Dusting Off the Messy Middle: Assessing Students' Inquiry Skills Through Doing and Writing

Haiying Li(✉), Janice Gobert, and Rachel Dickler

Graduate School of Education, Rutgers University,
New Brunswick, NJ 08904, USA
{Haiying.li, Janice.Gobert,
Rachel.Dickler}@gse.rutgers.edu

Abstract. Researchers are trying to develop assessments for inquiry practices to elicit students' deep science learning, but few studies have examined the relationship between students' ***doing***, i.e. ***performance assessment,*** and ***writing,*** i.e. ***open responses,*** during inquiry. Inquiry practices include generating hypotheses, collecting data, interpreting data, warranting claims, and communicating findings [1]. The first four practices involve "doing" science, whereas the last involves writing scientific explanations, i.e. arguing using evidence. In this study, we explored whether what students wrote in their constructed responses reflected what they did during science inquiry in the Inq-ITS system. Results showed that more than half of the students' writing did not match what they did in the environment. Findings revealed multiple types of students in the *messy middle*, which has implications for both teacher instruction and intelligent tutoring systems, such as Inq-ITS, in terms of providing real-time feedback for students to address the full complement of inquiry practices [1].

Keywords: Inquiry skills · Explanation skills · Log files · Constructed response · Doing science

1 Introduction

Next Generation Science Standards [1] and a framework for K-12 science education [2] expect students to demonstrate grade-appropriate proficiency in inquiry practices and understanding of core scientific ideas. These inquiry practices can be classified into two major categories: "doing" and "writing" scientific explanations (also called arguments or argumentation). The former consists of procedural knowledge including how to generate a research question, formulate a hypothesis, collect data from an experiment, analyze and interpret data, and select data to warrant claims. The latter involves constructing responses in order to communicate findings and argue a claim using evidence.

The original version of this chapter was revised. The spelling of the third author's name was corrected. The erratum to this chapter is available at https://doi.org/10.1007/978-3-319-61425-0_83

E. André et al. (Eds.): AIED 2017, LNAI 10331, pp. 175–187, 2017.
DOI: 10.1007/978-3-319-61425-0_15

To achieve the expectations of NGSS, researchers have developed intelligent tutoring systems (ITSs) [3] or 3-D videogames [4] to teach and assess science inquiry skills in computer-assisted learning and assessment environments.

These environments stealthily record a myriad of students' actions and behaviors that are saved in the form of log files. Typically, the log files record the forced answers that students select from multiple-choice questions, dropdown menus, clickable buttons, or drag-and-drops. Digital environments also record students' constructed responses, such as scientific explanations written in open response format. For all actions that students make, the log files record the corresponding response time. These log files provide researchers with substantial information on the processes that occur during inquiry as well as during the composition of explanations. Some researchers have examined experimentation data from log files to identify whether students designed controlled experimental trials (e.g., [5, 6]), specifically by changing one target variable at a time [7, 8]. A few researchers have analyzed experimentation data to evaluate performance on constructed causal explanations in the format of multiple-choice questions (e.g., [9]). Other researchers have examined constructed explanations using a content-reasoning matrix assessment framework to explicitly demonstrate the range of students' explanation skills from intermediary or middle knowledge to more sophisticated understanding [10].

Previous studies have primarily concentrated on either inquiry skills (such as experimenting) or written explanation skills. Few studies, to date, have developed an assessment of the full complement of inquiry practices to score student performance that includes both inquiry skills and explanation skills. It is uncertain whether students who are good at designing and conducting experiments can also produce strong scientific explanations, as their writing skills may not be sufficiently developed to do so. Likewise, some students are able to parrot what they have heard or read and can produce satisfactory explanations, but their understanding, as reflected and demonstrated by their experimentation, is lacking. In either case, an assessment could be negatively or positively biased depending on which data are used.

