necessary conceptual information, guiding them through the necessary procedural steps and, lastly, providing an instrumental "bottom-out" hint telling students the "procedure" to follow and why. An example sequence from orienting to instrumental scaffolding is highlighted below.

- Orienting scaffold: Orienting the student to his or her current task/step within the inquiry process is important as students have difficulty monitoring where they are in the inquiry process (de Jong, 2006; de Jong, et al., 2005). This scaffold refocuses the student to the task and/or to the information available. It allows students who have made a mistake, but do not need additional conceptual/procedural scaffolding because of their prior knowledge, the chance to reevaluate their work by themselves. In our scaffolds, orienting always comes first with the goal of reminding the student of their current step or task. For example:

  "You claim that as {IV} {increases/decreases}, the {DV} {increases/decreases/stays the same}. Check the data you selected again and see if this matches your claim."

- Conceptual scaffold: Providing the student with *conceptual* information needed for the inquiry task is important as they are learning in real time; for example, students may have difficulty translating and understanding how theoretical variables and manipulable variables relate to each other (van Joolingen & de Jong, 1997). All of these are examples of conceptual help that is needed for inquiry.

  "To back up your claim, you must have only one variable changing to see how it affects the experiment."

- Procedural scaffold: Providing students with information about how to conduct inquiry processes and procedures on the current step is critical as students are learning. Here, the pedagogical agent may focus the student to select only trials that relate to the hypothesis and using the control for variables strategy. Such procedural help is critical for

practices/skills in which students have difficulty like designing controlled experiments (Kuhn, 2005; Schunn & Anderson, 1998; Schunn & Anderson, 1999; Harrison & Schunn, 2004; McElhaney & Linn, 2008; McElhaney & Linn, 2010).

"Make sure all the trials you have selected have only the {IV} changing. Remember, this lets you tell for sure if the {IV} affects the {DV}.

- Instrumental scaffold: Telling the student exactly what to do on the current step, a "bottom-out hint", can be useful for students (Aleven & Koedinger, 2000). Providing help messages in these ways is based on principles from cognitive psychology (McKendree, 1990), and has been shown to benefit learning within learning environments for other domains (Beck, Chang, Mostow, & Corbett, 2008).

  "Your hypothesis is {supported/not supported} by your data. Change your claim so that it reflects this."

The following figure shows the different scaffolding branches for one of the categories of difficulty namely, "the data selected for warranting are not properly controlled". These figures were used during the pilot study as a script for the experimental researchers to deliver the scaffolds. Light blue boxes indicate actions to be taken by the pedagogical agent. Dark blue boxes indicate the responses that students can choose from after the pedagogical agent has given a scaffold (in the final implementation, these took the form of buttons on the computer screen that students could choose from and click, but in the pilot study, students were verbally prompted with these options by the researcher), or which subsequent scaffold gets fired if the student shows the same difficulty repeatedly.

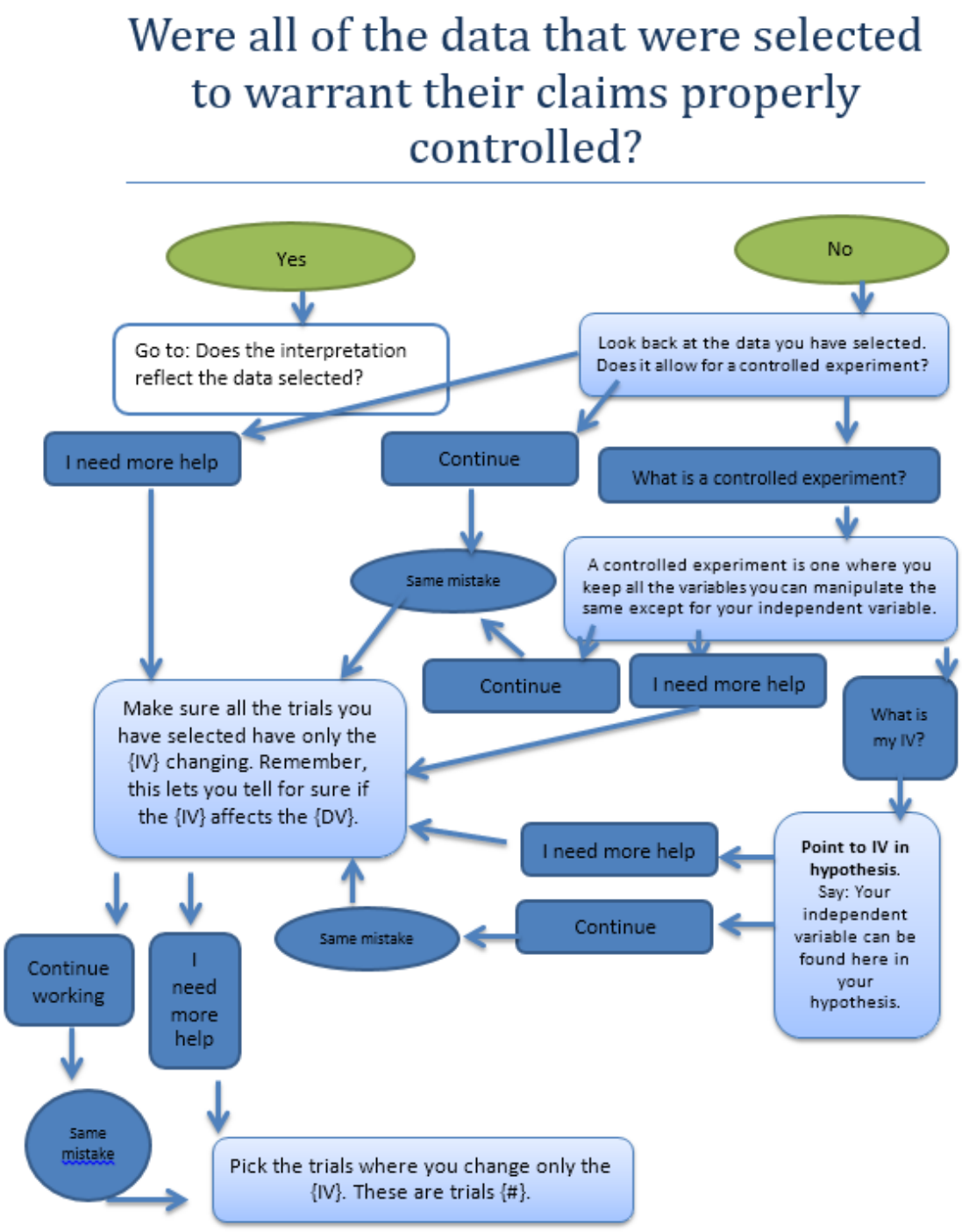


Figure 9: Scaffolding branches for difficulty with selecting controlled trials to warrant claims

6.9 Piloting the Data Interpretation Scaffolds

The designed scaffolds were piloted as a part of this study in February 2014 using think-aloud protocols with middle school students in order to refine the scaffolds.

Participants

The participants in this study were 10 middle school students (ages 12-15) of low SES from an urban, Central Massachusetts after-school program. The physical science labs in Inq-ITS are regularly used by 7th and 8th grade students, ranging in age from 12-14, so this group of students fits into the typical age of students who would use Inq-ITS. This group of students, however, had not previously used Inq-ITS.

Materials

In this study, two activities from the Phase Change microworld were used (Sao Pedro, Baker, & Gobert, 2013; Sao Pedro, Gobert, & Baker, 2014). These activities aim to foster understanding about the melting and boiling properties of ice. Each activity sets a different goal for the students. In the first activity, the goal was to determine if the amount of ice affected the boiling point; the second was to determine if the amount of ice affected the melting point.

When creating a hypothesis, students choose from a list of three IVs (amount of heat, amount of ice, and container size) and four DVs (melting point, boiling point, melting time, and boiling time). During their experimentation, students can manipulate any of the IVs to measure the effect on the DVs.

The two Phase Change activities students worked with had existing scaffolds (delivered automatically by the pedagogical agent) for the hypothesizing and experimentation stages (Sao

Pedro, Baker, & Gobert, 2013; Sao Pedro, Gobert, & Baker, 2014). This ensured that students had a testable hypothesis and at least one pairwise (but not necessarily sequentially collected) controlled trial that could be used for data interpretation and warranting, our practice/skill of interest.

Procedure

As students worked through the inquiry activities, they were encouraged to think-aloud throughout. Each student was paired with a researcher, who delivered the scaffolds during the data interpretation phase of inquiry, as needed by the student. All of the researchers were trained on how to deliver the scaffolds using paper copies of the scaffolds that included both the order in which the scaffolds should be presented and the text for each individual scaffold. If a student persisted with a difficulty past the bottom-out scaffold or if a difficulty was not addressed by a scaffold, the researcher interacted with the student to discern what the student was thinking and why they were continuing to exhibit a difficulty. This allowed for an iterative developmental approach for identifying refinements for the scaffolds (Larman & Basili, 2003).

Analysis

Eight students' think-aloud transcripts were analyzed by two researchers. The remaining two transcripts were not analyzed due to issues with recording clarity. These researchers examined how many students exhibited each type of "procedural" difficulty. An overall inter-rater reliability kappa of .915 was achieved, indicating strong agreement. The researchers also recorded how students responded to the scaffolds and what additional information students

required. As students worked across both activities, the researchers also recorded which scaffolds students were required by the students and which scaffolds were not required.

Analyzing Student Difficulties

All of the students exhibited at least one of the inquiry difficulties we had developed scaffolds for. These results are summarized in Table 4 and each difficulty is more fully analyzed in the following sections. Overall, this result supports prior research about the presence of these difficulties (Chinn & Brewer, 1993; Schunn & Anderson, 1999; Klahr & Dunbar, 1988; Quinn & Alessi, 1994) and indicates the importance of such scaffolds.

Table 6: Results from Piloting the Data Interpretation Scaffolds

ERROR ADDRESSED IN SCAFFOLDS	COUNT OF STUDENTS WITH ERROR ON FIRST ACTIVITY (OUT OF 8)	COUNT OF STUDENTS WITH ERROR ON SECOND ACTIVITY (OUT OF 5)	PERCENTAGE OF SCAFFOLDING NODES COVERED
IV/DV errors	**7**	**1**	**100%**
Errors warranting with controlled experiments	**6**	**0**	**100%**
Errors with claim	**6**	**4**	**60%**
Confirmation Bias	**2**	**0**	**100%**
Supports errors	**6**	**2**	**50%**

Difficulties with IV/DV (Claim IV/DV does not match Hypothesis IV/DV) (7 students). A total of seven students exhibited this type of difficulty. Of these seven students, five were able to correct themselves after scaffolding and fixed their claim so that the variables matched what was in the hypothesis. Two students did not know what independent and dependent variables were and needed their partner researchers to define them.

An example from the think-alouds of a student exhibiting this difficulty is shown below:

> ***Student**: "Oh. So, when I put less ice, the boiling time decreased."*

This example demonstrates that the student is able to identify and form a claim around a proper independent variable (amount of ice) and dependent variable (boiling time). However, in this instance, the student had hypothesized about the relationship between amount of ice and time to melt. Their exact hypothesis was, "If I change the amount of ice so that it decreases, the time the ice takes to melt decreases," so while they maintained the same IV in their claim, they changed their DV to "boiling time." This demonstrates that the student is not attending to their hypothesis (and the variables addressed there) when creating their claim, which is a known difficulty students have when building causal arguments (de Jong, 2006; Kang, Thompson, & Windschitl, 2014).

Difficulties with warranting claims with controlled trials (Data selected for warranting are not properly controlled) (6 students). A total of six students exhibited this difficulty. Of the six students exhibiting this type of difficulty, three students needed the researchers to provide more discussion regarding the definition of a controlled experiment. An example from the think-alouds of a student exhibiting this difficulty is shown below:

> ***Researcher**: Okay, can you compare, can you use this trial to compare to this trial? And again, was it a controlled experiment?*
> ***Student**: Because, yeah I can, because the independent variable is the amount of heat and I changed it.*
> ***Researcher**: Right, but did you keep everything else the same?*
> ***Student**: No. So I can use it?*

In this example, the student shows confusion in understanding what a controlled experiment is and the researcher paired with the student attempts to walk the student through each trial they have collected as a way of discovering which trials can be compared and thereby,

constitute a controlled pair. The student states that the trials are comparable because the target IV has changed, but does not pay attention to whether or not the other variables have changed and shows confusion in whether or not the changing of multiple variables means they can use the trials to warrant their claim.

Of the six students exhibiting this type of difficulty, three students needed additional explanation not provided in the scaffolds regarding how many trials were needed to warrant the claim.

The example below demonstrates the second difficulty that students exhibited, namely around how many trials they should select in order to warrant their claim. In this example, the student has just received scaffolding (from the researcher they are paired with) about choosing trials to warrant their claim. After receiving the scaffold, the student shows confusion about how many trials they need to choose and asks the researcher if they need to pick only one trial.

> ***Researcher****: Okay, now we need to drag trials to use the amounts from here to the evidence table. So to support your claim, you need to prove, you need to show some evidence. So you need to direct some experiments and the results to this evidence part.*
> ***Student****: Just one?*

Difficulties with claim (claim does not reflect the data selected) (6 students). A total of six students exhibited this difficulty. Of these six students, the existing scaffolds were effective for five students. One student required directed conversation during which each trial was individually interpreted and linked to the claim.

