Chapter 8
Real-Time Scaffolding of Students' Online Data Interpretation During Inquiry with Inq-ITS Using Educational Data Mining

Janice D. Gobert, Raha Moussavi, Haiying Li, Michael Sao Pedro, and Rachel Dickler

Abstract This chapter addresses students' data interpretation, a key NGSS inquiry practice, with which students have several different types of difficulties. In this work, we unpack the difficulties associated with data interpretation from those associated with warranting claims. We do this within the context of Inq-ITS (Inquiry Intelligent Tutoring System), a lightweight LMS, providing computer-based assessment and tutoring for science inquiry practices/skills. We conducted a systematic analysis of a subset of our data to address whether our scaffolding is supporting students in the acquisition and transfer of these inquiry skills. We also describe an additional study, which used Bayesian Knowledge Tracing (Corbett and Anderson. User Model User-Adapt Interact 4(4):253–278, 1995), a computational approach allowing for the analysis of the fine-grained sub-skills underlying our practices of data interpretation and warranting claims.

J. D. Gobert (✉)
Department of Educational Psychology, Rutgers Graduate School of Education, New Brunswick, NJ, USA

Apprendis, Berlin, MA, USA
e-mail: janice.gobert@gse.rutgers.edu; janice@apprendis.com

R. Moussavi
Teaching Systems Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: moussavi@mit.edu

H. Li · R. Dickler
Department of Educational Psychology, Rutgers Graduate School of Education, New Brunswick, NJ, USA
e-mail: haiying.li@gse.rutgers.edu; rachel.dickler@gse.rutgers.edu

M. Sao Pedro
Apprendis, Berlin, MA, USA
e-mail: mikesp@apprendis.com

M. E. Auer et al. (eds.), *Cyber-Physical Laboratories in Engineering and Science Education*, https://doi.org/10.1007/978-3-319-76935-6_8

Keywords Scaffold · Inquiry learning · Intelligent tutoring system · Microworld · Virtual lab

8.1 Introduction

Science educators and policy makers (NGSS Lead States 2013; OECD 2014) agree that richly integrating authentic inquiry with science content will promote well-honed learning strategies and allow students to apply and transfer their science knowledge in more flexible ways as is needed for tomorrow's jobs (Hilton and Honey 2011). As a result, as schools in the United States adopt the Next Generation Science Standards (NGSS), educators will need to (1) incorporate more inquiry experiences into instruction, (2) assess their students' inquiry practices/skills, and (3) ensure that each student demonstrates adequate progress on these.

Meeting these goals however poses significant challenges (Fadel et al. 2007). First, educators may not have adequate time, lab space, and/or physical materials for inquiry (Staer et al. 1998), particularly in schools with large class sizes (e.g., in Oregon there can be 50 students in a class). Second, grading inquiry is difficult, subjective, and time-intensive (Deters 2005). Third, teachers need *immediate* and *actionable data* to identify *which* of the many types of difficulties students are experiencing (Kuhn 2005) in order to foster students' growth (Shute 2008), but current assessments yield data too late for teachers to impact students' learning (Pellegrino et al. 2001). Fourth, developing authentic inquiry tasks and assessments is difficult due to its multifaceted, ill-defined nature (Williamson et al. 2006), and as a result, there are too few empirically tested resources to assess and support inquiry (Krajcik et al. 2000; Schneider et al. 2005). Lastly, since inquiry practices need to be honed over time, students need to engage in authentic inquiry multiple times across the school year, and without an automated solution, the burden on teachers to do grading is extremely onerous.

To add to these issues, the most recent student data on international comparisons of science performances show that American students continue to fall behind their peers. For example, in 2015, the United States ranked 25th worldwide on a key educational survey called the Program for International Student Assessment (PISA; Organization for Economic Cooperation and Development 2018). This is no doubt related, at least in part, to the many student difficulties that have been demonstrated for all of the inquiry skills identified by NGSS (2013). Specifically, students have trouble *forming testable hypotheses* (Chinn and Brewer 1993; Klahr and Dunbar 1988; Kuhn et al. 1995; Njoo and de Jong 1993; van Joolingen and de Jong 1997; Glaser et al. 1992) and difficulty *testing their hypotheses* (van Joolingen and de Jong 1991b, 1993; Kuhn et al. 1992; Schauble et al. 1991). They have difficulty *conducting experiments* (Glaser et al. 1992; Reimann 1991; Tsirgi 1980; Shute and Glaser 1990; Kuhn 2005; Schunn and Anderson 1998, 1999; Harrison and Schunn 2004; McElhaney and Linn 2008, 2010).

When interpreting data during inquiry, a key NGSS inquiry practice and the one addressed in this chapter, students have several different types of difficulties. They may draw conclusions based on confounded data (Klahr and Dunbar 1988;

Kuhn et al. 1992; Schauble et al. 1995), state conclusions that are inconsistent with their data (Kanari and Millar 2004), change ideas about causality (Kuhn et al. 1992), and/or have difficulty in making a valid inference and reconciling previous conceptions with their collected data, falling back on prior knowledge (Schauble 1990; Kanari and Millar 2004), thereby exhibiting confirmation bias during inquiry (Klayman and Ha 1987; Dunbar 1993; Quinn and Alessi 1994; Klahr and Dunbar 1988). They also fail to relate the outcomes of experiments to the theories being tested in the hypothesis (Schunn and Anderson 1999; Chinn and Brewer 1993; Klahr and Dunbar 1988).

When warranting their claims with evidence, one of the five essential features of classroom inquiry per NRC's (National Research Council 2011), they often provide little to no justification (McNeill and Krajcik 2011; Schunn and Anderson 1999) and create claims that do not answer the question posed (McNeill and Krajcik 2011). Students can also rely on theoretical arguments rather than on experimental evidence during warranting (Kuhn 1991; Schunn and Anderson 1999).

Lastly, they have difficulties developing rich explanations to explain *their findings* (Krajcik et al. 1998; McNeill and Krajcik 2007). When students provide reasoning for their claims, they often use inappropriate data by drawing on data that do not support their claim (McNeill and Krajcik 2011; Kuhn 1991; Schunn and Anderson 1999), make no mention of specific evidence (Chinn et al. 2008), or generally state that an entire data table is evidence (McNeill and Krajcik 2011; Chinn et al. 2008).

In this work, we sought to unpack the difficulties associated with data interpretation and warranting claims in particular.

8.2 Our Solution: Inq-ITS (Inquiry Intelligent Tutoring System; www.inqits.com)

In response to calls such as the Next Generation Science Standards, as well as teachers' assessment challenges and students' learning challenges, we have developed a solution that leverages schools' existing computing resources to help teachers with inquiry assessment by providing *automatic, formative data* and to help students learn these skills by providing *real-time, personalized scaffolds as they engage in inquiry*. Inq-ITS (Inquiry Intelligent Tutoring System) is a lightweight LMS, providing computer-based assessment and tutoring for science inquiry skills. It is a *no-install, entirely browser-based learning and assessment tool* created using evidence-centered design (Mislevy et al. 2012) in which middle school students conduct inquiry using science microworlds (Gobert 2015). Within Inq-ITS, which consists of different interactive simulations within microworlds, or virtual labs, for different domains in physical, life, and earth science, students "show what they know" by forming questions, collecting data, analyzing their data, warranting their claims, and explaining findings using a claim-evidence-reasoning framework, all key inquiry practices (NGSS Lead States 2013). As students work, the inquiry

work products they create and processes they use are automatically assessed using our patented assessment algorithms (Gobert et al. 2016a, b). These assessment algorithms were built and validated using student data (Sao Pedro et al. 2010, 2012a, 2013b, c, 2014; Gobert et al. 2012, 2013, 2015; Moussavi et al. 2015, 2016a). They have been shown to be robust when tested across inquiry activities with diverse groups of students and match human coders with high precision (precision values ranging from 84% to 99%; Sao Pedro et al. 2012a, b, 2013a, b, 2014, 2015).