The present study aims to examine whether students' inquiry skills for designing and conducting an experiment reflect their performance on writing scientific explanations within the Inq-ITS system (Inquiry Intelligent Tutoring System; inq-its.org). We use the term "*inquiry skills*" to refer to the behaviors involved in "doing" science that are captured in the log data. These behaviors consist of generating a hypothesis, collecting data, interpreting data, and warranting claims with data. We use the term "*explanation skills*" to refer to the scientific explanations constructed in an open response format. This study will significantly enhance science inquiry assessment for the following three reasons. First, it will provide a panoptic view of students' skills for science inquiry practices by integrating both doing science and writing a scientific explanation into the assessment. This method will allow for a clear investigation of the messy middle [10], as commonly acknowledged by assessment researchers, because using both types of data provides a complementary data set. This will also provide teachers, researchers, students, parents, and stakeholders with a more accurate form of assessment for the full complement of science inquiry practices. Second, Educational Data Mining (EDM), used as an automatic measure of inquiry skills [3], is able to capture student behaviors that are representative of authentic skills for science inquiry. Third, explanation skills are examined at the sublevel of knowledge components (KCs) instead of macro-level KCs to reduce

ambiguity for human grading (see Method section for details). Scoring sub-KCs of claim, evidence, and reasoning helps raters avoid subjective bias and judgment when grading, and hence yields higher agreement. For example, we used a general rubric [10] and our rubric with sublevel KCs to score students' reasoning, and interrater reliability as measured by Pearson correlation increased from .55 to .88.

This paper has four sections. First, we briefly review current approaches to the assessment of science inquiry, specifically based on doing science and writing explanations. Second, we describe how to assess inquiry skills and explanation skills in the Method section. Third, we display results and discuss the findings in terms of the relationship between inquiry and explanation skills. Fourth, implications for teachers and researchers are discussed.

1.1 *Doing* Science

Accurate and appropriate assessments can be used to guide teachers in making instructional decisions. The types of assessments adopted in classrooms range from the traditional elicitation-response-evaluation pattern, such as open-ended investigation (e.g., [11]), to newly-emerged assessments (e.g., [12]). Even though the latter form involves thinking and developing knowledge in disciplinary practices, this type of summative assessment could not capture the intermediary processes involved in science inquiry. Formative assessments that occur during the inquiry process allow for adapting and individualizing instruction to improve students' learning.

Many researchers have developed computer-assisted learning and assessment environments to evaluate science inquiry. The computer-assisted assessment saves students' actions and response times in log files. The log files provide not only students' inquiry products, but also their inquiry processes [13]. For example, Gobert et al. [3] developed automated measures for assessing science inquiry skills for designing and conducting experiments using EDM on students' log files. This approach combined text replay tagging and educational data mining to develop a detector to assess science inquiry skills based on what students did during inquiry. Even though log files are collected in a nonintrusive way [14] and provide an informative progression of inquiry practices [3], to date, most researchers do not include performance assessment based on log data. This is probably due to the volume and complexity of log data and the challenge in analyzing it [15]. Instead, most researchers continue to focus on assessments based on a final product.

1.2 Writing Explanations

Scientific explanations in inquiry practices are purported to assess students' core conceptual understandings and reasoning about key scientific ideas used in inquiry [1, 2]. Scientific explanations require students to construct responses that can elicit critical thinking and involve making connections between scientific concepts and evidence [10, 16]. This in turn requires assessment of complex, higher-order cognitive processes [17, 18]. Toulmin's [19] model of argumentation is widely used as a framework for

scientific explanations. The modified version consists of three components: claim (a statement that establishes a conclusion for the investigated question), evidence (data or observations that support or refute the claim), and reasoning (the scientific principle that connects data to the claim and makes visible the reason why the evidence supports or refutes the claim) [10, 16]. Prior research has shown that it is difficult for students to communicate their knowledge about science (i.e. articulating and justifying their claims with sufficient and appropriate evidence [20, 21]), distinguish evidence from theory [16], link their claim and evidence to scientific ideas [16], or use evidence to support their claim [22].

Researchers have assessed inquiry by examining content knowledge with procedural understanding [11, 23], content knowledge with reasoning skills [10], or predicting causal explanations generated by multiple-choice questions based on experimentation behaviors [9]. No studies have investigated procedural performance via doing science and performance on causal explanation via writing in science inquiry.