The following interaction between a researcher and a student demonstrates an example of this difficulty:

> ***Researcher****: "So, you claim that as the amount of ice increases, the melting point of ice increased. So check the data you selected and see if this matches your claim."*
> ***Student****: "I think I need the melting point of ice decreased."*

In this example, the researcher reads an orienting scaffold related to the claim difficulty to the student because they have just made an incorrect claim ("As the amount of ice increases, the melting point of ice increased"). While the student realizes that they need to change something in their claim, they demonstrate difficulty in creating a scientifically valid claim by changing their incorrect claim to another incorrect claim ("As the amount of ice increases, the melting point of ice decreased"). In this case, the student was able to fix their error by receiving another scaffold directing them to look at their data, focusing on what happens to the DV (melting point).

Difficulties with Supports (claim is incorrectly said to support/refute the hypothesis) (6 students). A total of six students exhibited this difficulty. The existing scaffolds were effective for all the students. Four students were able to remediate this after receiving the orienting scaffold and the other two were able to remediate this after receiving the bottom-out scaffold. The example below shows that a scaffold addressing the support difficulty (delivered by the researcher) was able to help a student remediate this difficulty.

> ***Researcher**: "Good interpretation, but check your claim again and see whether it supports or does not support your hypothesis."*
> ***Student**: "When I change the amount of ice so that it increased, the melting point of the ice did not change. Oh no, it isn't. Because it says it decreases, so…"*

In this example, after receiving the orienting scaffold from the researcher, the student looks back at their hypothesis and realizes that in it they had hypothesized that the melting point would decrease, which is not supported by their current claim ("When I change the amount of ice so that it increased, the melting point of the ice did not change").

Analyzing Student Actions across the Two Activities

We also examined if receiving scaffolding in the first activity would positively affect student performance in the second activity. If students exhibited fewer difficulties on the second activity and/or did not exhibit any of the same difficulties, this would suggest that the scaffolding helped foster skill acquisition for that skill/practice. We analyzed protocols from five students who had completed both activities. Of these students, two students repeated difficulties in the second activity. The repeated difficulties were ones in which the claim did not reflect the data selected and in which the claim was incorrectly said to support/not support the hypothesis. This may potentially indicate a need for further scaffolding and/or experience with these subskills.

Analyzing Coverage of Scaffolds

We also examined how many of the nodes (each individual scaffold in the decision tree, see Figure 9) were used at least once during the two activities to see if all the scaffolds had been tested. All of the nodes in the scaffolding trees for IV/DV difficulties and difficulties with warranting were seen by at least one student (100% coverage). However, the scaffolds for claim difficulties and supports difficulties had a lower coverage (60% and 50% coverage, respectively) (Table 4).

6.10 Revising the Data Interpretation Scaffolds

Based on the results from the pilot study, a number of changes were made to the scaffolds prior to automating them into the Inq-ITS system. The majority of the changes consisted of clarifying words as well as adding definitions of terms that students could request. The results from the

pilot study also led to the creation of a new category of inquiry difficulty that needed to be scaffolded. These changes are listed below:

1. Providing an explanation of an independent variable in the scaffolds addressing the variables in the claim
2. Providing an explanation of a dependent variable in the scaffolds addressing the variables in the claim
3. Providing an explanation of a controlled trial in the scaffolds addressing the use of controlled trials to warrant the claim
4. Adding a new set of scaffolds that address how many trials need to be selected to adequately warrant the claim

To address the fourth change, another scaffolding category was created that ensured students select at least two controlled trials to warrant their claim. While the evaluation rules for selecting trials for warranting required that students select at least two trials, this minimum number of trials needed was not clear in the scaffolds, as they previously existed. Therefore, it was necessary to create a new scaffolding category that addressed this difficulty exhibited by students during this study and this new scaffolding category also addresses prior research that has shown that one of the characteristics used to rate the quality of evidence is sufficiency (i.e., whether or not students have gathered enough data to support the claim) (McNeill & Krajcik, 2011).

Determining if there are sufficient data that can be used to support a claim is a critical aspect of creating explanations and can help build students' scientific literacy (McNeill & Krajcik, 2011). Correspondingly, the use of evidence has been a focus of the explanation and

argumentation literature and a consistent finding in this literature is that students tend to make claims with little to no justification (McNeill & Krajcik, 2011; Schunn & Anderson, 1999; Chinn et al., 2008; Kanari & Millar, 2004; Sandoval, 2003; Sandoval & Reiser, 1997; Kuhn, Schauble, & Garcia-Mila, 1992).

With the addition of this scaffolding category, there were a total of 5 categories of scaffolds that were then integrated within the scaffolding engine in Inq-ITS for automatic delivery, with the second category being introduced after this pilot study. These categories are:

(1) Claim IV/DV does not match hypothesis IV/DV;

(2) The data selected for warranting are not sufficient;

(3) The data selected for warranting are not properly controlled;

(4) The claim does not reflect the data selected;

(4) Incorrect claim due to confirmation bias; and

(5) Claim is incorrectly said to support/not support the hypothesis

These categories align with the following subskills assessed in Inq-ITS:

Table 7: Data Interpretation and Warranting Subskills aligned to the Revised Categories of Difficulty

Category of Difficulty	**Relevant Subskills**
Claim IV/DV does not match Hypothesis IV/DV	Aligned IV
	Aligned DV
Claim does not reflect the data selected	Interpreting the IV/DV relationship
	Warranting the IV/DV relationship
	Target IV
Data selected for warranting are not sufficient	Evidence
Data selected for warranting are not controlled	Controlled trials
Claim is incorrectly said to support/not support the Hypothesis	Interpreting the hypothesis/claim relationship
	Warranting the hypothesis/claim relationship

7 - Study 2

For study 2, the main goal was to assess more broadly via a real classroom intervention the efficacy of the data interpretation scaffolds in supporting students to learn the data interpretation and warranting subskills of interest. The main research question for this study was: Can data interpretation scaffolding from Rex help students acquire the assessed data interpretation and warranting subskills across the activities within one physical science microworld as measured by students' performance on data interpretation and warranting subskills?

7.1 Method

Participants

Data was collected from 160 8th grade students from the same school in the Northeast of the United States using the Density activities in Inq-ITS. All the students had previously used Inq-ITS, but not with its scaffolding capacity.

Materials

In this study, three activities from the Density virtual lab were used. These activities aim to foster understanding about the density of different liquid substances (water, oil, and alcohol). All students received the same three density activities. In the first activity, the goal was to determine if the shape of the container affected the density of the liquid; the goal in the second activity was to determine if the amount of liquid affected the density; and the goal in the third activity was to determine if the type of liquid affected the density. When creating their hypothesis, students can choose from three IVs (shape of container, amount of liquid, and type of liquid) to test the effects

on one DV (density of liquid). During experimentation, students could manipulate the IVs to measure the effect on the DV.

Procedure

Students worked on the density virtual lab in a computer lab at their school for the length of one science class (approximately 50 minutes). Each student worked independently on a computer at their own pace, meaning that not all students completed the entire set of activities by the end of the class period. Once students clicked to start the virtual lab, they were automatically randomly placed by the Inq-ITS system into either the "Interpretation Scaffolding" (n=78) or "No Interpretation Scaffolding" (n=82) condition.

For the first activity, none of the students, regardless of condition, received scaffolding. This allowed us to collect a baseline for each student on the targeted data interpretation and warranting subcomponents. For the next two activities, the students in the "Interpretation Scaffolding" condition received scaffolding during hypothesizing, data collection, and data interpretation. The students in the "No Interpretation Scaffolding" condition only received scaffolding during hypothesizing and data collection; the scaffolding during hypothesizing and data collection phases of inquiry ensured that all students, regardless of scaffolding condition, had both a testable hypothesis and relevant, controlled data with which they could correctly engage in data interpretation/analysis.

Table 8: Procedure for Study 2

Condition	Activity 1	Activity 2	Activity 3
Interpretation Scaffolding (n = 78)	No scaffolding present	Scaffolding present in: • Hypothesizing • Data Collection • Data Analysis	Scaffolding present in: • Hypothesizing • Data Collection • Data Analysis
No Interpretation Scaffolding (n = 82)	No scaffolding present	Scaffolding present in: • Hypothesizing • Data Collection	Scaffolding present in: • Hypothesizing • Data Collection

Data Analysis

Again, the main question in this study was: Can data interpretation scaffolding from Rex help students acquire the assessed data interpretation and warranting subskills across the activities within one physical science microworld? To address this question, Bayesian Knowledge Tracing (BKT) was used as a model for estimating student knowledge and learning. BKT was used because of its ability to track individual skills, look at learning across time (by capturing performance and latent knowledge) to trace the likelihood that a student knows a particular skill, evaluate the efficacy of interventions, and model student knowledge in intelligent tutoring systems (Corbett & Anderson, 1995; Pardos & Heffernan, 2010).

Due to the complexities and subskills inherent in the inquiry practices of data interpretation and warranting claims, it was necessary to employ an advanced analytical method to assess student learning and the role our data interpretation scaffolding played. We used an extension of the Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1995) framework that considers the presence of scaffolding as a cognitive modeling approach for approximating mastery learning of the inquiry practices/skills of interest (Sao Pedro et al., 2013), which is better suited to address the effects of scaffolding on students' learning and transfer of inquiry subskills (Sao Pedro et al., 2013).

Prior to using BKT, we first analyzed the frequency with which scaffolds were used across the activities as well as the error rates for the subskills rolled up to the level of the four main difficulties introduced in Study 1 – IV/DV relationship, claim is incorrectly said to support/not support the hypothesis, the data selected for warranting are not sufficient, and the data selected for warranting are not properly controlled – as a way to look at students' progress with and without scaffolding. BKT was then used as a model for estimating student knowledge and learning for all of the data interpretation and warranting subskills.

Bayesian Knowledge Tracing

Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1995) estimates the likelihood that a student knows a particular skill and disentangles between "knowing" a skill and "demonstrating" a skill based on prior opportunities in which students attempt to demonstrate a particular skill. BKT assumes that knowledge of a skill is binary (either a student knows they skills or does not) and that skill demonstration is also binary (either a student demonstrates a skill or does not). We note that our assessment algorithms that score the subcomponents of each practice are viewed as binary and are purposefully operationalized at a fine enough grain-size such that performance on them can be attributed *solely to that subcomponent* and no other subcomponents, as is typically done in intelligent tutoring systems. For a fuller discussion on this, see Anderson et al. (1995).

BKT can be and has been used in a variety of computer learning environments – from narrative centered environments for science inquiry to computer-based reading tutors (Beck & Sison, 2006; Beck, Chang, Mostow, & Corbett, 2008). Not only can BKT be used as a useful analysis tool in its basic form, but by adapting certain parts of BKT – as was done here – it can also be used for other learning environments that operate under different assumptions (Beck et

al., 2008; Sao Pedro et al., 2013). Because BKT assesses students' knowledge across a series of practice opportunities (i.e. attempts or submissions), it incorporates the potential for students to learn the skill within the learning environment. As such, BKT is useful in tracking knowledge within a system that supports learning as does Inq-ITS.

Mathematically, four parameters are used to model this sequence of knowledge states (as seen in Figure 10; Sao Pedro et al, 2013). These four parameters are: L_0, *T*, *G*, and *S* (Corbett & Anderson, 1995). L_0 is the probability of initial knowledge (of a skill), which is the chance that the student is already in the "learned state" (i.e. knows the skill) before they start the first problem. *T* is the probability of learning, which is the chance that the student goes from the "unlearned state" (i.e. does not know the skill) to the "learned state" over the course of doing all of the problems in the sequence. *G* is the probability of guessing, which is the chance that a student in the "unlearned state" answers the problem correctly. *S* is the probability of slipping, which is the chance that a student in the "learned state" answers the problem incorrectly (Corbett & Anderson, 1995). The parameters of *G* and *S* mediate the difference between "knowing" a skill and "showing" a skill. A student who shows the skill may not actually know it, contributing to *G*. Conversely, a student who knows the skill may not always show it, contributing to *S*. Accordingly, *G* and *S* take into account one of the main assumptions of BKT – there is no forgetting (Corbett & Anderson, 1995). Once a student is in the "learned state", they cannot forget and go back to the "unlearned state". Instead, if a student in the "learned state" does not "show" a skill at a specific practice opportunity (i.e. one attempt or submission), they are considered by the model to have "slipped," i.e. they were not able to show the skill at that time despite knowing it. This then affects the *S* parameter, but does not change what state the student is considered to be in. Therefore, once a student learns a skill and is considered to be in the

"learned state", they will always know that skill and be in that state, regardless of whether or not they show the skill at any following practice opportunity.