8.3 Others' Prior Research on Scaffolding Inquiry

Given student difficulties with inquiry as previously described, providing support to students for inquiry is critical if the Next Generation Science Standards (2013) or other policies emphasizing authentic science practices (e.g., OECD 2018) are to be realized. Scaffolds for inquiry can help students achieve success they could not on their own (Kang et al. 2014; McNeill and Krajcik 2011) and can lead to a better understanding of scientific concepts and the purpose of experimentation, as well as the inquiry skills used in experimentation (Kirschner et al. 2006). For example, providing scaffolding for a PhET simulation on circuit construction lead students to be more explicit in their testing (such as adding a voltmeter or connecting an ammeter in the circuit); this systematicity also transferred once scaffolding was removed (Roll et al. 2014). Additionally, the specific skill of collecting controlled trials, a lynchpin skill of inquiry, can be learned via strategy training and transfers to other topics (Klahr and Nigam 2004). Scaffolding can also be used to help students make connections between experimental data and real-world scenarios (Schauble et al. 1995). Lastly, scaffolding students' explanations during inquiry can yield positive effects on learning (Edelson et al. 1995; McNeill et al. 2006). Taken together, these results demonstrate the potential for deeper inquiry learning when students are provided with adequate support.

One drawback, however, to many of these studies is that the scaffolding is either provided by a teacher, is in the form of text-based worksheets, or in some other form that is either not scalable or fine-grained, i.e., operationalized at the sub-skill level. Additionally, these approaches typically require a student to know when they need help; however, students may not have the metacognitive skills needed to do so (Aleven and Koedinger 2000; Aleven et al. 2004).

In our system, by contrast, we use an automated approach that detects students' problems with inquiry and provides computer-based scaffolding in real time in order to support the acquisition and development of inquiry skills/practices (Gobert et al. 2013; Sao Pedro et al. 2013b, c, 2014; Gobert and Sao Pedro 2017). These scaffolds are designed to address specific aspects of scientific inquiry on a fine-grained level and can help students receive the help they need by targeting the exact sub-skill on which they are having difficulty. Our identification of each of the sub-skills underlying each of the science practices described by the NGSS (2013) is described elsewhere (Gobert and Sao Pedro 2017). This approach provides both scalable assessment of science inquiry practices as well scalable guidance so that

students can get help while they are having difficulty. Scaffolding in real time has been shown to better support students' learning in general (Koedinger and Corbett 2006) and in inquiry learning in particular (Gobert et al. 2013; Gobert and Sao Pedro 2017). This approach has a great benefit over the others in that it is scalable so that NGSS practices, as described, can be learned.

8.4 Inq-ITS' Prior Work on Efficacy of Scaffolding

In our work, we have shown that our scaffolding can help students who did not know two skills related to planning and conducting experiments (NGSS Lead States 2013) – testing hypotheses and designing controlled experiments – acquire these skills and transfer them to a new science topic. These findings were robust both within the topic in which students were scaffolded *and* across topics for each domain studied (physical, life, and earth science), with scaffolded students maintaining and/or improving their skills in new topics when scaffolding was removed compared to those who did not receive scaffolding (Sao Pedro et al. 2013a, b, 2014).

With regard to the inquiry practices of interest in this chapter, namely, interpreting data and warranting claims, we recently conducted a systematic analysis of a subset of our data to address whether our scaffolding with Rex is supporting students in the acquisition and transfer of these inquiry skills. Later in the chapter, we provide an additional study, using Bayesian Knowledge Tracing (BKT) (Corbett and Anderson 1995), a computational approach allowing for the analysis of the fine-grained sub-skills underlying our practices of data interpretation and warranting claims.

Our data were drawn from 357 students in six middle school classes in the Northeast of the United States. Students completed two microworlds (Flower and Density) in either the Rex ($N = 156$) or No Rex ($N = 201$) condition. Mixed repeated measures ANOVAs on both interpretation skill and warranting skill were performed. An independent variable of time phase (repeated) was included in order to account for how participants consecutively completed two microworlds: Flower and Density. In the Flower virtual lab, none of the students received scaffolding from Rex, so performance in this virtual lab was used as the baseline. In the Density virtual lab, students were randomly assigned to either the Rex or No Rex condition. The Rex condition meant that Rex was available to assist students as they engaged in the microworld, whereas the No Rex condition meant that Rex was not available and could not be triggered. The results focused on the interactive effects of time × condition. Effect size was calculated using Cohen's d. All significance testing for the primary analyses was conducted with an alpha level of .05. Our main interest was the effect of Rex's scaffolding on learning.

Table 8.1 illustrates the estimated means of interpretation skill and warranting skill in the Rex and No Rex conditions as well as standard errors, lower and upper bound with 95% confidence interval, F values, and the effect size of Cohen's d in the pairwise analyses, respectively.

Table 8.1 Statistics for condition × time in the Flower and Density virtual labs

					95% CI			Cohen's
Skills	Time	Condition	Mean	SE	Lower	Upper	*F*	*d*
Interpreting data	1	No Rex	0.68	0.02	0.65	0.71	0.48	0.074
		No Rex	0.66	0.02	0.63	0.70		
	2	No Rex	0.74	0.02	0.71	0.77	18.11***	0.454
		Rex	0.84	0.02	0.81	0.88		
Warranting claims	1	No Rex	0.37	0.02	0.32	0.41	3.63	0.203
		No Rex	0.30	0.03	0.25	0.35		
	2	No Rex	0.68	0.02	0.64	0.73	7.06**	0.283
		Rex	0.77	0.03	0.72	0.82		

$N = 714$; $df = 1, 710$. Time 1 is Flower and Time 2 is Density
SE standard error, *CI* confidence interval
$^{***}p < .001$. $^{**}p < .01$

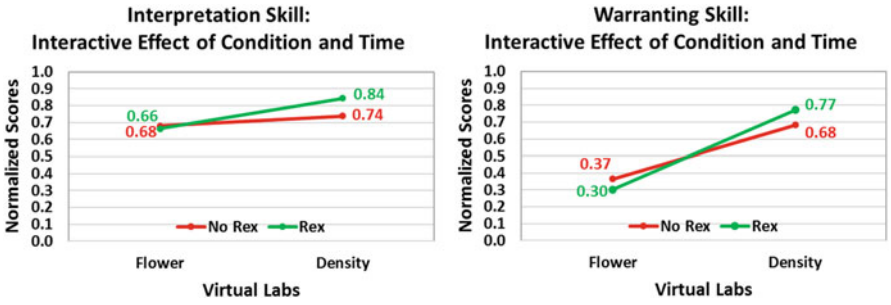

Fig. 8.1 Estimated means of *condition* × *time* in Flower and Density microworlds, respectively

8.4.1 Data Interpretation

There was a significant two-way interaction between condition × time for data interpretation skill, $F(2, 710) = 12.25$, $p < 0.001$ (see Table 8.1 and Fig. 8.1). The pairwise comparisons showed that students' interpretation substantially improved in both the No Rex (mean increased from .68 to .74, $p = .010$, $d = .26$) and Rex conditions (mean increased from .66 to .84, $p < .001$, $d = .79$). This implies that students' interpretation skills improved when they used the virtual lab *even without scaffolding from Rex*. In the second virtual lab, Density, students who received scaffolding from Rex achieved higher scores on interpreting data in the Rex condition than in the No Rex condition with a medium effect size. These findings confirm that students who received Rex's support experienced greater improvement on interpretation skills relative to students who did not receive support from Rex.