This study investigated three research questions: (1) to what extent do students' inquiry skills reflect their explanation skills? (2) what distribution is displayed in terms of high versus low inquiry skills and high versus low explanation skills? and (3) to what extent does performance on inquiry and explanation differ among the four groups (High-High, High-Low, Low-High, Low-Low with inquiry before explanation)? We hypothesize that inquiry performance can explain part of explanation performance because both of these skills may require certain domain-specific conceptual knowledge. However, as experimentation involves procedural understanding [24], doing experimentation may have its own unique features that do not reflect explanation skills. Similarly, as explanations involve connecting theory with data using reasoning, writing explanations may have unique characteristics involved in coherently synthesizing information. The second question may illustrate that there are some students who have developed good inquiry skills, but are not good at articulating their understanding as represented by their explanation. Many highly spatial science/math students could fall into this category. Under current assessment tests, such as state multiple-choice tests, these students are at risk for being assessed as not knowing science when they are actually highly skilled at conducting key inquiry practices. Conversely, those who are unskilled at inquiry but skilled at writing explanations are likely students who are parroting what they have "learned" in science class. Under current assessment tests, these students are at risk for being assessed as knowing science when their understanding is very superficial.

2 Method

2.1 Participants and Materials

293 middle school students from 18 classes in six public middle schools completed one Inq-ITS density virtual lab (inqits.com). Inq-ITS is a web-based intelligent tutoring and assessment system for Physical, Life, and Earth science that automatically assesses scientific inquiry practices at the middle school level in real time within interactive microworld simulations [3]. Within each microworld, inquiry practices proposed in the

NGSS for middle school are assessed including: hypothesizing, collecting data, analyzing data, interpreting data, warranting claims, and communicating findings. The Density Virtual Lab contained three activities aimed to foster understanding about the density of different liquid substances (water, oil, and alcohol), different amounts of liquid (quarter, half, and full), and different shapes of the container (narrow, square, and wide). This study analyzed the data in the last activity, the shape of the container.

Students completed four stages of inquiry over the course of the Density Virtual Lab, as illustrated in Fig. 1 and also in demos on the Inq-ITS website (inqits.com). During the Hypothesis stage, students used a widget (dropdown menu) to formulate a hypothesis that measured an activity goal. In the Collect Data phase, students used a widget (clickable buttons) to manipulate the independent variables in a simulation while a data table automatically recorded their findings. During the Analyze Data stage, students used a widget (dropdown menu) to state their claim, identified whether or not their claim supported their hypothesis, and selected evidence that supported their claim (clickable). Communicate Findings was the final inquiry stage where students responded to three open response questions in order to explain their claim, evidence, and reasoning for how their evidence supported their claim (writing). The first three stages are involved in doing science and we refer to the skills involved in doing science as inquiry skill. The last stage involves writing a scientific explanation and we refer to the skills involved in writing as explanation skill.

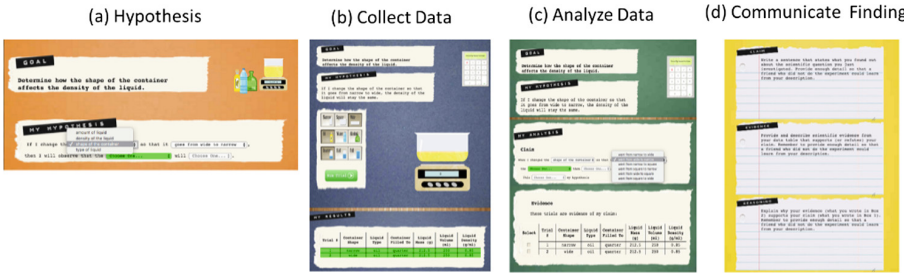

Fig. 1. Stages of the density virtual lab in Inq-ITS.