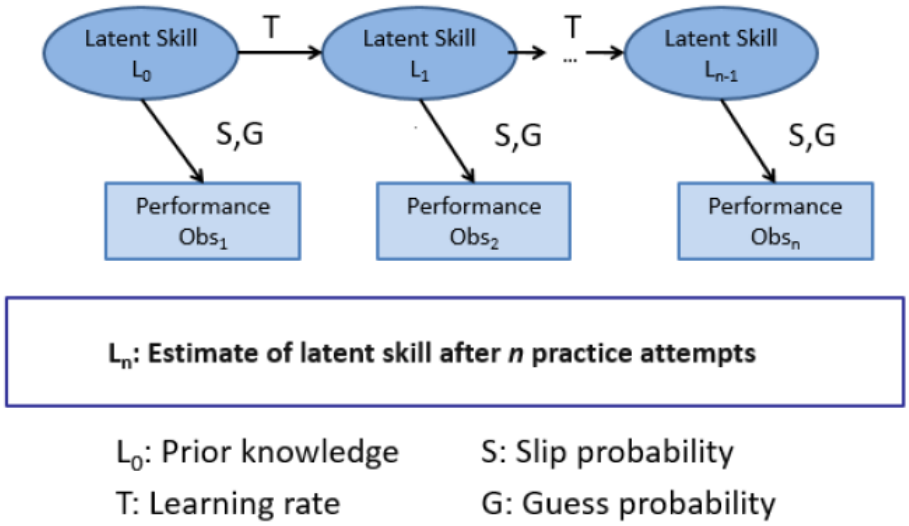


Figure 10: Bayesian Knowledge Tracing Model

Similar to the incorporation of tutor intervention in a BKT model done by Beck et al. (2008), Sao Pedro, Baker, and Gobert (2013) extended the traditional BKT model to account for the presence of a tutor intervention by introducing the dichotomous observable variable of Scaffolded = {True, False} and conditioning only the learning rate (T), which led to two distinct learning rate parameters – $T_{scaffolded}$ and $T_{unscaffolded}$. This resulted in the following equations for computing P(Ln) (Sao Pedro et al., 2013):

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

This work draws from the prior work done by Sao Pedro et al. (2013) and utilizes this scaffolding extension of the BKT framework to measure students' inquiry practice/skill development, specifically focusing on the practices/skills used for data interpretation and warranting and the scaffolds designed to address difficulties with these practices/skills. The goal of this work was to first fit the BKT model to a complex inquiry task (data interpretation) and then to apply and interpret the extended BKT model to assess whether or not the scaffolding approach was effective.

One of the main assumptions of BKT is that skills are considered to be independent. This means that each skill that we want to track has to be modeled separately. Because of this, there were certain design considerations that we had to make when fitting this data to the BKT model, specifically with regards to how the scaffolding condition was defined and how practice opportunities were defined. These considerations are discussed in the following section.

Fitting Data to the Bayesian Knowledge Tracing Model

The data logged here differs from typical data logs due to how the data interpretation scaffolds were integrated into the Inq-ITS system. In the system, all of the data interpretation and warranting subskills are designed to be evaluated at once. That is, in Inq-ITS, when students click to submit their data interpretation, the system records all of the actions as one practice opportunity and evaluates all of the subcomponents (Gobert et al., 2013). However, the data interpretation scaffolds are designed to only address one subcomponent at a time in order to give directed support, as described above. For example, if a student submits their analysis and is evaluated as both choosing an incorrect IV and an incorrect IV/DV relationship, even though they will have been evaluated on every data interpretation and warranting subcomponent, they

will only receive the scaffold for one of their errors, in this case the error of the incorrect IV. Once the student revises their analysis and submits again, they are once again evaluated on all of the data interpretation and warranting subcomponents, regardless of what specific aspects of their analysis they changed.

Considering this and since in the BKT analysis every subskill is considered separately and has its own model, it was important to consider how the BKT framework defined the scaffolding condition and practice opportunity to create an accurate model, as described in more detail below.

Determining Scaffolding Condition.

Not all of the 78 students in the Interpretation Scaffolding condition needed the data interpretation scaffolds and while some students only used one scaffold, others used multiple scaffolds targeting multiple subskills. Since BKT operates under the assumption of independence of skills, it would not be appropriate to label all of these students as having been scaffolded and, arguably, it is more important to model the scaffolds students received on a per subskill basis, rather than simply considering the students as scaffolded or not. The subskills considered here are the eight data interpretation and warranting claims subskills evaluated within Inq-ITS, as listed in the table below.

Table 7: Data Interpretation and Warranting Subskills with Criteria for Evaluation

Data Interpretation and Warranting Subskill	Criteria
Claim IV	Ensure the independent variable selected is a variable manipulable by the student (not measured). For example, choosing "amount of ice" as the independent variable would be scored as correct, whereas choosing "melting point" as the independent variable would be scored as incorrect.
Claim DV	Ensure the dependent variable selected is a variable measured (not manipulated by the student). For example, choosing "melting point" as the independent variable would be scored as correct, whereas choosing "amount of ice" as the independent variable would be scored as incorrect.
Interpreting the IV/DV relationship	Ensure that the relationship between the independent and dependent variables stated in the student's claim is scientifically correct. For example, a correct claim would be "When I increased the amount of ice, the melting point of ice stayed the same", whereas an incorrect claim would be "When I increased the amount of ice, the melting point of ice increased."
Interpreting the hypothesis/claim relationship	Ensure that the claim correctly supports or refutes the hypothesis, as stated. For example, if the claim is the same as the hypothesis, it should be said to support the hypothesis. If the claim is different from the hypothesis, it should be said to refute the hypothesis.
Evidence	Ensure that more than one trial from the evidence table is chosen by the student to warrant their claim.
Controlled trials	Given the analysis, ensure that every trial selected by the student is controlled.
Warranting the IV/DV relationship	Ensure that the trials selected by the student reflect the relationship stated in the student's claim.
Warranting the hypothesis/claim relationship	Ensure that the trials selected by the student serve as evidence for determining if the claim is linked to the hypothesis.

Because of this, scaffolding was considered at the subskill level so that any scaffolds a student received for one specific subskill had no bearing on the student's scaffolding classification for the other subskills. This means that in the BKT model for the Claim DV subskill, for example, a student will only be considered to have been in the scaffolding condition if they ever received the specific scaffold directly addressing the Claim DV subskill, regardless of any other scaffold they may or may not have received. In this way, a student who received scaffolds only for Claim DV and Claim IV would be in the scaffolding condition in the BKT models for those two subskills, but would be in the unscaffolded condition in the BKT models for the other six subskills. This allowed us to directly measure the effect of a scaffold on the subskill it directly addresses, which gives us a cleaner analysis of the efficacy of each scaffold.

Determining Number of Practice Opportunities.

As previously stated, in Inq-ITS students click to submit their data interpretation after which the system records all of the actions as one practice opportunity and evaluates all of the subskills at once (Gobert et al., 2013). Students who have been evaluated as incorrectly demonstrating any subskill receive scaffolding and are redirected to their data interpretation. Any subsequent actions students perform (up until submitting again) are considered part of a new practice opportunity for all the subskills regardless of which specific subskill(s) were worked on, which can make it seem as though students require more practice opportunities to master a subskill than they actually do. For example, as shown in Table 8, based on the evaluations, it looks like after three practice opportunities, the student is still incorrectly demonstrating the "claim" and "support" subskills. However, if we look at the student's actions, we can see that the student was only focused on correctly demonstrating the "DV" subskill (due to the scaffolding received) and was not actually working on the other two subskills. Therefore, it would not be accurate to say

that the student had three practice opportunities for the "claim" and "support" subskills. This, then, needed to be accounted for in the BKT models in order to more accurately assess students' probability of learning.

Table 8: Example of Practice Opportunity Succession Before and After Collapsing

Student Presses Submit			
Skills	**Evaluation**	**Practice Opportunity**	**Scaffolding Given**
IV	1	1	False
DV	0	1	False
Claim	0	1	False
Supports	0	1	False
Student receives scaffolding for DV, only changes DV (still incorrect), and submits			
Skills	**Evaluation**	**Practice Opportunity**	**Scaffolding Given**
IV	1	2	False
DV	0	2	True
Claim	0	2	False
Supports	0	2	False
Student received scaffolding for DV, only changes DV (correctly this time), and submits			
Skills	**Evaluation**	**Practice Opportunity**	**Scaffolding Given**
IV	1	3	False
DV	1	3	True
Claim	0	3	False
Supports	0	3	False

↓

Collapsed evaluation after student submits after three practice opportunities and after receiving scaffolding for DV		
Skills	**Evaluation**	**Scaffolding Given**
IV	1	False
DV	0	True
Claim	0	False
Supports	0	False

The option considered here was to collapse student evaluations for each subskill within each activity into one practice opportunity. This acts as a "pre-smoothing" of data and while it looks at the data in a slightly coarser way because of the rolling up of practice opportunities, it yields an easier-to-interpret model with fewer parameters. In collapsing students' evaluations, all

of the evaluations for one subskill within an activity were examined and a student would receive a correct evaluation for a particular subskill only if they always had correct evaluations for that subskill. Consequently, if a student gets a subskill wrong at ANY part of the scaffolding sequence then it is counted as not demonstrating in the collapsed version. This was done because if a student ever incorrectly demonstrated a subskill, it could be assumed that the student most likely did not know the subskill at the onset of inquiry. This resulted in the student's evaluations in the above figure to be collapsed into one practice opportunity as shown in Table 8.

Before performing the BKT analysis, the data set was halved – with half of the data being used to build the BKT models and the other half used to validate them. Then, in order to determine the parameters for each of the BKT models (one model per targeted data interpretation and warranting subskill), we used a brute force grid search approach (Baker et al., 2010). This method does two passes through the data in order to find a finer estimation (at a 0.01 grain-size) of the parameter values based on the lowest sum of squared residuals (SSR) between the probability of demonstrating a skill and the student's actual data.

Once the BKT parameters were determined, they were applied to the model and then its predictive performance was tested against the held out set of data that was not used to construct the model. Here, performance was measured using A' (Hanley & McNeil, 1982), which is the probability that the detector will able to correctly label two examples of students' skill evaluation when in one the student is correctly demonstrating the skill and in the other the student is not. An A' of 0.5 is indicative of chance performance and an A' of 1.0 is indicative of perfect performance.

7.2 Results

Table 9 shows the number of students who received any data interpretation scaffold in an activity and the total number of scaffolds triggered in an activity. Not all the students were able to finish the third activity within the time frame of their science class, contributing to the lower number of students in Activity 3. Looking at these numbers, we can see that by the third activity, a fewer number of students received scaffolds and that these students, overall, required less scaffolding support to successfully demonstrate the data interpretation and warranting subskills that we evaluate. This gives an initial indication that the scaffolding support, in its entirety, is helping students successfully interpret the data they collected and warrant their claims with data.

Table 9: Students using any data interpretation scaffold

	Activity 2	*Activity 3*
# of students in Interpretation Scaffolding condition who completed activity	76	64
# of students who used scaffolds	25	12
Total # of scaffolds triggered	207	32

We next looked at the error rates for the data interpretation and warranting claims subskills rolled up (i.e., aggregated) to the level of the four main difficulties introduced in Study 1 – IV/DV relationship, claim is incorrectly said to support/not support the hypothesis, the data selected for warranting are not sufficient, and the data selected for warranting are not properly controlled. The graphs below (Figure 11) show the error rate of students in each of the two conditions (Interpretation Scaffolding condition and No Interpretation Scaffolding condition) as

they worked through the three activities. As shown in these graphs (Figure 11), student difficulty/error was present in each of these difficulties, with the difficulty "Interpreting correct IV/DV relationship" and "Interpreting hypothesis/claim relationship" having the highest initial error rates, regardless of condition. Furthermore, this analysis revealed that students in the "Interpretation scaffolding" condition started with a higher error rate, but ended with a lower error rate. For example, for the difficulty "Warranting with controlled trials", on their first opportunity, students in the Interpretation Scaffolding condition had an error rate of 0.33 compared to an error rate of 0.26 exhibited by the students in the No Interpretation Scaffolding condition. However, by their third opportunity, students in the Interpretation Scaffolding condition had a much lower error rate of 0.05, which was less than the error rate of 0.16 exhibited by the students in the No Interpretation Scaffolding condition. This indicates that students in the Interpretation Scaffolding condition are improving faster than the students in the No Interpretation Scaffolding condition.

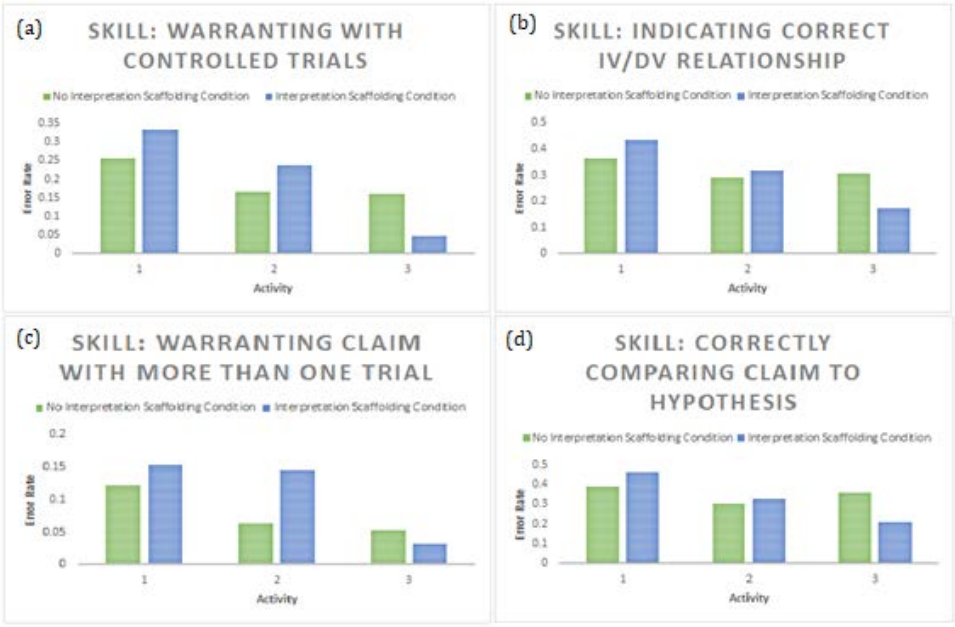


Figure 11: Error rate analysis for rolled up subskills: (a) warranting with controlled skills, (b) indicating correct IV/DV relationship, (c) warranting claim with more than one trial, (d) correctly comparing claim to hypothesis

Finally, the BKT analysis was performed for each of the assessed data interpretation subskills, using the scaffolding extension of the BKT framework (Sao Pedro et al., 2013) and a brute force grid search approach (Baker et al., 2010), as previously described. The A' values for this analysis (as seen in Table 10) were between 0.68 and 0.79, which allows us to interpret the BKT model parameters as they indicate that the model is a good predictor of student performance.