8.4.2 Warranting Claims

There was a significant two-way interaction between condition × time for warranting skill, $F(2, 710) = 10.40$, $p = 0.001$ (see Table 8.1 and Fig. 8.1). The pairwise comparisons showed that students' performance on warranting claims substantially improved in both the No Rex (mean increased from .37 to .68, $p < .001$, $d = 1.02$) and Rex conditions (mean increased from .30 to .77, $p < .001$, $d = 1.51$). This implies that students' skills at warranting claims improved when they used the virtual lab with or without scaffolding from Rex. Results also showed that there were no significant differences in students' skills at warranting claims when they conducted the first virtual lab, Flower, without Rex scaffolding. However, in the second virtual lab, Density, students who received scaffolding from Rex achieved higher scores on warranting claims in the Rex condition than in the No Rex condition with a small effect size. These findings further confirm that students who received Rex's support experienced greater improvement on warranting claims skills relative to students who did not receive support from Rex.

8.4.3 Using Advanced Analytical Approaches to Study the Fine-Grained Effects of Scaffolding on Students' Data Interpretation and Warranting Claims

In this study, we hypothesized that an automated scaffolding approach that provides personalized feedback would help students learn data interpretation skills and warranting claims skills. As such, we developed scaffolds within Inq-ITS that react when students have difficulty on these key skills and sub-skills (McNeill and Krajcik 2011; Gotwals and Songer 2009; Kang et al. 2014; Berland and Reiser 2009).

8.4.4 Method

8.4.4.1 Participants

Data were collected from 160 eighth grade students from the same school in the Northeast of the United States using Inq-ITS Density activities. All the students had previously used Inq-ITS, but not with this new scaffolding capacity.

8.4.4.2 Materials

Inq-ITS Density Virtual Lab Activities For each Inq-ITS virtual lab, there are typically three or four inquiry activities, consisting of driving questions that help

guide students through the inquiry phases. Within each activity, students conduct inquiry by first articulating a testable hypothesis using a hypothesis widget with pulldown menus. They then experiment by collecting data with an interactive simulation through the manipulation of variables (Fig. 8.2a). Once they have collected all of their data, they interpret the results of their experiment by forming a claim in claim widget (similar to that used for hypothesizing) and selecting trials as evidence (Fig. 8.2b). Finally, students write a short open-response report that summarizes their findings from their inquiry using a claim-evidence-reasoning format (McNeill and Krajcik 2011).

In this study, three Density virtual lab activities were used. These activities aim to foster understanding about the density of different liquid substances (water, oil, and alcohol). In the first activity, the goal was to determine if the shape of the container affected the density of the liquid; the second was to determine if the amount of liquid affected the density; and the third was to determine if the type of liquid affected the density.

8.4.4.3 Procedure

Students worked on the Density activities in a computer lab at their school for the length of one science class (approximately 50 min). 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. Students were randomly assigned to one of two conditions: 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 sub-skills. 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 ensured that all students, regardless of scaffolding condition, had both a testable hypothesis and relevant, controlled data with which they could correctly undergo data interpretation (this design also allows us to isolate and systematically study the effects of scaffolding for data interpretation skills, as opposed to the two that proceed it in the inquiry process). Thus, students in both conditions worked in the same environment and on the same activities with access to hypothesizing and data collection scaffolding. The only difference was the presence of data interpretation scaffolding for one condition (Interpretation Scaffolding condition).

Evaluation of Inquiry Sub-skills For data interpretation and warranting claims, there are eight main sub-skills that are evaluated in the system using the work products students create. These work products are their claim (selecting the appropriate variables and relationship between them) and supporting evidence (selecting

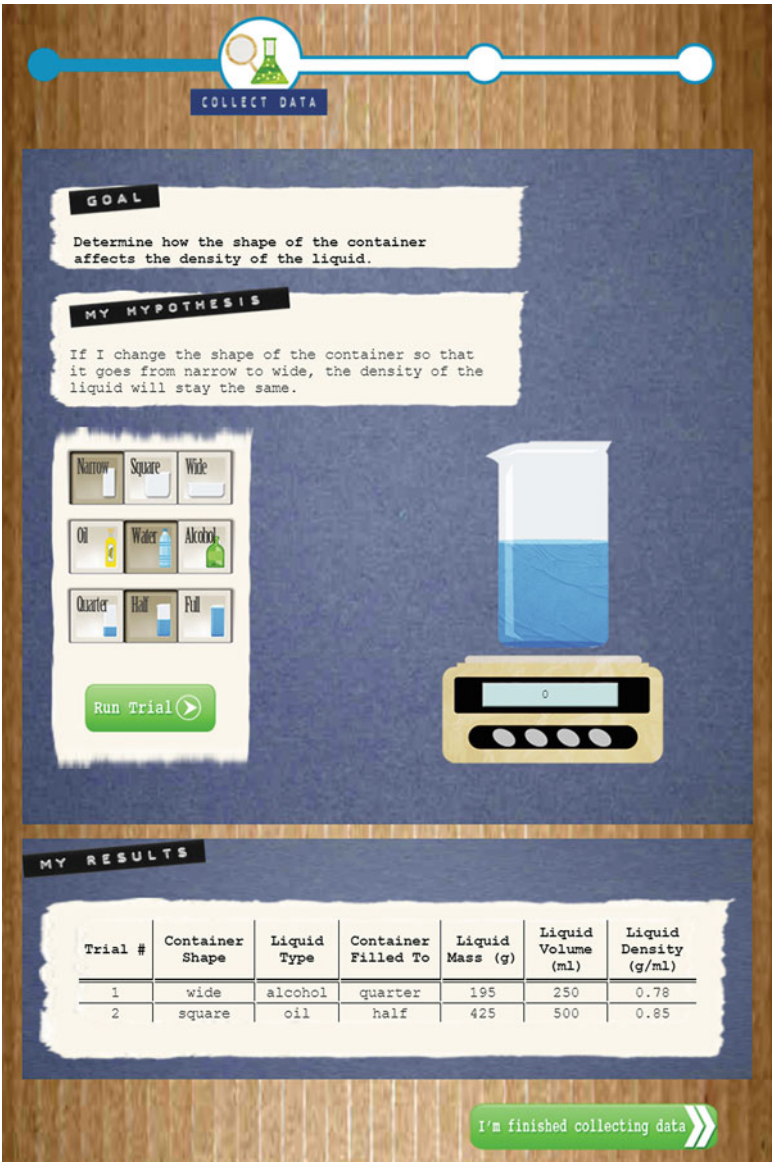

Fig. 8.2a In the "collect data" phase of the Inq-ITS Density virtual lab, students collect to test their hypothesis

relevant, controlled trials from their data table that reflect the relationship stated in their claim). These sub-skills and the specific criteria with which they are evaluated can be seen in Table 8.1. Since these sub-skills, defined in the context of this activity, are well-defined (Gobert and Sao Pedro 2017), they are evaluated using