Table 10: A' values of Data Interpretation and Warranting Subskills

	Scaffolded BKT Model	Classic BKT Model
Skill	**A'**	**A'**
Claim IV	0.68	0.66
Claim DV	0.7	0.7
Interpreting the IV/DV Relationship	0.74	0.73
Interpreting the Hypothesis/Claim Relationship	0.72	0.72
Controlled Trials	0.79	0.78
Evidence	0.79	0.78
Warranting the IV/DV Relationship	0.76	0.75
Warranting the Claim/Hypothesis Relationship	0.73	0.73

The results from the BKT analysis indicate that the data interpretation scaffolds were effective in supporting the acquisition of the data interpretation and warranting subskills. This can be seen through the values of the probability of learning for each of the data interpretation and warranting claims subskills. The probability of learning represents the chance that the student goes from the unlearned state to the learned state over the course of activities. As can be seen in the data table below (Table 11), the probability of learning for students receiving data interpretation scaffolding is higher for five of the evaluated subskills – Claim DV, Controlled

Trials, Evidence, Warranting the IV/DV Relationship, and Warranting the Hypothesis/Claim Relationship. For two other subskills – Interpreting the IV/DV Relationship and Interpreting the Hypothesis/Claim Relationship – the probability of learning for students receiving data interpretation scaffolding is lower than for students not receiving data interpretation scaffolding, but the probabilities are very close (within .04). This would suggest that, for these subskills, the data interpretation scaffolding was as effective as the system by itself is at helping students acquire these subskills. The remaining subskill, selecting an IV for the claim, has a very low probability of learning for students who received data interpretation scaffolding. However, this subskill also has a high probability of initial knowledge. Since students only receive the appropriate scaffolding when they exhibit difficulty with a subskill, this high level of initial knowledge suggests that most students already knew this subskill and therefore, would not have received any scaffolding for it, meaning that they would not be in the data interpretation scaffolding condition for this subskill. Also, compared to another subskill with a relatively high probability of initial knowledge such as the Evidence subskill, the Claim IV subskill can be a noisier variable to assess, as it is very related to the content presented in each activity because knowing that you need more than one trial to warrant your claim (i.e. the Evidence subskill) is necessary for any claim regardless of domain, whereas recognizing what an independent variable is (i.e., the Claim IV subskill) is more domain-specific, i.e., a relevant independent variable in density is different from a relevant independent variable in free fall.

Table 11: BKT parameters for each subskill

Skill	Probability of Initial Knowledge	Probability of Guessing	Probability of Slipping	Probability of Learning	
				No Interpretation Scaffolding Condition	Interpretation Scaffolding Condition
Claim IV	0.934	0.249	0.014	0.99	0.001
Claim DV	0.75	0.3	0.051	0.7	0.99
Interpreting the IV/DV Relationship	0.657	0.004	0.1	0.4	0.39
Interpreting the Hypothesis/Claim Relationship	0.626	0.029	0.1	0.35	0.31
Controlled Trials	0.73	0.16	0.042	0.23	0.54
Evidence	0.862	0.174	0.013	0.02	0.71
Warranting the IV/DV Relationship	0.674	0.041	0.1	0.28	0.51
Warranting the Hypothesis/Claim Relationship	0.631	0.047	0.1	0.3	0.31

7.3 Discussion

The goal of this work was to test the efficacy of our data interpretation scaffolding on the sub-skills underlying the skills/practices of data interpretation and warranting claims within one domain/topic. This was tested by using an innovative extension to Bayesian Knowledge Tracing (BKT) that considers the presence of scaffolding for approximating mastery learning for each of the subskills of interest (Sao Pedro et al., 2013b).

This study contributes to the development of a fine-grained method for unpacking the effect of scaffolding via logged, process data. In this study, an extension to BKT was used as a modeling paradigm to track students' data interpretation and warranting subskill development with and without scaffolding. The modifications that were developed to this framework allowed it to be applied when condition and practice opportunity can be defined on different levels. This is important because this study was done within a complex domain of science inquiry in which the student data, number of practice opportunities, and evaluated skills were not as clearly delineated as in previous studies in which BKT was used (Beck & Sison, 2006; Beck et al., 2008). While other systems force one skill to be applied at one time in order to meet BKT's

requirement of skill independence, this makes for inauthentic inquiry because authentic inquiry requires the simultaneous coordination of a number of subskills. Therefore, that approach could not be used here because in Inq-ITS, all the skills are applied at the same time, which is a more authentic in terms of inquiry. For this reason, each subskill is modeled in this study with a separate BKT model, which allows for the mastery of each subskill to unfold at different rates, preserving their independence. Thus, using this modified BKT framework, a student could be shown to master one subskill more quickly than other subskills or master some subskills, but not others. In addition, defining what constitutes as a practice opportunity further helped to disentangle both the BKT model and the assessment of the many inquiry subskills. As such, analyzing these data in this kind of authentic inquiry environment – consisting of many subskills – is one of the main contributions of this work and this work provides a framework for how data in these complex environments can be treated before BKT can be used.

This study also builds on prior work (Kang, Thompson, & Windschitl, 2014; McNeill & Krajcik, 2011; Schauble, 1990) regarding the assessment and scaffolding of the explanation framework (i.e., claim, evidence, and reasoning) by conceptualizing and assessing the components of data interpretation and warranting as part of the explanation framework. Doing this in an automated and fine-grained way with component subskills can help us unpack and target known difficulties documented by previous research and more deeply elaborated on in Study 1. It also helps us support students in the components of data interpretation and warranting (Gotwals & Songer, 2009; McNeill & Krajcik, 2011; Schunn & Anderson, 1999). As such, this work provides a scalable solution towards the assessment and scaffolding of these practices and in doing so represents a scalable solution to supporting teachers and students in NGSS practices.

8 – Study 3

In this study, the data interpretation scaffolds tested in study 2 were assessed in terms of their efficacy in aiding in the transfer of the interpreting data and warranting claims practices/skills as students move within a microworld domain, i.e., from one driving question in a domain to another driving question within the same domain.

The specific research questions for this analysis were:

1. Can data interpretation scaffolding from Rex help students acquire the assessed interpreting data and warranting claims practices/skills and transfer them to a new activity within the same microworld (i.e., across driving questions)?
2. Are there any differences in students' transfer within different microworlds? (i.e., Is transfer across the activities within the Density microworld the same as it is within the Free Fall microworld?)

This study will look at four different physical science microworlds: (1) Collisions, (2) Density, (3) Free Fall, and (4) Phase Change. By examining transfer of practices/skills in different domains, we can see if practices/skills are more easily transferable in specific domains, providing some insight about the interaction between content knowledge and inquiry practices/skills (Greiff et al., 2014; McNeill & Krajcik, 2009; Schunn & Anderson, 1999).

8.1 Data Collection

The data for this study were collected during the 2015-2016 school year from students' interactions in physical science microworlds with and without Rex. Data were collected from two different schools across the country.

Data from one school consisted of 8th grade students from two teachers' classes. These students completed three physical science microworlds using a Rex/No Rex randomized control method. These three microworlds were: Density, Phase Change, and Free Fall: Speed.

The second school had two sets of data – one set collected from 8th grade students from two teachers' classes and another set of data collected from 7th grade students from two other teachers' classes. The 8th grade students completed three physical science microworlds using a Rex/No Rex randomized control method. These three microworlds were: Density, Free Fall: Speed, and Collisions.

The 7th grade students completed two physical science microworlds using a Rex/No Rex randomized controlled method. These two microworlds were: Phase Change and Density. These data allowed for the analysis of skill transfer within four microworlds. These analyses are described in the table below.

Table 12: Microworld Analysis

Microworld	School(s)	Grade(s)	N (Rex/No Rex)
Collisions	2	8th	16 / 28
Density	1	8th (Teacher A)	7 / 20
	1	8th (Teacher B)	18 / 24
	2	7th	21 / 20
	2	8th	14 / 29
Free Fall	1	8th	14 / 17
	2	8th	20 / 51
Phase Change	1	8th	8 / 34
	2	7th	27 / 36

8.2 Data Analysis

The goal of this analysis was to investigate whether students improved on their performance of interpreting data and warranting claims inquiry practices/skills with scaffolding help from Rex. Data from each microworld - collisions, density, free fall, and phase change - were analyzed through a series of Rex/No Rex repeated measures ANOVAs and examined for skill transfer in the practices/skills of interpreting data and warranting claims.

These analyses looked at student performance at the level of the aggregated practices/skills of interpreting data and warranting claims (meaning the sub-components for each respective practice were aggregated). This allowed us to disentangle any differences between the two practices/skills, as it is possible that one will be more domain-independent and easily

transferrable than the other. For example, it is possible that data interpretation will be more domain-specific than warranting claims as it is more dependent on the specific variables in a domain. In order to aggregate the analyzed practices/skills, the six data interpretation subcomponents were consolidated into the aggregated data interpretation practice/skill. These subcomponents are:

- Claim IV
- Claim DV
- Aligned IV/DV
- Interpreting the IV/DV relationship
- Warranting the IV/DV relationship
- Target IV

Since interpreting data is made up of six subcomponents (each dichotomously scored), the total possible score for interpreting data is six.

The four warranting claims subcomponents were consolidated into the aggregated warranting claims practice/skill. These subcomponents are:

- Evidence
- Controlled trials
- Interpreting the hypothesis/claim relationship
- Warranting the hypothesis/claim relationship

Since warranting claims is made up of four subcomponents (each dichotomously scored), the total possible score for warranting claims is four.

Table 13: Data Interpretation and Warranting Difficulties, Relevant Subcomponents, and Aggregated Practice

Category of Difficulty	**Relevant Subcomponent**	**Aggregated Practice/Skill**
Claim IV/DV does not match Hypothesis IV/DV	Claim IV Claim DV	Interpreting data Interpreting data
	Aligned IV/DV	Interpreting data
Claim does not reflect the data selected	Interpreting the IV/DV relationship	Interpreting data
	Warranting the IV/DV relationship	Warranting claims
	Target IV	Interpreting data
Data selected for warranting are not sufficient	Evidence	Warranting claims
Data selected for warranting are not controlled	Controlled trials	Warranting claims
Claim is incorrectly said to support/not support the Hypothesis	Interpreting the hypothesis/claim relationship	Interpreting data
	Warranting the hypothesis/claim relationship	Warranting claims

The variables that were considered in the repeated measures analysis were: (1) Rex condition and (2) student performance on the two aggregated practice/skills on their first opportunity in each of the activities in the microworld. For this analysis, students were defined as being in the scaffolding (Rex) condition if they were in the scaffolded condition for the given microworld (as randomly determined by the Inq-ITS system) and triggered scaffolding during any of the activities in the microworld.

In this study, student performance on only the first opportunity in each of the activities in the microworld was considered. The schematic below illustrates the student data that were collected and used for the analyses. In this example, the student is working in a microworld that has three activities/driving questions.

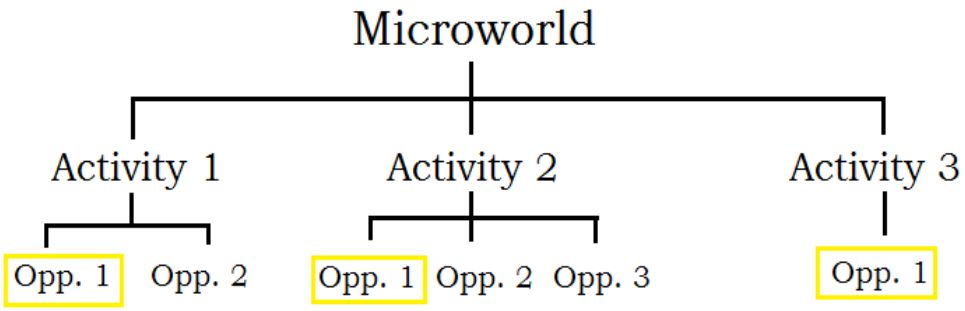


Figure 12: Data used for analysis

Only students' evaluations from their first opportunity were considered and included in data analysis because scaffolding support for each activity is initially triggered only after students' first opportunity. Therefore, by only considering performance on the first opportunity, we were able to measure students' performance prior to their receiving scaffolding support.

8.3 Results

The results from this analysis are presented in the following sections. The results for the interpreting data aggregated practice/skill will be discussed first, followed by the results for the warranting claims aggregated practice/skill.