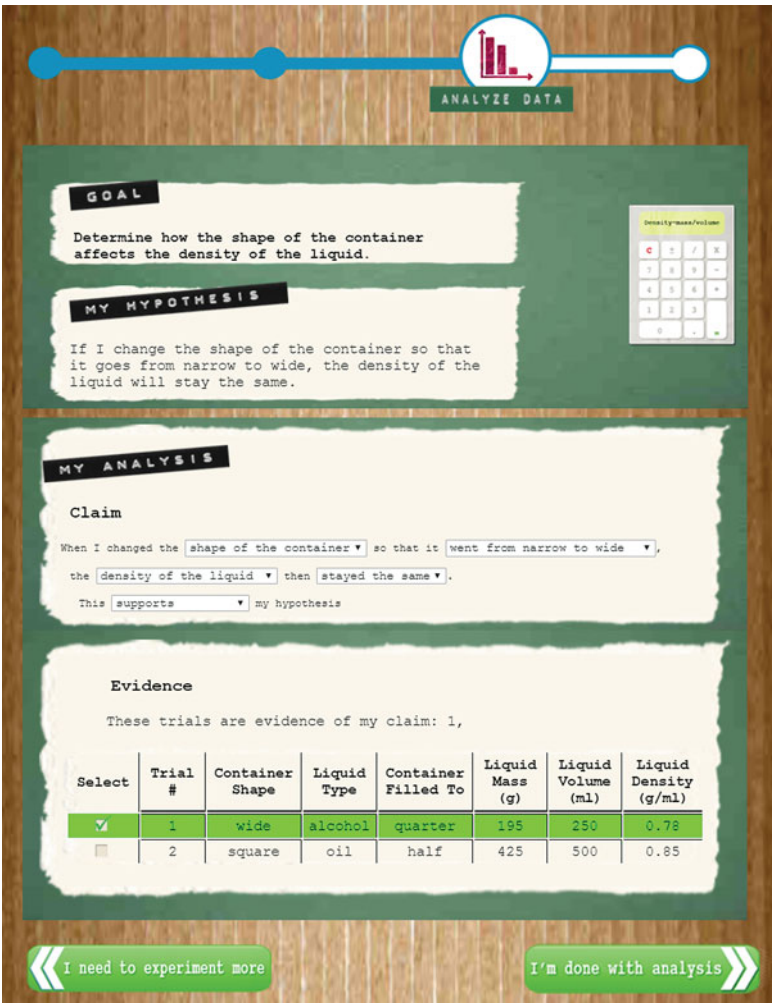

Fig. 8.2b After collecting data, students analyze their data. They review the data they collected, use pulldown menus to describe the trends found in their data, and select the evidence (trials) to support their claim

knowledge-engineered rules that specify if the sub-skill has been demonstrated. For example, for the sub-skill "Claim DV" shown in Table 8.2, the system evaluates whether or not the student has correctly chosen a variable that is measured, not changeable, within the simulation (a dependent variable) in the appropriate part of the claim. Within the context of the Density virtual lab, the appropriate dependent variable is "density of the liquid." So if the student states "density of the liquid" as the dependent variable, they would be marked as correctly demonstrating the DV sub-skill. However, if the student chooses another variable, such as one of

Table 8.2 Data interpretation sub-skills

Data interpretation sub-skills	Criteria
Interpreting the IV/DV relationship	Is the IV DV relationship correct?
Claim IV	Did the student correctly select an IV when making a claim?
Claim DV	Did the student correctly select a DV when making a claim?
Interpreting the hypothesis/claim relationship	Is the choice of whether the claim supports (or refutes) the hypothesis correct?
Controlled trials	Are all the selected trials controlled?
Warranting the IV/DV relationship	Do the selected trials support the stated IV/DV relationship?
Evidence	Did the student select more than one trial as evidence?
Warranting the hypothesis/claim relationship	Do the selected trials support the student's statement on whether their interpretation supports their hypothesis?

the independent variables like "type of liquid," as the dependent variable, then they would be scored as incorrectly demonstrating the DV sub-skill. As another example, for the sub-skill "interpreting the IV/DV relationship," a rule checks that the relationship between the independent and dependent variables specified in the claim is reflected in the data collected by the student. Elaborating further, if a student claims that "When I increased the size of the container the density of the liquid stayed the same" and their data reflects that relationship, that sub-skill would be scored as correct. If the data they collected did not reflect that relationship, the sub-skill would be scored as incorrect. The evaluation rules yield binary measure of correctness on each sub-skill (i.e., the results are presented as being correct or incorrect rather than having levels of correctness). This allows us to tease apart separate components (the sub-skills) within the broader skill of analyzing data.

Scaffolds in Inq-ITS Inq-ITS delivers scaffolds to students in text format via a pedagogical agent named Rex, a cartoon dinosaur (Fig. 8.3). Scaffolding is triggered automatically when a student completes their data analysis and at least one of the sub-skills is incorrectly demonstrated (evaluated by the knowledge-engineered rules discussed previously). This proactive scaffolding approach helps to support students in their inquiry processes (Schauble 1990; deJong 2006) by preventing students from engaging in ineffective behaviors (Buckley et al. 2006; Sao Pedro 2013). This proactive approach is also important because students may not be aware that they need help (Aleven and Koedinger 2000; Aleven et al. 2004). Once scaffolding is triggered, students may also ask Rex for additional clarification and support.

The scaffolds are designed to adapt to students' skill level by both providing multiple levels of automatic scaffolds and allowing students to request for further help or clarification (once support is auto-triggered), as needed. In this way, the

Fig. 8.3 Example scaffold delivered by Rex during data interpretation

scaffolds personalize each student's learning, recognizing that different students may need different amounts of help to successfully hone different sub-skills. The data interpretation scaffolds address four categories of procedurally-oriented difficulties that focus on the eight aforementioned sub-skills evaluated within data interpretation and warranting claims (Moussavi et al. 2015). These data interpretation and warranting claims scaffold categories are:

1. The Claim IV/DV does not match the hypothesis IV/DV.
2. The trials selected for warranting are not properly controlled or relevant to the claim.
3. The claim does not reflect the data selected.
4. The claim is incorrect as to whether it supports/does not support the hypothesis.

Since students may require scaffolding support for none, one, or many of these sub-skills, the scaffolds are designed to address these in the order listed above, so

that each step of data interpretation is completed before moving onto the next. 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 until the sub-skill is demonstrated correctly before another difficulty is addressed. On the other hand, if a student also 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.

When students make multiple errors within the same category, we follow a sequence that increases the level of feedback given to the student. For the first error, a scaffold is provided to orient students to the current task. If the same error is repeated, they are then guided through the necessary procedural skills. Finally, the system provides a "bottom-out" hint telling students the procedure to follow. In this way, the student receives more and more targeted support, similar to cognitive tutors (e.g., Anderson et al. 1995; Corbett and Anderson 1995; Koedinger and Corbett 2006).

In sum, these scaffolds are designed to adapt to students' skill level by both providing multiple levels of automatic scaffolds and allowing students to request for further help or clarification (once support is auto-triggered), as needed. In this way, the scaffolds personalize each student's learning, recognizing that different students may need different amounts of help to successfully hone different sub-skills.

Data Analysis Approaches Due to the complexities and sub-skills inherent in the inquiry practices of data interpretation and warranting claims, an advanced analytical method using an extension of Bayesian Knowledge Tracing (Corbett and Anderson 1995) is better suited to address the effects of scaffolding on students' learning and transfer of sub-skills of inquiry under investigation here (Sao Pedro et al. 2013b). Bayesian Knowledge Tracing (BKT henceforth), a cognitive modeling approach to approximating the mastery of sub-skills in intelligent tutoring systems, is a powerful technique, and its prediction of student performance is as good as or better than similar algorithms that aggregate performance over time in order to infer student skill (e.g., Baker et al. 2011). Additionally, our group has shown that this approach is effective for modeling students' learning of inquiry, both with and without the presence of scaffolding (Sao Pedro et al. 2013b).