Significant, positive effects for Rex on Interpreting Data

Students in the Rex condition from the datasets discussed in this section outperformed students in the No Rex condition on their performance of the interpreting data aggregated practice/skill. That is, they exhibited higher gains as they moved through the microworld from the first activity to the last, i.e., students in the Rex condition had a higher overall increase in the evaluation of their interpreting data score.

The first dataset is that from School 2 in Collisions. This dataset had a significant positive effect for Rex, p = .001, on the interpreting data practice/skill. Students in the Rex condition in this dataset increased their scores on interpreting data by 1.6 (compared to an overall increase of 0.2 by students in the No Rex condition).

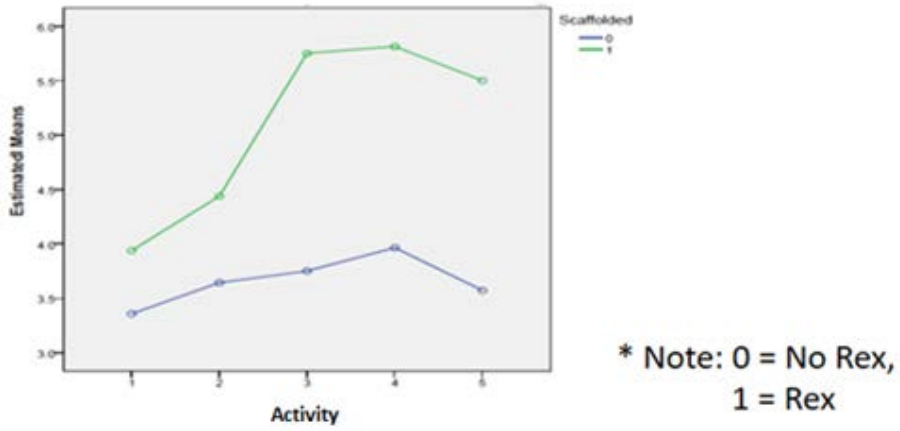


Figure 13: Student Performance in the Collisions Microworld on the Interpreting Data practice/skill (green = Rex condition, blue = No Rex condition)

Two datasets from the density microworld – school 1, teacher B and school 2, grade 8 – had significant positive effects for Rex on the interpreting data practice/skill (school 1, teacher B: p = .003; school 2, grade 8: p = .005). For the dataset from school 2, grade 8, students in the Rex condition increased their scores in interpreting data by 0.8 (compared to an increase of 0.7 by students in the No Rex condition).

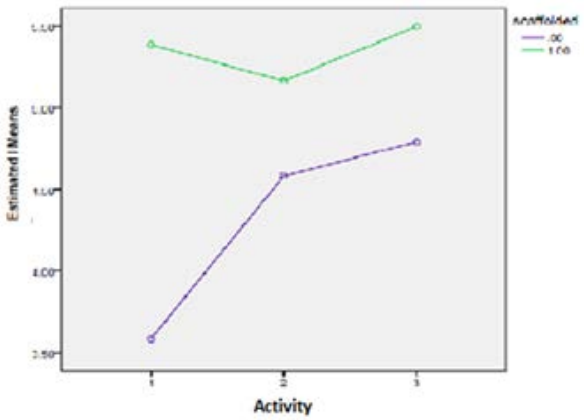


Figure 14: School 1, Teacher B - Student Performance in the Density Microworld on the Interpreting Data practice/skill (green = Rex condition, blue = No Rex condition)

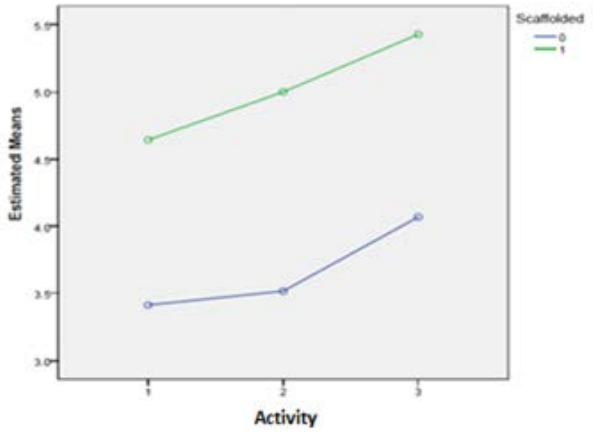


Figure 15: School 2, Grade 8 - Student Performance in the Density Microworld on the Interpreting Data practice/skill (green = Rex condition, blue = No Rex condition)

One dataset from the free fall microworld – school 2, grade 8 – had a significant positive effect for Rex on the interpreting data practice/skill ($p = .011$). In this dataset, students in the Rex condition increased their scores in interpreting data by 0.8 (compared to a gain of 0.5 by students in the No Rex condition).

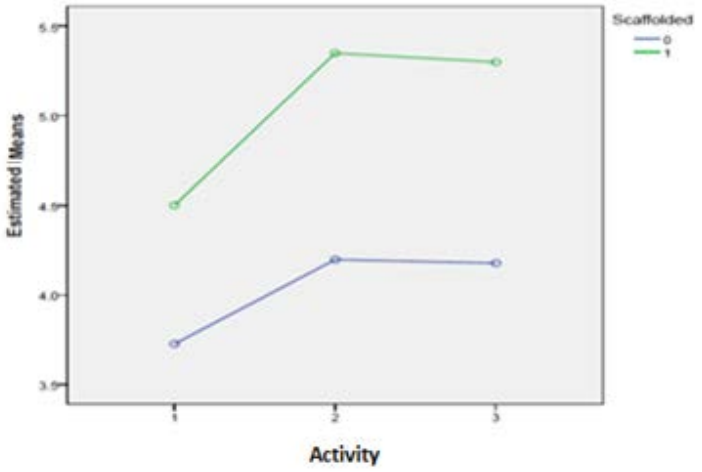


Figure 16: School 2, Grade 8 - Student Performance in the Free Fall Microworld on the Interpreting Data practice/skill (green = Rex condition, blue = No Rex condition)

One dataset from the phase change microworld – school 2, grade 7 – had a significant positive effect for Rex on the interpreting data practice/skill ($p < .001$). In this dataset, students in the Rex condition increased their scores in interpreting data by 1.3 (compared to a gain of 0.6 by students in the No Rex condition).

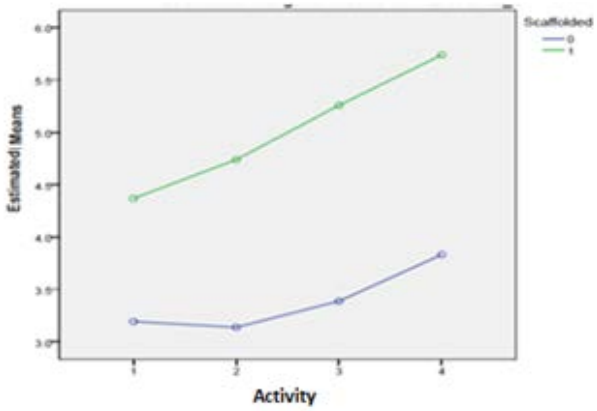


Figure 17: School 2, Grade 7 - Student Performance in the Phase Change Microworld on the Interpreting Data practice/skill (green = Rex condition, blue = No Rex condition)

Significant, positive effects for Rex on Warranting Claims

Students in the Rex condition from the datasets discussed in this section outperformed students in the No Rex condition on their performance of the warranting claims aggregated practice/skill. That is, they exhibited higher gains as they moved through the microworld from the first activity to the last, i.e., students in the Rex condition had a higher overall increase in the evaluation of their warranting claims score.

The first dataset is that from School 2 in Collisions. This dataset had a significant positive effect for Rex, p = .002, on the warranting claims practice/skill. Students in the Rex condition in this dataset increased their scores on warranting claims by 1.8 (compared to an overall increase of 0.2 by students in the No Rex condition).

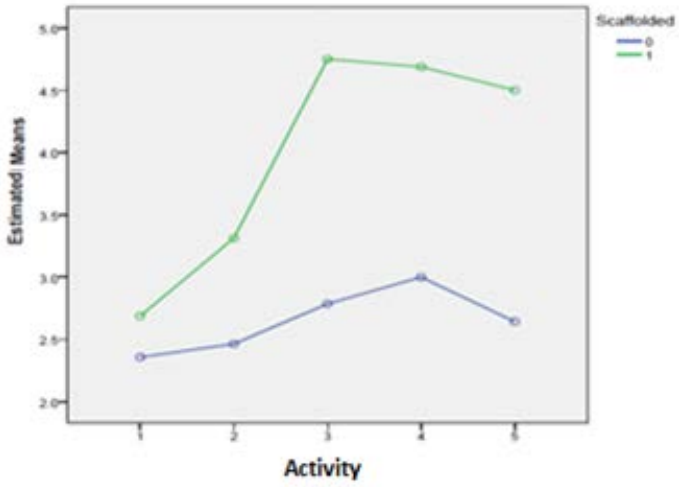


Figure 18: Student Performance in the Collisions Microworld on the Warranting Claims (green = Rex condition, blue = No Rex condition)

One dataset from the free fall microworld – school 2, grade 8 – had a significant positive effect for Rex on the warranting claims practice/skill (p = .005). For this dataset, students in the

Rex condition increased their scores in warranting claims by 1.1 (compared to a gain of 0.3 by students in the No Rex condition).

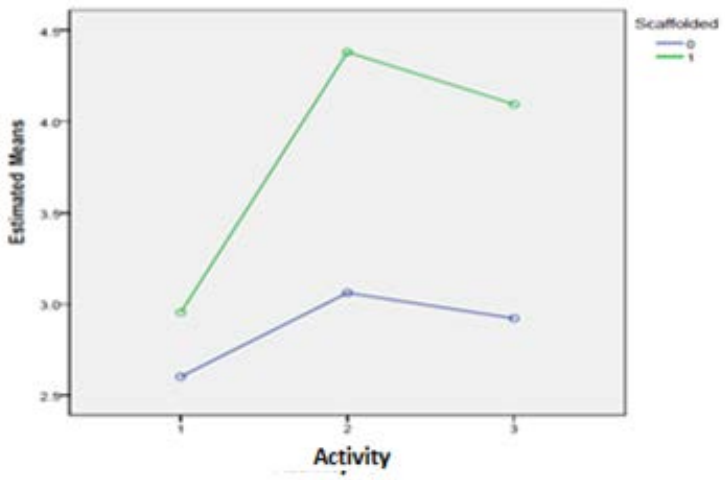


Figure 19: School 2, Grade 8 - Student Performance in the Free Fall Microworld on Warranting Claims (green = Rex condition, blue = No Rex condition)

Two datasets from the phase change microworld – school 1 and school 2, grade 7 – had a significant positive effect for Rex on the warranting claims practice/skill (school 1: $p = .041$; school 2, grade 7: $p < .001$). For the dataset from school 1, students in the Rex condition increased their scores in warranting claims by 0.4 (compared to a gain of 0.1 by students in the No Rex condition). For the dataset from school 2, grade 7, students in the Rex condition increased their scores in warranting claims by 2.1 (compared to a gain of 0.7 by students in the No Rex condition).

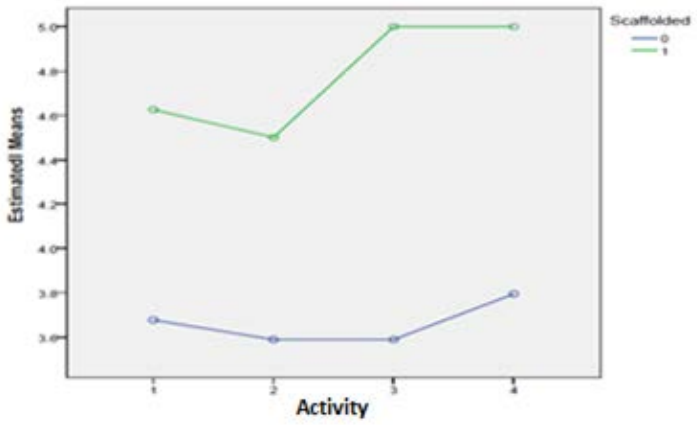


Figure 20: School 1 - Student Performance in the Phase Change Microworld on the Warranting Claims practice/skill (green = Rex condition, blue = No Rex condition)

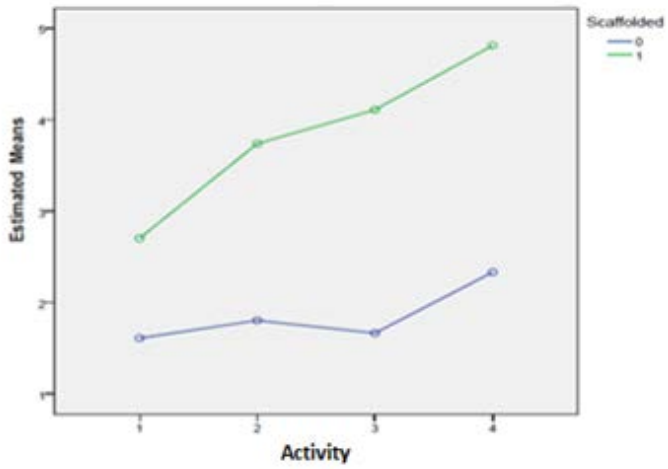


Figure 21: School 2, Grade 7 - Student Performance in the Phase Change Microworld on the Warranting Claims practice/skill (green = Rex condition, blue = No Rex condition)

Other (Non-significant) Results for Rex on Interpreting Data and Warranting Claims

In some of the other datasets, even though the effect of Rex was not significant, the results were still notable in helping to understand how students in the Rex condition were performing.

The dataset from school 1 in the free fall activity did not result in a significant effect for Rex. However, throughout the microworld, students in the Rex condition outperformed students

in the No Rex condition. Furthermore, although students in both conditions decreased their scores in interpreting data during the third activity, students in the Rex condition demonstrated a smaller decrease. These students decreased their scores in interpreting data by 0.1 (compared to a decrease of 0.7 by students in the No Rex condition). This shows that even when all the students encountered difficulty, students in the Rex condition still were better at transferring their knowledge of the evaluated inquiry practices.

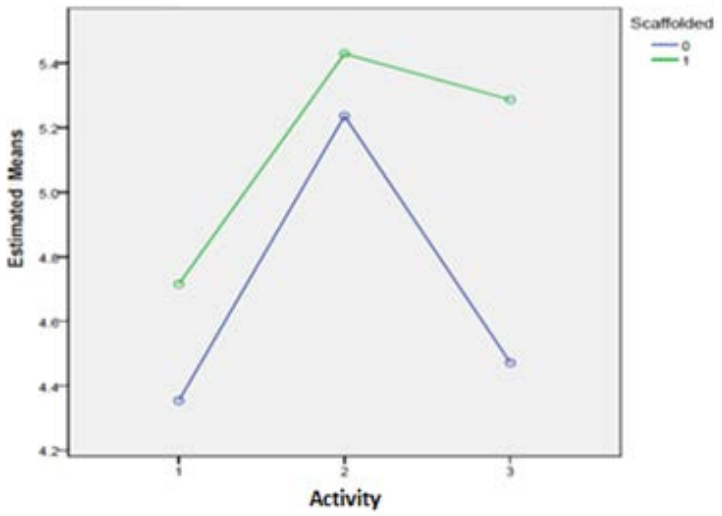


Figure 22: School 1 - Student Performance in the Free Fall Microworld on the Interpreting Data practice/skill (green = Rex condition, blue = No Rex condition)

The same trend was seen on students' warranting claims evaluations. Once again, although students in both conditions decrease their scores on warranting claims during the third activity, students in the Rex condition demonstrated a smaller decrease. These students decreased their scores on warranting claims by 0.2 (compared to a decrease of 0.6 by students in the No Rex condition).

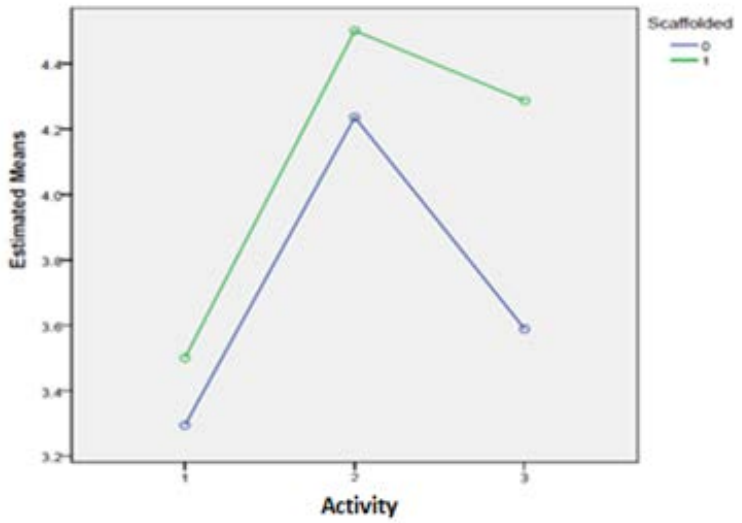


Figure 23: School 1 - Student Performance in the Free Free Fall Microworld on the Warranting Claims practice/skill (green = Rex condition, blue = No Rex condition)

In another result, students in the Rex condition in school 2, grade 8 in the density microworld exhibited a positive trend in their evaluations and increased their scores on warranting claims by 1.2 (compared to an increase of 0.7 by students in the No Rex condition), despite the non-significant effect of Rex.

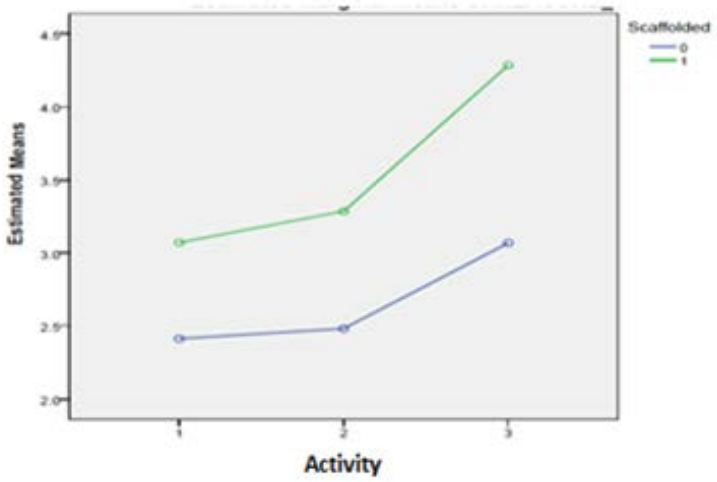


Figure 24: School 2, Grade 8 - Student Performance in the Density Microworld (green = Rex condition, blue = No Rex condition)

8.4 Discussion

In this study, the scaffolds designed and piloted in Study 1 and classroom tested in Study 2 for efficacy were evaluated in terms of their efficacy in aiding in the transfer of interpreting data and warranting claims subskills as students move within a domain. Overall, the results from this analysis suggest that data interpretation scaffolding can help students acquire and transfer both the interpreting data and warranting claims practices/skills through multiple activities within a single microworld. While in many of the activities and microworlds, students in both the Rex and No Rex condition are shown to acquire and transfer these practices/skills since the students in the No Rex condition are still learning as they are working through the microworld, students in the Rex condition are generally shown to have higher gains in their scores from the first to the last activity.

Even when the analyses showed student performance decreasing from one activity to the next, students in the Rex condition exhibited smaller decreases. Furthermore, these dips in performance can be attributed to the content difficulty of the activity in which students are working (Clement et al., 1989; Eryilmaz, 2002; Nickerson, 1998).

For example, students show a dip in their performance for both evaluated practices in the third activity in Free Fall (Figures 22 and 23), with a driving question of: Determine how the mass of the ball affects the time to drop. The effect of mass on time to drop is a classic misconception, i.e. preconception (c.f. Clement et al., 1989) that students have in this content area, as students believe that heavier items take longer to hit the ground (Eryilmaz, 2002) and often struggle to understand that physical properties of an object do not affect the speed of free fall (Halloun & Hestenes, 1985). This can go undetected in the classroom due to the emphasis on formula-based learning and assessment (Clement et al., 1989) and is likely to be exhibited in an

authentic inquiry environment. Therefore, given the driving question relating mass of the ball and time to drop, students may state that a change in the mass affects the time to drop, even though their data will show this to be incorrect. In other words, students may be exhibiting confirmation bias. This type of incorrect claim and data analysis could affect students' interpreting data and warranting claims scores, resulting in a dip in performance as seen in the third activity.

Similarly, students from school 1 show a decrease in performance for both evaluated practices in the second activity in Phase Change (Figure 20), with a driving question of: Determine how the amount of ice affects the boiling point of water. This driving question centers around a classic misconception that students have regarding the concepts of heat and temperature in phase change (Abraham et al., 1992) and the difficulty they have with being able to explain the physical phenomenon that accompanies the boiling of liquids (Nakleh, 1992). Many students think that the amount of ice affects the boiling point of water and that by adding more ice to the container, the boiling point will be higher, even though they have valid evidence that this is not the case. This type of incorrect claim and data analysis would hinder students' inquiry and their interpreting data and warranting claims scores would decrease as seen in the second activity.

These results suggest that the practices/skills of interpreting data and warranting claims, while transferable within a domain, are not completely domain general skills and can be affected by domain knowledge and misconceptions (i.e., alternative conceptions, preconceptions). These analyses show that students, regardless of condition, performed differently in each of the two microworlds. If the two evaluated practices were wholly domain-general, one would expect students to show a similar trajectory in the development and transfer of these practices in both microworlds, which was not the case. Instead, the trends of the students' data in each microworld

differ and student difficulty with the practices in each microworld corresponds to the documented misconceptions specific to that domain. This suggests that domain knowledge and domain misconceptions affect students' competencies at interpreting data and warranting claims. This, in turn, suggests that these practices cannot be classified as wholly domain-general. Rather, as suggested by previous research (e.g., Kuhn et al., 1992; McNeill & Krajcik, 2009), these practices may lie between a domain-generality/specificity and have some aspects of both.

9 – Discussion

This work built on the prior work on explanation and argumentation, two of the main goals of science (Windschitl, Thompson, & Braaten, 2008) and a key practice in national science standards (McNeill & Krajcik, 2011; Kang, Thompson, & Windschitl, 2014). These practices/skills are important for students to develop as they have been found to be an essential aspect of scientific thinking and direct attention to the development of these practices/skills has been found to be necessary (Kuhn et al., 2008; McNeill & Krajcik, 2011). It is also important to support the acquisition of these practices/skills as prior research has shown that explanation/argumentation can be difficult for many students (McNeill & Krajcik, 2011; Schunn & Anderson, 1999; Kuhn, Schauble, Garcia-Mila, 1992; Kuhn, 1991).

This goal of this work was to apply a scaffolding approach to the data interpretation and warranting subskills underlying explanation and present within the Inq-ITS environment for scientific inquiry in order to facilitate students' acquisition and transfer of these practices/skills.

Specifically, this work addressed the:

1. Design of scaffolds for data interpretation and warranting subskills as informed by prior research

2. Efficacy of these data interpretation scaffolds in supporting students to learn data interpretation and warranting subskills
3. Transfer of data interpretation and warranting subskills from one topic to another and how this transfer is best captured and measured, as well as which data interpretation and warranting subskills are most impacted by the use of data interpretation scaffolds

The development and testing of these data interpretation scaffolds was done in an iterative approach across three studies.

- *Study 1*: The design of scaffolds for data interpretation and warranting subskills as informed by prior research
- *Study 2*: Assessing the efficacy of data interpretation scaffolds in supporting students to learn data interpretation and warranting subskills
- *Study 3*: Assessing the transfer of data interpretation and warranting practices/skills within a domain

The contributions of this work are: (1) The conceptualization of data interpretation and warranting subskills as practices underlying the explanation/argumentation practices necessary for communicating science findings; (2) The design and iterative refinement of automated real-time scaffolding to address and support the data interpretation and warranting subskills of interest; (3) The application and modification of the Bayesian Knowledge Tracing model for assessing student acquisition of data interpretation and warranting skills with and without scaffolding; and (4) The transfer and domain-generality of data interpretation and warranting skills.

9.1 Contributions

Inquiry subskills

When it comes to unpacking the broad components of explanation, Toulmin's (1958) model of argumentation is typically used (McNeill & Krajcik, 2011; Gotwals & Songer, 2009; Kang, Thompson, & Windschitl, 2014; Berland & Reiser, 2009). This model breaks down explanation into three main components: the use of claims, evidence, and reasoning. Thus, the data interpretation and warranting practices (NGSS Lead States, 2013) of creating a claim and using evidence to warrant that claim are both critical components of explanation processes. To that end, as part of this work, the data interpretation and warranting claims skills/practices used here were conceptualized/framed as underlying the explanation practices necessary for communicating science findings.

It is important to bring together data interpretation and warranting subskills with the overall explanation practices because if students are having problems analyzing their data, they will not be able to engage successfully in explanation. Students have to be able to address the subcomponents of explanation – such as analyzing data and warranting claims – before they can successfully do explanation. Therefore, it is important to unpack and concretize the subskills of data interpretation and warranting claims within the frame of explanation so that it gives us the nuance to examine these complex practices (as set forth by NGSS) and allows us to look at specifically what aspects students are having difficulty with and work to target those exact difficulties.

As such, one of the main goals for the first study of this work was to identify the different subskills in data interpretation and warranting. As part of this work, a categorization of five types

of difficulties in data interpretation and warranting was created and refined using pilot data from middle-school students and then aligned with the dichotomous subskills assessed in Inq-ITS.

In this way, we were able to support the components of students' data interpretation and warranting – including their creation of a claim and use of evidence – in an automated and fine-grained way with appropriate subskills that allowed us to target known difficulties documented by previous research (Gotwals & Songer, 2009; McNeill & Krajcik, 2011; Schunn & Anderson, 1999) and support students who are having problems with specific aspects of their data interpretation and warranting, which allows them to build their knowledge and skills such that they can in turn successfully engage in the overall practices of explanation.