8.4.4.4 Bayesian Knowledge Tracing

Bayesian Knowledge Tracing (BKT) (Corbett and Anderson 1995) estimates the likelihood that a student knows a particular skill (or sub-skill) and disentangles between "knowing" and "demonstrating" that skill (or sub-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 the skill or they do not) and that skill demonstration is also binary (either a student demonstrates a skill or they do not).

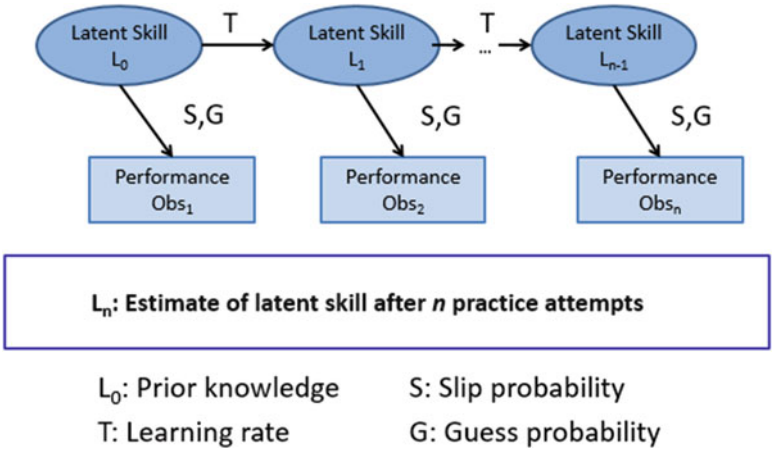

Fig. 8.4 Bayesian Knowledge Tracing model

Mathematically, four parameters are used to model whether a student knows a skill: L_0, T, G, and S (Corbett and Anderson 1995). L_0 is the probability of initial knowledge that the student is already in the "learned state," i.e., before they start the first problem. T is the probability of learning, i.e., the chance that the student goes from the "unlearned state" to the "learned state" over the course of doing all of the problems in the sequence. G is the probability of guessing, i.e., the chance that a student in the "unlearned state" answers the problem correctly. Lastly, S is the probability of slipping, i.e., the chance that a student in the "learned state" answers the problem incorrectly (Corbett and 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. BKT, in this formulation, assumes that skills are not forgotten (Corbett and 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, 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. See Fig. 8.4.

Prior work by Sao Pedro (2013) extended the traditional BKT model to account for the presence of a tutor intervention, similar to that of Beck et al. (2008). To incorporate scaffolding into the BKT framework, they introduced the dichotomous observable variable of Scaffolded = {True, False} and conditioned the learning rate (T) on that observable leading to two distinct learning rate parameters – $T_{\text{scaffolded}}$ and $T_{\text{unscaffolded}}$. This resulted in the following equations for computing $P(L_n)$, the likelihood of knowing a skill (Sao Pedro 2013):

$$P\left(L_n|\text{Scaffolded}_n = \text{True}\right) = P\left(L_{n-1}|\text{Prac}_n\right) + \left(1 - P\left(L_{n-1}|\text{Prac}_n\right)\right)_* P\left(T_{\text{scaff}}\right)$$

$$P\left(L_n|\text{Scaffolded}_n = \text{False}\right) = P\left(L_{n-1}|\text{Prac}_n\right) + \left(1 - P\left(L_{n-1}|\text{Prac}_n\right)\right)_* P\left(T_{\text{unscaff}}\right)$$

We follow this approach to determine whether data interpretation scaffolds are supporting students' learning.

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 our data to the BKT model, specifically with regard to how scaffolding condition was defined and practice opportunities were defined. These considerations are discussed in the following section.

8.4.4.5 Data Preparation Extensions to Leverage the BKT Framework

The data logged here differs from typical data logs due to how the data interpretation scaffolds were integrated into the system. In the system, all of the data interpretation sub-skills are designed to be evaluated at once. However, the data interpretation scaffolds are designed to only address one sub-skill 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 sub-skill, 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 more evaluated on all of the data interpretation sub-skills, regardless of what specific aspects of their analysis they changed.

Considering this and the fact that in BKT analysis every sub-skill is considered separately and has its own model, it became important to consider how the BKT framework defined the scaffolding condition and practice opportunity in order to create an accurate model. These design decisions for the BKT model are described in more detail below.

8.4.4.6 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 sub-skills. 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. Arguably, it is more important to model the scaffolds students received on a per skill basis, rather than simply considering them as scaffolded or not. Because of this, scaffolding was considered at the sub-skill level so that any scaffolds a student received for one specific sub-skill had no bearing on the student's scaffolding classification for the other sub-skills. This

means that in the BKT model for the Claim DV sub-skill, 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 sub-skill, regardless of any other scaffold they may or may not have received. This makes it so that a student may only be in the scaffolding condition in the BKT model for one sub-skill or may be in the scaffolding condition in multiple BKT models on different sub-skills.

8.4.4.7 Determining Number of Practice Opportunities

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 sub-skills (Gobert et al. 2013). When scaffolding is being used, students who have been evaluated as "incorrectly demonstrating any sub-skill" 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 sub-skills regardless of what specific sub-skill(s) were worked on, which can make it seem as though students require more practice opportunities to master a sub-skill than they actually do. For example, as shown in Table 8.3, based on the evaluations, it looks like after three practice opportunities, the student is still incorrectly demonstrating the "claim" and "support" sub-skills. However, if we look at the student's actions, we can see that the student was only focused on correctly demonstrating the "DV" sub-skill (due to the scaffolding received) and was not actually working on the other two sub-skills. Therefore, it would not be accurate to

Table 8.3 Example of practice opportunity succession

Student presses submit		
Sub-skills	**Evaluation**	**Practice opportunity**
IV	1	1
DV	0	1
Claim	0	1
Supports	0	1
Student receives scaffolding for DV, only changes DV (still incorrect), and submits		
Sub-skills	**Evaluation**	**Practice opportunity**
IV	1	2
DV	0	2
Claim	0	2
Supports	0	2
Student receives scaffolding for DV, only changes DV (correctly this time), and submits		
Sub-skills	**Evaluation**	**Practice opportunity**
IV	1	3
DV	1	3
Claim	0	3
Supports	0	3

Table 8.4 Example of collapsed evaluation

Sub-skills	Evaluation
IV	1
DV	0
Claim	0
Supports	0

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

The option considered here was to collapse student evaluations for each sub-skill 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 model with fewer parameters. In collapsing students' evaluations, all of the evaluations for one sub-skill within an activity were examined, and a student would receive a correct evaluation for a particular sub-skill only if they always had correct evaluations for that sub-skill. This was done because if a student ever incorrectly demonstrated a sub-skill, it could be assumed that the student most likely did not know the sub-skill to begin with. This resulted in the student's evaluations in the above figure to be collapsed into one practice opportunity as shown in Table 8.4.

Therefore, the BKT analysis was performed for each of the assessed data interpretation and warranting claims sub-skills, using the scaffolding extension of the BKT framework developed by Sao Pedro (2013), as previously described.

8.4.4.8 Fitting BKT Model Parameters

To learn the parameters (L_0, T_{Scaff}, T_{Unscaff}, G, S) from student data for each of the BKT models (one model per targeted data interpretation sub-skill), we used a brute force grid search approach (Baker et al. 2010) to find the parameters that minimize the lowest sum of squared residuals (SSR) between the probability of demonstrating a skill and the actual data, as done in Sao Pedro et al. (2013b).

8.4.4.9 Determining Goodness of the BKT Models

Once the BKT parameters were determined, they were applied to the model, and then its predictive performance was tested against the same set of data used to construct the model. Although cross-validation helps to ensure that the models are accurate and can be applied to new students, it requires a held-out validation data set collected from a similar population. Since this work is exploratory in nature in that it is examining the first set of data collected with the data interpretation scaffolds, we did not have a held-out data set that could be used for this purpose. As such, the same set of data used for training was also used for validation, which can lead

to over-fitting of the model. In ongoing work, we are addressing this limitation by using a held-out test set to test the models.

As in Sao Pedro et al. (2013b), performance was measured using A′ (Hanley and 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.

8.5 Results

Our goal is to determine whether our automated scaffolding approach helps students acquire data interpretation sub-skills. We first look at a descriptive analysis of the frequency with which scaffolds were used across the activities. We also look at error rates for the sub-skills to get an initial look at students' progress with and without scaffolding. Then, as mentioned, we used our BKT extensions to approximate student learning of the data interpretation sub-skills and to make inferences about whether scaffolding was effective.

Descriptive Analysis Table 8.5 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 sub-skills 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.

We next looked at the error rates for four of the data interpretation sub-skills most tightly related to the evaluations that trigger the scaffolds. Error rate is defined as the percentage of students who demonstrated that error in each activity. The graphs in Fig. 8.5 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.

Table 8.5 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

Activity 1 is not presented, because scaffolding was not available in that activity

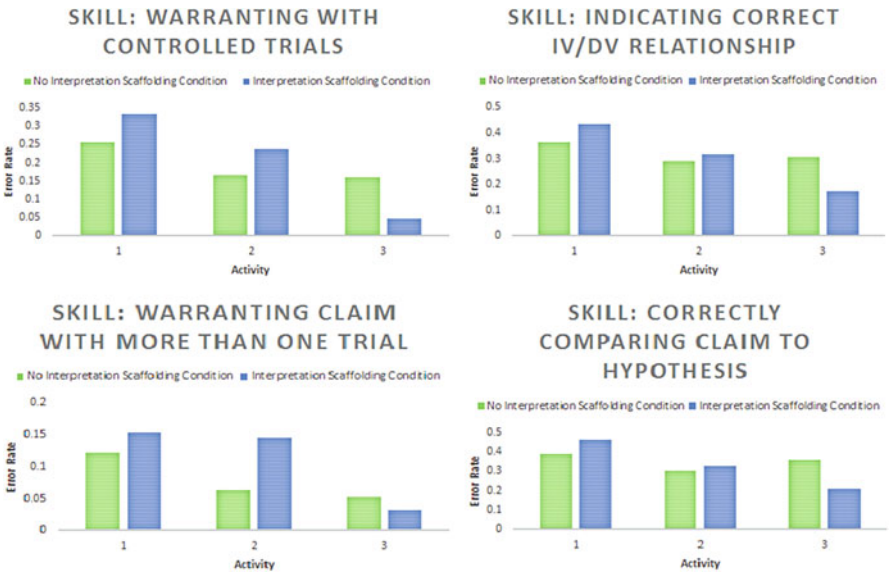

Fig. 8.5 Error rate analysis

As shown in these graphs (Fig. 8.5), student difficulty/error was present in each of these sub-skills, with the sub-skill "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 start with a higher error rate but end with a lower error rate. For example, for the sub-skill "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.

The descriptive analyses suggest that scaffolding appears to be effective at helping students acquire these sub-skills. We next conduct a deeper inferential analysis using the BKT modeling framework described previously.

Inferential Analysis with Bayesian Knowledge Tracing As described previously, we fit BKT models using the student data collected and use A′ (Hanley and McNeil 1982) to measure the goodness of the models. Recall that an A′ of 0.5 is indicative of chance performance and an A′ of 1.0 is indicative of perfect performance. The A′ values for this analysis can be seen in Table 8.6. In this case, performance was measured to be relatively high for all of the sub-skills with A′ values between

Table 8.6 A′ values showing high performance of the BKT models

Sub-skill	A′
Interpreting the IV/DV relationship	0.73
Claim IV	0.70
Claim DV	0.69
Interpreting the hypothesis/claim relationship	0.72
Controlled trials	0.79
Warranting the IV/DV relationship	0.73
Evidence	0.81
Warranting the hypothesis/claim relationship	0.72

Table 8.7 BKT parameters for each sub-skill

				No Interpretation Scaffolding condition	Interpretation Scaffolding condition
Sub-skill	Probability of initial knowledge	Probability of guessing	Probability of slipping	Probability of learning	
Claim DV	0.72	0.30	0.04	0.69	0.71
Claim IV	0.94	0.21	0.01	0.83	0.36
Interpreting the IV/DV relationship	0.61	0.13	0.09	0.24	0.62
Interpreting the hypothesis/claim relationship	0.59	0.14	0.10	0.20	0.55
Controlled trials	0.71	0.12	0.04	0.27	0.79
Warranting the IV/DV relationship	0.62	0.10	0.10	0.22	0.64
Evidence	0.81	0.22	0.00	0.22	0.84
Warranting the hypothesis/claim relationship	0.59	0.13	0.10	0.20	0.53

0.69 and 0.81, allowing for parameter interpretation. However, again, since cross-validation was not done, it is possible that some of these models may be over-fitting to some student data (c.f. Sao Pedro et al. 2013).

The results from the BKT analysis indicate that the data interpretation scaffolds were effective in supporting the acquisition of data interpretation sub-skills. This can be seen through the values of the probability of learning. This value 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 in Table 8.7, the probability of

learning for students receiving data interpretation scaffolding is higher for all but one of the evaluated sub-skills. This sub-skill, selecting an IV for the claim, also has a high probability of initial knowledge, which could indicate that the sub-skill is not being learned because so many students already know it (e.g., Sao Pedro et al. 2014). Also, compared to another sub-skill with a relatively high probability of initial knowledge – such as the Evidence sub-skill – the Claim IV sub-skill is noisier to assess, likely because it might be highly related to the content in each activity.

8.6 Discussion

The goal of this work was to test the efficacy of our data interpretation scaffolding on the sub-skills underlying the skill practices underlying data interpretation and warranting claims. We tested this in two ways, both using analysis of variance on the aggregate scores for each practice (data interpretation and warranting claims), as well as an innovative extension to Bayesian Knowledge Tracing (BKT) that considers the presence of scaffolding approximating mastery learning for each of the sub-skills of interest (Sao Pedro et al. 2013b). We also developed modifications to this framework, which allow it to be applied when condition and practice opportunity can be defined on different levels (i.e., activity level vs. skill level).

In developing our BKT extension, this work contributes a fine-grained method for unpacking the effect of scaffolding via logged, process data. Our extension to BKT was used as a modeling paradigm to track the sub-skills underlying data interpretation and warranting claims. This study was done within a complex domain of science inquiry whereby the student data, number of practice opportunity counts, and evaluated skills were not as clearly delineated as in previous studies in which BKT was used to evaluate educational interventions (Koedinger et al. 2010). This work provides a framework for how data in these complex environments can be treated before BKT can be used.

This work also explores modifying the BKT framework to represent and track students' learning of the targeted data interpretation sub-skills with and without scaffolding. Further analyses are needed to determine the efficacy of this model and its accuracy in comparison to other models. As the data used for this work was collected as an initial study of the data interpretation/warranting claims scaffolds, additional data will be used to cross-validate the predictive performance of the models used here and provide greater assurance in interpreting the parameters of the model. This method could then be used as students work through multiple domains with scaffolding to assess the efficacy of these scaffolds across a larger number of practice opportunities (e.g., Sao Pedro et al. 2014). This will also allow us to assess how scaffolding can impact the transfer of these skills from one science domain to another. Additionally, we will use this method on studies without scaffolding, which will give us data to better understand how this skill develops naturally.