Automated scaffolding tied to subskills

This work also built on work done in other environments by scaffolding data interpretation and warranting subskills in real-time and at multiple levels. Prior research showed that scaffolding aids with the acquisition of inquiry practices/skills (Roll, Yee, & Briseno, 2014) and the use of computer-based support can further help facilitate this acquisition. Specifically, when it comes to data interpretation and warranting, scaffolding can support articulation, reflection, and the revision of explanations by facilitating ongoing articulation through the externalization of ideas, findings, or interpretations; supporting explanation building and hypothesis generation; and providing opportunities for learners to organize, reflect upon, and revise their project artifacts (Land & Zembal-Saul, 2003). Furthermore, when these scaffolds are automatic and fired in real-time, they can help students in their inquiry processes by: preventing them from engaging in ineffective inquiry behaviors (Buckley et al., 2006; Gobert & Schunn, 2007); offering help even if they may not be aware that they need help (Aleven & Koedinger, 2000; Aleven et al., 2004);

personalizing the learning processes by providing multiple levels of automatic scaffolds and allowing students to request for further help or clarification, as needed; and providing both scalable assessment and guidance. However, much of the existing scaffolding in learning environments is provided in the form of text-based worksheets, or in some other similar form that is either not scalable, fine-grained, or personalized. Additionally, these approaches typically require a student to know when they need help; however, students may not have the metacognitive skills needed to recognize this (Aleven & Koedinger, 2000; Aleven et al., 2004)

This work addressed these drawbacks by using a computer-based scaffolding approach aimed at inquiry subskills so that learner support can be utilized automatically and in real-time in order to better support students in their inquiry. The scaffolding for this work was integrated into the Inq-ITS system in such a way that tied it to the automatic and ongoing evaluations of the inquiry subskills. Thus, these scaffolds were triggered not by a fixed schedule decided by a designer or by students when they believed they need help, but by the actual inquiry products the students exhibited such that students had to demonstrate successfully that they no longer needed the scaffolds and could perform the inquiry practice/skill correctly before they stopped receiving the scaffolds. This allowed for personalization to each student's learning, recognizing that different students may need different amounts of help to hone successfully different subskills.

Furthermore, the overall scaffolding strategy in this work was to provide students with more and more targeted support through multiple levels of scaffolding (Corbett & Anderson, 1995; Koedinger & Corbett, 2006; Anderson et al., 1995). Therefore, the end product of a set of automatic, real-time scaffolds that address data interpretation and warranting as science inquiry practices/skills with different levels that adapt to students' difficulties and that are tied to

assessment is a major contribution of this work and addresses a gap in previous work in scaffolding.

Application of BKT as a method for measuring student acquisition

This work also contributes towards the development of a fine-grained method for unpacking the effect of scaffolding via logged, process data. In study 2, data was collected through a randomized control study and an extension of the Bayesian Knowledge Tracing (BKT) framework that considers the presence of scaffolding as a cognitive modeling approach for approximating mastery learning of the inquiry subskills of interest (Sao Pedro et al., 2013b) was used. Modifications were also developed to this framework, which allowed it to be applied when condition and practice opportunity can be defined on different levels (i.e. activity level vs. subskill level).

This study was done within a complex domain of science inquiry whereby the student data, number of practice opportunities, and evaluated skills were not as clearly delineated as in previous studies in which BKT was used to evaluate educational interventions. If Inq-ITS was built to evaluate students after every action they took (e.g. evaluate after the student chooses an IV, evaluate after a student chooses a DV, etc.), the assessment and scaffolding of these subskills would be much easier as it would not introduce any disentangling between subskills. However, this approach would not be representative of authentic inquiry where everything is done at once. Therefore, to keep the environment and the assessment authentic to science inquiry, in Inq-ITS all the subskills are evaluated at once. As previously mentioned, this then introduces a level of complexity to the BKT analysis as scaffolding condition and practice opportunity are not as clear

cut. As such, this work provides a framework for how data in these complex environments can be treated before BKT can be used.

Transfer of skills (domain generality)

This work also contributes to the understanding of the domain generality and transferability of data interpretation and warranting practices/skills. In study 3, the scaffolds designed and tested in studies 1 and 2 were assessed in terms of their efficacy in aiding in the transfer of the assessed interpreting data and warranting claims practices/skills as students move within a domain.

When it comes to categorizing specific skills as domain-general or domain-specific, there is some debate as to how distinct the categories of domain-generality and domain-specificity are (Greiff et al., 2014; McNeill & Krajcik, 2009; Schunn & Anderson, 1999). Many researchers believe that this is better described as a continuum from domain specific to domain general, rather than a dichotomy (Greiff et al. 2014; McNeill & Krajcik, 2009; Schunn & Anderson, 1999).

However, more research is needed regarding what constitutes a domain-general practice/skill, which could be tested by specifically testing the domain generality of different types of inquiry practices/skills (e.g. data collection vs. data interpretation), inquiry subskills (e.g., claim DV, claim IV, etc.), and topics within a domain (e.g. density vs. free fall), as well as what factors (e.g., content, domain familiarity, etc.) can facilitate or impede transfer.

The results from study 3 showed that data interpretation scaffolding can help students acquire the assessed interpreting data and warranting claims practices/skills and transfer them to a new activity within the same microworld. The results also showed differences among different

datasets/classes and domains/microworlds. The effect of prior knowledge and misconceptions could also be seen on student performance on certain activities within the microworlds.

These results built on prior work regarding the domain generality of data interpretation and warranting practices/skills and suggested that the practices/skills of interpreting data and warranting claims while transferable within a domain are not completely domain general and can be affected by domain knowledge and misconceptions (i.e., alternative conceptions, preconceptions). The analyses show that students (from different schools), regardless of condition, performed differently in each of the microworlds, suggesting that domain knowledge and domain misconceptions affect students' competencies at interpreting data and warranting claims and that these practices lie between a domain-generality/specificity spectrum. As such, this work adds to the growing literature on domain generality and adds evidence to the view that data interpretation and warranting practices/skills have aspects of both domain generality and domain specificity.

9.2 Limitations & Future Work

Accounting for Prior Experience in Study 3

The data used in Study 3 were collected from two different schools across the country. Data from one school consisted of 8th grade students from two teachers' classes. These students were assigned three microworlds over the course of the year. Data from the other school consisted of one 8th grade teacher's students and one 7th grade teacher's students. The 8th grade students were assigned three microworlds over the course of the year and the 7th grade students were assigned two microworlds over the course of the year. Each data set from each of the microworlds was analyzed separately in order to account for the fact that different microworlds were completed at

different times. This analysis did not consider students' prior experience with scaffolding. While this allowed for a domain-specific analysis of student performance across driving questions, as a limitation of this work, it ignored any potential effects from prior experience with the data interpretation and warranting claims scaffolds. .

A potential analysis consideration for this work could have been to add a factor to the statistical model that accounts for any prior experience with scaffolding. However, these data sets do not contain enough student data to statistically account for this. For example, in the data sets from the 8th grade students in school 2, there were 89 students who were assigned the three microworlds. However, of these 89 students, only 18 students completed all three of the assigned microworlds. Of the remaining students, 33 students completed two microworlds and 38 students completed only one microworld.

Furthermore, of the students who completed two microworlds and the students who completed one microworld, there was variability in which microworld(s) they completed. The table below shows how there were three different ways the 33 students who completed two microworlds did so – some students completed the first two microworlds, some students completed the last two microworlds, and some students completed the first and third microworlds.

Table 14: Students who completed two microworlds

	Density	Free Fall	Collisions
Student 1	✓	✓	✗
Student 2	✓	✗	✓
Student 3	✗	✓	✓

Similarly, the table below shows how there were three different ways the 38 students who completed one microworld did so – some students completed the first microworld, some students completed the last microworld, and some students completed the second microworld.

Table 15: Students who completed one microworld

	Density	Free Fall	Collisions
Student 1	✓	✗	✗
Student 2	✗	✗	✓
Student 3	✗	✓	✗

The variability in both the number of completed microworlds and in which microworlds were completed complicates the data as it introduces the variable of time to students' prior experiences. While a student can be coded as having had prior experience, that prior experience could be experience from one microworld or two microworlds, but it also could be experience from a microworld directly prior to the microworld being analyzed or a microworld completed before that. Furthermore, when prior experience with scaffolding condition is included, even more variability is introduced. For example, looking at the students who completed two microworlds, it is possible that they received scaffolding in neither, both, or one of those microworlds. This makes it even clearer that prior experience with scaffolding is very much variable by individual student and there are not enough students in each condition to statistically account for any effects. Therefore, accounting for prior experience in this situation is complex and not covered in this work as a limitation. Future analyses should look at data sets in which

students have either no prior experience or the same prior experience in order to better capture the effects of scaffolding in each domain.

Measuring and Supporting Reasoning Subskills

This work focused on the first two parts of explanation – claim and evidence – and did not address the third part of explanation – reasoning – as it was beyond the scope of this work to address students' reasoning. However, students' reasoning is an important part of their inquiry and an important practice in explanation. While prior research has shown that even young children have the basic tools necessary for inquiry and are capable of developing new skills in scientific reasoning (Windschitl, 2000; Schauble, Glaser, Duschl, Schulze, & John, 1995), students often display difficulty providing reasoning for their claims (McNeill & Krajcik, 2011). This is due to the fact that in order to gain expertise in reasoning, students must be able to detect complex patterns, deal with abstract objects and phenomena, and use models (Windschitl, 2000). The difficulty behind reasoning may also stem from the fact that scientific reasoning explicitly entails the coordination of existing theories with new evidence bearing on them (Kuhn et al., 1992).

As such, reasoning as an inquiry practice/skill should be examined in future work the same way that claim and evidence were examined here; i.e., the subskills in reasoning should be identified, assessed, and scaffolded. Other work (e.g. Li, Gobert, & Dickler, 2017) is currently addressing this.

Transferring Across Domains in Inq-ITS

This work focused on skill and subskill transfer within one domain in Inq-ITS and did not cover transfer of skills and/or subskills across multiple domains. More research is needed on what constitutes a domain-general practice/skill and, more specifically, if there are differences in domain generality at the subskill level, which could affect how students can transfer these practices from one topic to another.

This could be tested in future work by specifically testing the domain generality of different types of inquiry practices/skills (e.g., data collection vs. data interpretation) and inquiry subskills (e.g., claim DV, claim IV, etc.) across domains (e.g., density to free fall). The specific research questions for such a study would be:

1. Can data interpretation scaffolding help students acquire the assessed interpreting data and warranting claims practices/skills and/or subskills and transfer them to a new microworld?
2. Does students' transfer of interpreting data and warranting claims practices/skills and/or subskills change as they move between different microworlds? (i.e., Is transfer the same between the density and phase change microworlds as between the phase change and free fall microworlds).

By examining transfer of skills between different domain pairs, we can see if skills are more easily transferable to certain domains. For example, students who work in the density microworld may be better able to transfer the assessed skills to the phase change microworld rather than the free fall microworld. These analyses would examine which microworld pairs, if any, are better suited to transfer.

3. Are there any differences in transfer between the practices/skills and/or subskills?

It is possible that either the interpreting data or warranting claims skill or a specific subskill therein, due to specific difficulty or domain-dependency, may be more difficult for students to transfer to a new microworld topic.

Transfer Outside of Inq-ITS

One of the main considerations of this work is that it exists within the Inq-ITS environment. This inherently affects the type of work students can do and consequently the type of data that can be collected. As previously mentioned, Inq-ITS has built-in implicit scaffolds such as widgets and drop-down menus to help focus students on the inquiry practices/skills of interest. For example, the widgets in use for the claim make it so that students do not need to worry about creating their own sentence structure and can focus on choosing the appropriate variables and relationships. This does mean that students are never asked to practice writing their own claims from scratch as they are in other environments, such as Belvedere, and as they may be asked to do in their actual classroom environments.

However, I would argue that the subskills and processes of inquiry that students engage with in Inq-ITS are key foundational pieces of inquiry as a whole – the ability to identify an IV, identify a DV, specify the relationship between those variables, and use proper data to warrant a claim will always be key elements of inquiry regardless of what other layers are added (i.e., communicating a claim through writing, including reasoning, etc.). Therefore, the better students are able to master these subskills, the better foundation they have in doing inquiry in other environments that are less scaffolded. That being said, future work should look at students' mastery of these subskills in the Inq-ITS environment and see how they transfer to environments

outside of Inq-ITS – such as another virtual environment that is more open-ended or a classroom assignment.

References

AAAS. (1993). *Benchmarks for science literacy.*

Abraham, M. R., Grzybowski, E. B., Renner, J. W., & Marek, E. A. (1992). Understandings and misunderstandings of eighth graders of five chemistry concepts found in textbooks. Journal of Research in Science Teaching, 29(2), 105-120.

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In *Intelligent tutoring systems* (pp. 292-303). Berlin Heidelberg: Springer.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking. In *Intelligent tutoring systems.* Berlin Heidelberg: Springer.

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

Anderson, R. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1-12.

Baker, R., Corbett, A., Gowda, S., Wagner, A., MacLaren, B., Kauffman, L., et al. (2010). Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. In 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.S.J.d., Gowda, S., & Corbett, A.T. (2011). Towards predicting future transfer of learning. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 23-30.

Baker, R.S.J.d., Gowda, S.M., & Corbett, A.T. (2011). Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. *Proceedings of the 4th International Conference on Educational Data Mining*, 179-188.

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(1), 153.

Beck, J.E., Chang, K., Mostow, J., & Corbett, A. (2008). Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. *Intelligent Tutoring Systems*, 383-394.

Beck, J. E., & Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Intelligent Tutoring Systems* (pp. 353-362). Springer Berlin Heidelberg.

Beck, J.E., & Sison, J. (2006). Using Knowledge Tracing in a Noisy Environment to Measure Student Reading Proficiencies. *International Journal of Artificial Intelligence in Education*, 16(2), 129-143.

Bell, P. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE.

Bell, P. (2004). Promoting students' argument construction and collaborative debate in the science classroom. *Internet environments for science education*, 115-143.