Regarding inquiry, this work builds on prior research (Kang et al. 2014; McNeill and Krajcik 2011; Schauble 1990) on the nature of data interpretation and warranting claims skills, their assessment, and scaffolding. This work makes a contribution to the prior research on argumentation practices for inquiry by conceptualizing and framing the data interpretation and warranting claims practices as *necessary but not sufficient for* appropriate scientific argumentation.

When it comes to unpacking the broad components of explanation, Toulmin's (1958) model of argumentation is typically used (McNeill and Krajcik 2011; Gotwals and Songer 2009; Kang et al. 2014; Berland and Reiser 2009), breaking down argumentation into three main components: the use of claims, evidence, and reasoning. The interpretation of evidence and the creation of an evidence-based explanation or argument are both key practice in national science standards and essential for fostering students' science literacy (McNeill and Krajcik 2011; Kang et al. 2014).

We feel that unpacking the inquiry practices associated with data interpretation and warranting claims *separately* from students' data on claims, evidence, and reasoning, as expressed in open response format, is important because if students are having problems analyzing their data, they won't be able to successfully engage in explanation and argumentation. Our prior work has shown that a number of students are not able to articulate a correct explanation or argument despite knowing the data interpretation skills (Li et al. 2017). Moreover, there are large numbers of students who are being mis-assessed when their open responses are used as the only source of assessment: there are students who are skilled at science but cannot convey what they know in words (i.e., false negatives), as well as students who are skilled at parroting that they have read or heard but do not understand the science they are writing about (i.e., false positives; Gobert 2016). In short, using solely students' writing for assessment is only an accurate way of measuring what students know *if* they are good at articulating words.

To this end, we conceptualize/frame data interpretation and warranting claims practices as underlying the argumentation practices necessary for communicating science findings and thus find it necessary to study these skills separately from students' overall written explanations and arguments. Conceptualizing and supporting students on the components of the explanation framework – claim, evidence, and reasoning – in an automated and fine-grained way with appropriate sub-skills can help us unpack and target known difficulties documented by previous research (Gotwals and Songer 2009; McNeill and Krajcik 2011; Schunn and Anderson 1999). While we could make the assessment of these skills easier by designing activities that only target one skill at a time, this would be a much less authentic way of conducting inquiry. This work attempts to disentangle the effects of learning support delivered via automatic scaffolds that apply to individual sub-skills in an environment where multiple performance-based skills are being practiced and assessed at once. This 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 before moving on to students' claims, evidence, and reasoning.

Lastly, this work provides a scalable solution toward the assessment and scaffolding of these practices and in doing so represents a scalable solution to supporting teachers and students in NGSS practices.

References

Aleven, V., & Koedinger, K. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Proceedings of the 5th international conference on intelligent tutoring systems* (pp. 292–303). Berlin: Springer.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In J. C. Lester, R. M. Vicario, & F. Paraguaçu (Eds.), *Proceedings of seventh international conference on intelligent tutoring systems* (pp. 227–239). Berlin: Springer.

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

Baker, R., Corbett, A., Gowda, S., Wagner, A., MacLaren, B., Kauffman, L., Mitchell, A., & Giguere, S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. In *Proceedings of the 18th annual conference on user modeling, adaptation, and personalization* (pp. 52–63). Berlin: Springer.

Baker, R., Gowda, S., & Corbett, A. (2011) Automatically detecting a student's preparation for future learning: Help use is key. In *Proceedings of the 4th international conference on educational data mining* (pp. 179–188).

Beck, J., Chang, K. M., Mostow, J., & Corbett, A. (2008). Does help help? Introducing the bayesian evaluation and assessment methodology. In *Intelligent tutoring systems* (pp. 383–394). Berlin: Springer.

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

Buckley, B. C., Gobert, J. D., & Horwitz, P. (2006). *Using log files to track students' model-based inquiry.* Paper presented at the 7th international conference of the learning sciences, Bloomington, IN.

Chinn, C. A., & Brewer, W. F. (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. A., Duschl, R. A., Duncan, R. G., Buckland, L. A., & Pluta, W. J. (2008, June). A microgenetic classroom study of learning to reason scientifically through modeling and argumentation. In *Proceedings of the 8th international conference on International conference for the learning sciences* (Vol. *3,* pp. 14–15). International Society of the Learning Sciences.

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

de Jong, T. (2006). Computer simulations: Technological advances in inquiry learning. *Science, 312*, 532–533.

Deters, K. M. (2005). Student opinions regarding inquiry-based labs. *Journal of Chemical Education, 82*(8), 1178–1180.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science: A Multidisciplinary Journal, 17*(3), 397–434.

Edelson, D. C., O'Neill, D. K., Gomez, L. M., & D'Amico, L. (1995). A design for effective support of inquiry and collaboration. In *The first international conference on computer support for collaborative learning* (pp. 107–111). Mahwah: Erlbaum.

Fadel, C., Honey, M., & Pasnick, S. (2007). Assessment in the age of innovation. *Education Week, 26*(38), 34–40.

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific reasoning across different domains. In E. DeCorte, M. Linn, H. Mandl, & L. Verschaffel (Eds.), *Computer-based learning environments and problem-solving* (pp. 345–371). Heidelberg: Springer.

Gobert, J. (2015). Microworlds. In R. Gunstone (Ed.), *Encyclopedia of science education* (pp. 638–639). Netherlands: Springer.

Gobert, J. D. (2016). *Op-Ed: Educational data mining can be leveraged to improve assessment of science skills*. US News & World Report. http://www.usnews.com/news/articles/2016-05-13/op-ed-educational-data-mining-can-enhance-science-education.

Gobert, J. D., & Sao Pedro, M. A. (2017). Inq-ITS: Design decisions used for an inquiry intelligent system that both assesses and scaffolds students as they learn. Invited chapter in A. A. Rupp, & J. Leighton (Co-Eds), *Handbook of cognition and assessment*. New York: Wiley/Blackwell.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., 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. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *The Journal of the Learning Sciences, 22*(4), 521–563.

Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity, 18*, 81–90.

Gobert, J. D., Baker, R. S., & Sao Pedro, M. A. (2016a). *U.S. patent no. 9,373,082*. Washington, DC: U.S. Patent and Trademark Office.

Gobert, J., Sao Pedro, M., Betts, C., & Baker, R. S. (2016b). *U.S. patent no. 9,564,057*. Washington, DC: U.S. Patent and Trademark Office.

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.

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. M., & Schunn, C. D. (2004). The transfer of logically general scientific reasoning skills. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 541–546). Mahwah: Erlbaum.

Hilton, M., & Honey, M. A. (Eds.). (2011). *Learning science through computer games and simulations*. Washington, DC: National Academies Press.

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.

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.

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*, 1–48.

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

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211–228.

Koedinger, K., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). New York: Cambridge University Press.

Koedinger, K., Pavlik Jr, P. I., Stamper, J., Nixon, T., & Ritter, S. (2010). Avoiding problem selection thrashing with conjunctive knowledge tracing. In *Educational data mining 2011*.

Krajcik, J., Blumenfeld, P., Marx, R., Bass, K., Fredricks, J., & Soloway, E. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. *The Journal of the Learning Sciences, 7*, 313–350.

Krajcik, J., Marx, R., Blumenfeld, P., Soloway, E., & Fishman, B. (2000). *Inquiry based science supported by technology: Achievement among urban middle school students*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

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

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

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

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S., Klahr, D., & Carver, S. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development, 60*(4), 1–157.

Li, H., Gobert, J., & Dickler, R. (2017). Dusting off the messy middle: Assessing students' inquiry skills through doing and writing. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), *Lecture Notes in Computer Science* (Vol. 10331, pp. 175–187). Cham: Springer.

McElhaney, K., & Linn, M. (2008). Impacts of students' experimentation using a dynamic visualization on their understanding of motion. In *Proceedings of the 8th international conference of the learning sciences* (pp. 51–58). Netherlands: International Society of the Learning Sciences.

McElhaney, K., & Linn, M. (2010). Helping students make controlled experiments more informative. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th international conference of the learning sciences* (pp. 786–793). Chicago: International Society of the Learning Sciences.

McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 233–265). New York: Taylor & Francis Group, LLC.

McNeill, K. L., & Krajcik, J. S. (2011). *Supporting grade 5–8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. Upper Saddle River: Pearson.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences, 15*(2), 153–191.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining, 4*(1), 11–48.

Moussavi, R., Kennedy, M., Sao Pedro, M. A., & Gobert, J. D. (2015). *Evaluating a scaffolding design to support students' data interpretation skills within a simulation-based inquiry environment*. Presented at the meeting of the American Education Research Association, Chicago.

Moussavi, R., Sao Pedro, M., & Gobert, J. D. (2016a). *Evaluating the efficacy of real-time scaffolding for data interpretation skills.* Paper presented at the meeting of the American Education Research Association, Washington, DC.

Moussavi, R., Sao Pedro, M., & Gobert, J. D. (2016b). *The effect of scaffolding on the immediate transfer of students' data interpretation skills within science topics.* Presented at the 12th International Conference of the Learning Sciences, Singapore.

National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. Washington, D.C.: National Academies Press.

Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Njoo, M., & de Jong, T. (1993). Exploratory learning with a computer simulations for control theory: Learning processes and instructional support. *Journal of Research in Science Teaching, 30*, 821–844.

Organization for Economic Cooperation and Development. (2018). *PISA 2015 results in focus: What 15-year-olds know and what they can do with what they know*. Paris: Organization for Economic Cooperation and Development.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Quinn, J., & Alessi, S. (1994). The effects of simulation complexity and hypothesis-generation strategy on learning. *Journal of Research on Computing in Education, 27*(1), 75–91.

Reimann, P. (1991). Detecting functional relations in a computerized discovery environment. *Learning and Instruction, 1*(1), 45–65.

Roll, I., Yee, N., & Briseno, A. (2014). Students' adaptation and transfer of strategies across levels of scaffolding in an exploratory environment. In *Intelligent tutoring systems* (pp. 348–353). Switzerland: Springer International Publishing.

Sao Pedro, M. (2013). *Real-time assessment, prediction, and scaffolding of middle school students' data collection skills within physical science simulations* (Doctoral dissertation). Worcester: Worcester Polytechnic Institute.

Sao Pedro, M. A., Baker, R. S., Montalvo, O., Nakama, A., & Gobert, J. D. (2010). Using text replay tagging to produce detectors of systematic experimentation behavior patterns. In R. Baker, A. Merceron, & P. Pavlik (Eds.), *Proceedings of the 3rd international conference on educational data mining* (pp. 181–190).

Sao Pedro, M., Baker, R., & Gobert, J. (2012a). Improving construct validity yields better models of systematic inquiry, even with less information. In *Proceedings of the 20th conference on user modeling, adaptation, and personalization* (pp. 249–260). Berlin: Springer.

Sao Pedro, M., Gobert, J., & Baker, R. (2012b). *Assessing the learning and transfer of data collection inquiry skills using educational data mining on students' log files.* Paper presented at The Annual Meeting of the American Educational Research Association, Vancouver.

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

Sao Pedro, M., Baker, R., & Gobert, J. (2013b). What different kinds of stratification can reveal about the generalizability of data-mined skill assessment models. In *Proceedings of the 3rd conference on learning analytics and knowledge.*

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (2013c). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction, 23*(1), 1–39.

Sao Pedro, M. A., Gobert, J. D., & Betts, C. G. (2014). Towards scalable assessment of performance-based skills: Generalizing a detector of systematic science inquiry to a simulation with a complex structure. In *Intelligent tutoring systems* (pp. 591–600). Switzerland: Springer International Publishing.

Sao Pedro, M., Gobert, J., Toto, E., & Paquette, L. (2015). Assessing transfer of students' data analysis skills across physical science simulations. In I. Bejar (Chair), *The state of the art in automated scoring of science inquiry tasks.* Symposium conducted at the meeting of the American Education Research Association, Chicago.

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

Schauble, L., 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.

Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *The Journal of the Learning Sciences, 4*(2), 131–166.

Schneider, R., Krajcik, J., & Blumenfeld, P. (2005). Enacting reform-based science materials: The range of teacher enactments in reform classrooms. *Journal of Research in Science Teaching, 42*(3), 283–312.

Schunn, C. D., & Anderson, J. R. (1998). Scientific discovery. In J. R. Anderson (Ed.), *The atomic components of thought* (pp. 385–428). Mahwah: Lawrence Erlbaum Associates.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science, 23*(3), 337–370.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
Shute, V., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 55–71.
Staer, H., Goodrum, D., & Hackling, M. (1998). High school laboratory work in Western Australia: Openness to inquiry. *Research in Science Education, 28*(2), 219–228.
Toulmin, S. (1958). *The uses of argument*. New York: Cambridge University Press.
Tsirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51*(1), 1–10.
van Joolingen, W. R., & de Jong, T. (1991a). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science, 20*(5), 389–404.
van Joolingen, W. R., & de Jong, T. (1991b). Characteristics of simulations for instructional settings. *Education and Computing, 6*(3-4), 241–262.
van Joolingen, W. R., & de Jong, T. (1993). Exploring a domain through a computer simulation: Traversing variable and relation space with the help of a hypothesis scratchpad. In D. Towne, T. de Jong, & H. Spada (Eds.), *Simulation-based experiential learning* (pp. 191–206). Berlin: Springer.
van Joolingen, W. R., & de Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science, 25*(5), 307–346.
Williamson, D., Mislevy, R., & Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah: Lawrence Erlbaum Associates.

Janice D. Gobert, Ph.D., is a Professor of Learning Sciences and Educational Psychology at Rutgers Graduate School of Education. She is also a Co-founder and the CEO of Apprendis, whose main products are Inq-ITS and Inq-Blotter.

Raha Moussavi is a Ph.D. candidate in Learning Sciences & Technology at Worcester Polytechnic Institute. She is completing Ph.D. with Dr. Janice Gobert. She is presently employed in the Teaching Systems Lab, Massachusetts Institute of Technology.

Haiying Li, Ph.D., is a Postdoc with Dr. Janice Gobert in the Graduate School of Education at Rutgers University.

Michael Sao Pedro, Ph.D., is CTO and Co-founder of Apprendis and was a former graduate student of Dr. Janice Gobert at Worcester Polytechnic Institute. He is Co-inventor with Dr. Janice Gobert on the two patents for Inq-ITS and Inq-ITS.

Rachel Dickler is a Ph.D. student with Dr. Janice Gobert in the Graduate School of Education at Rutgers University.

Acknowledgements
This research is funded by the Department of Education (R305A120778). Any opinions expressed are those of the authors and do not necessarily reflect those of the funding agencies.