Bell, P., & Davis, E. A. (2000). Designing Mildred: Scaffolding students' reflection and argumentation using a cognitive software guide. In *Fourth international conference of the learning sciences* (pp. 142-149). Mahwah, NJ: Erlbaum.

Belland, B.R. (2010). Portraits of middle school students constructing evidence-based arguments during problem-based learning: The impact of computer-based scaffolds. *Educational technology research and development,* 58(3), 285-309.

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. In *Handbook of research on educational communications and technology* (pp. 505-518). Springer New York.

Belland, B. R., Glazewski, K. D., & Richardson, J. C. (2011). Problem-based learning and argumentation: Testing a scaffolding framework to support middle school students' creation of evidence-based arguments. *Instructional Science*, *39*(5), 667-694.

Belland, B. R., Walker, A. E., Olsen, M. W., & Leary, H. (2015). A Pilot Meta-Analysis of Computer-Based Scaffolding in STEM Education. *Journal of Educational Technology & Society, 18*(1), 183-197.

Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation.*Journal of Research in Science Teaching*, *49*(1), 68-94.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26-55.

Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education, 92*(3), 473-498.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational researcher, 18*(1), 32-42.

Bruner, J. (1961). The act of discovery. *Harvard Educational Review, 31*(1), 21-32.

Buckley, B., Gobert, J., & Horwitz, P. (2006). Using log files to track students' model-based inquiry. *In Proceedings of the 7th International Conference of the Learning Sciences* (pp. 57-63). International Society of the Learning Sciences.

Carr, C. S. (2000). *The effect of computer-supported collaborative argumentation (CSCA) on argumentation skills in second-year law students*(pp. 1-152). The Pennsylvania State University.

Cerbin, B. (1988). *The nature and development of informal reasoning skills in college students.*

Chinn, C.A. & Clark, D. B. (2013). Learning through collaborative argumentation. *The international handbook of collaborative learning*, 314.

Chinn, C. A., Duschl, R. A., Duncan, R. G., Buckland, L. A., & Pluta, W. J. (2008). A microgenetic classroom study of learning to reason scientifically through modeling and argumentation. *Proceedings of the 8th international conference on International conference for the learning sciences* (pp. 14-15). International Society of the Learning Sciences.

Chinn, C., & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*, 1-49.

Chinn, C., & Malhotra, B. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology, 94*(2), 327.

Cho, K. L., & Jonassen, D. H. (2002). The effects of argumentation scaffolds on argumentation and problem solving. *Educational Technology Research and Development*, *50*(3), 5-22.

Clement, J., Brown, D. E., & Zietsman, A. (1989). Not all preconceptions are misconceptions: finding 'anchoring conceptions' for grounding instruction on students' intuitions. International journal of science education, 11(5), 554-565.

Collins, A. (1996). Design issues for learning environments. In S. Vosniadou, E.D. Corte, R. Glaser, & H. Mandl (Eds.), *International perspectives on the design of technology-supported learning environments* (pp. 347-362). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Davis, E. A. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education*, *22*(8), 819-837.

de Jong, T. (2006). Technological Advances in Inquiry Learning. *Science*.

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., Beishuizen, J. J., Hulshof, C. D., Prins, F., Rijn, H., Someren, M. W., . . . Wilhem, P. (2005). Determinants of discovery learning.

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.* Washington, D.C.: National Academies Press.

Eryilmaz, A. (2002). Effects of conceptual assignments and conceptual change discussions on students' misconceptions and achievement regarding force and motion. Journal of research in science teaching, 39(10), 1001-1015.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., . . . Strijbos, J. W. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28-45.

Giere, R. (1991). *Understanding scientific reasoning.* Fort Worth, TX: Holt, Rinehart, & Winston.

Gijlers, H., & de Jong, T. (2009). Sharing and confronting propositions in collaborative inquiry learning. *Cognition and Instruction*, *27*(3), 239-268.

Gobert, J., & Schunn, C. (2007). Supporting inquiry learning: a comparative look at what matters. *In a symposium presented at the Annual Meeting of the American Educational Research Association.* Chicago, IL.

Gobert, J., Sao Pedro, M., Baker, R., Toto, E., & Montalvo, O. (2012). Leveraging Educational Data Mining for Real Time Performance Assessment of Scientific Inquiry Skills within Microworlds. *Journal of Educational Data Mining, 4*(1), 111-143.

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

Gotwals, A. W., & Songer, N. B. (2009). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*(2), 259-281.

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, *13*, 74-83.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. American journal of Physics, 53(11), 1043-1055.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Harrison, A., & Schunn, C. (2004). The transfer of logically general scientific reasoning skills. *Proceedings of the Cognitive Science Society*, *26.*

Hsu, Y. S., Lai, T. L., & Hsu, W. H. (2014). A Design Model of Distributed Scaffolding for Inquiry-Based Learning. *Research in Science Education, 45*(2), 241-273.

Kammerer, Y., Meier, N., & Stahl, E. (2016). Fostering secondary-school students' intertext model formation when reading a set of websites: The effectiveness of source prompts. *Computers & Education*, *102*, 52-64.

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.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: the role of scaffolding in assessment tasks. *Science Education, 98*(4), 674-704.

Kim, M.C., & Hannafin, M.J. (2011). Scaffolding 6th graders' problem solving in technology-enhanced science classrooms: A qualitative case study. *Instructional Science*, 39(3), 255-282.

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., and Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*. 15 (10).

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review, 94*(2), 211.

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.

Kollar, I., Fischer, F., & Slotta, J. (2007). Internal and external scripts in computer-supported collaborative inquiry learning. *Learning and Instruction, 17*(6), 708-721.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological review*, *96*(4), 674.

Kuhn, D. (1991). *The skills of argument.* Cambridge University Press.

Kuhn, D. (2005). *Education for Thinking.* Harvard University Press.

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills.* Academies Press.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development, 23*(4), 435-451.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills?. *Cognition and Instruction*, *26*(4), 512-559.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction, 9*(4), 285-327.

Kyza, E.A. (2009). Middle-school students' reasoning about alternative hypotheses in a scaffolded, software-based inquiry investigation. *Cognition and Instruction*, 27(4), 277-311.

Lajoie, S. P., Guerrera, C., Munsie, S. D., & Lavigne, N. C. (2001). Constructing knowledge in the context of BioWorld. *Instructional Science*,*29*(2), 155-186.

Land, S. M., & Zembal-Saul, C. (2003). Scaffolding reflection and articulation of scientific explanations in a data-rich, project-based learning environment: An investigation of progress portfolio. *Educational Technology Research and Development, 51*(4), 65-84.

Larman, C., & Basili, V. (2003). Iterative and incremental developments: A brief history. *Computer, 36*(6), 47-56.

Lazonder, A. W., Hagemans, M. G., & De Jong, T. (2010). Offering and discovering domain information in simulation-based inquiry learning. *Learning and Instruction*, *20*(6), 511-520.

Li, D. D., & Lim, C. P. (2008). Scaffolding online historical inquiry tasks: A case study of two secondary school classrooms. *Computers & Education, 50*(4), 1394-1410.

Li, H., Gobert, J., & Dickler, R. (2017) Dusting Off the Messy Middle: Assessing Students' Inquiry Skills Through Doing and Writing. In Proceedings of the 18th International Conference on Artificial Intelligence in Education. Wuhan, China (pp. 175-187).

Loh, B., Radinsky, J., Reiser, B. J., Gomez, L. M., Edelson, D. C., & Russell, E. (1997, December). The progress portfolio: Promoting reflective inquiry in complex investigation environments. In *Proceedings of the 2nd international conference on Computer support for collaborative learning* (pp. 176-185). International Society of the Learning Sciences.

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

McElhaney, K., & Linn, M. (2008). Impacts of students' experimentation using a dynamic visualization on their understanding of motion. *Proceedings of the 8th International Conference for the Learning Sciences* (pp. 51-58). International Society of the Learning Sciences.

McElhaney, K., & Linn, M. (2010). Helping students make controlled experiments more informative. *Proceedings of the 9th International Conference of the Learning Sciences* (pp. 786-793). International Society of the Learning Sciences.

McGinn, M. K., & Roth, W. M. (1999). Preparing students for competent scientific practice: Implications of recent research in science and technology studies. *Educational Researcher, 28*(3), 14-24.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human-computer interaction, 5*(4), 381-413.

McNeill, K., & Krajcik, J. (2006). Supporting students' construction of scientific explanation through generic versus context-specific written scaffolds. *Annual Meeting of the American Educational Research Association.* San Francisco, CA.

McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. *Thinking with data*, 233-265.

McNeill, K. L., & Krajcik, J. (2009). Synergy between teacher practices and curricular scaffolds to support students in using domain-specific and domain-general knowledge in writing arguments to explain phenomena. *The journal of the learning sciences*, *18*(3), 416-460.

McNeill, K., & Krajcik, J. (2011). *Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing.* Pearson.

Metcalf, S. (1999). *The design of guided learner-adaptable scaffolding in interactive learning environments (Ph.D. Dissertation).* University of Michigan, Ann Arbor, MI.

Millar, R., & Lubben, F. (1996). Investigative work in science: The role of prior expectations and evidence in shaping conclusions. *Education, 24*(1), 28-34.

Moore, E. B., Herzog, T. A., & Perkins, K. K. (2013). Interactive simulations as implicit support for guided-inquiry. *Chemistry Education Research and Practice, 14*(3), 257-268.

Nakhleh, M. B. (1992). Why some students don't learn chemistry: Chemical misconceptions. Journal of Chemistry Education, 69(3), 191.

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

Nelson, B., Ketelhut, D. J., Clarke, J., Bowman, C., & Dede, C. (2005). Design-based research strategies for developing a scientific inquiry curriculum in a multi-user virtual environment. *Educational Technology, 45*(1), 21-27.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States .* Washington, DC: The National Academies Press.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2*(2), 175.

Paolucci, M., Suthers, D., & Weiner, A. (1996, June). Automated advice-giving strategies for scientific nquiry. In *International Conference on Intelligent Tutoring Systems* (pp. 372-381). Springer, Berlin, Heidelberg.

Pardos, Z. A., & Heffernan, N. T. (2010, June). Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 255-266). Springer, Berlin, Heidelberg.

Pea, R. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences, 13*(3), 423-451.

Puntambekar, S., & Hubscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational psychologist, 40*(1), 1-12.

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., . . . Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337-386.

Reiser, B. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences, 13*(3), 273-304.

Roll, I., Yee, N., & Briseno, A. (2014, June). Students' adaptation and transfer of strategies across levels of scaffolding in an exploratory environment. *International Conference on Intelligent Tutoring Systems* (pp. 348-353). Springer, Cham.

Sadler, T. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of research in science teaching, 41*(5), 513-536.

Sandoval, W. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences, 12*(1), 5-51.

Sandoval, W. A., & Reiser, B. J. (1997). Evolving Explanations in High School Biology.

Sandoval, W., & Reiser, B. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*(3), 345-372.

Sao Pedro, M., Baker, R., & Gobert, J. (2013). Incorporating Scaffolding and Tutor Context into Bayesian Knowledge Tracing to Predict Inquiry Skill Acquisition. *In S.K. D'Mello, R.A. Calvo, & A. Olney (Eds.) Proceedings of the 6th International Conference on Educational Data Mining*, (pp. 185-192). Memphis, TN.

Sao Pedro, M., Gobert, J., & Baker, R. (2014). Impacts of Automatic Scaffolding on Students' Acquisition of Data Collection Inquiry Skills. *Paper presented at the Annual Meeting of the American Education Research Association.* Philadelphia, PA.

Sao Pedro, M., Gobert, J., & Baker, R. (2012). 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*. Vancouver, BC, CA: Retrieved April 15, 2012, from the AERA Online Paper Repository. Best Student Paper Award - AERA SIG Advanced Technologies for Learning/Learning Sciences

Sao Pedro, M., Gobert, J., Heffernan, N., & Beck, J. (2009). Comparing Pedagogical Approaches for Teaching the Control of Variables Strategy. In N.A. Taatgen & H. vanRijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1294-1299). Austin, TX: 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*(1), 31-57.

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. *Journal of the Learning Sciences, 4*(2), 131-166.

Schauble, L., Klopfer, L. E., & 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., & Anderson, J. (1998). Scientific Discovery. *The atomic components of thought*, 385-427.

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

Toulmin, S. (1958). *The Uses of Argument.* Cambridge, England: Cambridge University Press.

van Joolingen, W. R., & de Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science, 25*(5), 307-346.

Voss, J. F., Wolfe, C. R., Lawrence, J. A., & Engle, R. A. (1991). From representation to decision: An analysis of problem solving in international relations.

Windschitl, M. (2000). Supporting the development of science inquiry skills with special classes of software. *Educational Technology Research and Development, 48*(2), 81-95.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education, 92*(5), 941-967.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry, 17*(2), 89-100.

Xun, G. E., & Land, S. M. (2004). A conceptual framework for scaffolding III-structured problem-solving processes using question prompts and peer interactions. *Educational Technology Research and Development*, *52*(2), 5-22.

Yore, L. D., Henriques, L., Crawford, B., Smith, L., Gomez-Zwiep, S., & Tillotson, J. (2007). Selecting and using inquiry approaches to teach science: The influence of context in elementary, middle, and secondary schools. *Inquiry in the classroom: Realities and Opportunities*, 39-87.