2.2 Measures

Inquiry skills were measured by four components using educational data mining techniques in Inq-ITS (see [3] for details). Each component contained sublevel KCs: (1) hypothesis (a. the identification of the independent variable (IV) and b. the identification of the dependent variable (DV)), (2) data collection (a. designing controlled experiment, b. testing hypothesis, and c. changing pairwise IV and controlled variable), (3) data interpretation (a. interpreting IV, b. interpreting DV, c. interpreting IV-DV relationship, and d. supporting hypothesis), and (4) warranting claims (a. warranting IV-DV relationship, b. number of single trial, c. supporting hypothesis, and d. all controlled trials). Each sublevel KC was automatically graded with binary scores, 0 for failing or 1 for succeeding at a skill. The inquiry score adopted the total of all the sublevel KC scores, with 0 as the minimum score and 13 as the maximum score.

The scientific explanation consisted of three components: claim, evidence, and reasoning. Each component was graded according to a scoring rubric that was modified based on the previous rubrics used by researchers (e.g., [10]) (see Table 1). The previous rubrics for claim and reasoning classified responses into incorrect or correct, but failed to specify the extent to which the claim was correct or incorrect. Similarly, the previous rubrics for evidence categorized three types of evidence: complete, incomplete, and incorrect. To determine which category evidence belonged to depended on raters' subjective ratings. This ambiguity reduces the agreement among human raters. Since previous coding schemes were too general to apply to Inq-ITS claim, evidence, and reasoning, we created a more specific coding scheme, as displayed in Table 1.

Table 1. Rubrics for scoring claim, evidence, and reasoning.

Type	KC		2 points	1 point	0.5 points	0 points
Claim (0–4)	IV		X	Shape	X	Incorrect IV
	IVR		X	2 shapes	1 shape	Incorrect IVR
	DV		X	Density	X	Incorrect DV
	DVR		X	Stays the same	X	Incorrect DVR
Evidence (0–4)	Sufficient		2 shapes	1 shape	X	No shape
	Appropriate	Mass + Volume	X	Data of mass & volume	Data of mass or volume	Incorrect data
		Density	X	Data of density	X	No density
Reasoning (0–6)	Theory		Mass/volume = density or property of substance	Mass + density or volume + density or partial property	Mass or volume	Incorrect theory
	Connection		X	Data supports/refutes claim	Partial connection	Incorrect connection
	Data	IV/IVR	X	Shape or 2 shapes	1 shape	Incorrect IV/IVR
		DV	X	Density	X	Incorrect DV
		DVR	X	Stays the same	X	Incorrect DVR

Note. 2 shapes = two of three types of shape (narrow, square, wide). 1 shape = any one of three types of shape. Shape means mentioning the word "shape."

In Inq-ITS, the widget claim is constructed with four knowledge components (KCs), IV, IVR (IV relationship, namely, any two of three types of shape; e.g., from narrow to wide), DV, and DVR (DV relationship, namely, state of density; e.g., stays the same). Therefore, the written claim was graded with the same four KCs. Written evidence was graded in terms of sufficiency and appropriateness [21]. Sufficiency was a measure of whether students provided sufficient evidence, namely, whether students specified changing the shape of container from one shape to another. Mentioning only one specific shape was considered insufficient evidence and not mentioning any specific shape was considered incorrect. During data collection, a data table displayed the values for mass,

volume, and density for each trial a student ran. Appropriateness was a measure of whether students provided appropriate data, specifically the data for mass, volume, and density. Reasoning was composed of three sublevel KCs: theory, connection of data to theory, and data. Theory referred to whether students stated the nature of density, namely, density is only affected by the property of substance or the ratio of mass to volume. Data referred to whether students generalized the data, such as "The shape of the container does not affect the density of the liquid." Data-Theory connection referred to whether students specified that their data supports or refutes their claim.

Two expert raters discussed the rubrics and then graded for each KC or sublevel KC. The maximum score for claim and evidence was 4 points, respectively. The maximum score for reasoning was 6 points. Thus, the total possible score for explanation was 14 points. Inter-rater reliability was assessed by the intraclass correlation coefficient with a two-way random model and absolute agreement type [25]. The interrater-reliabilities by Cronbach's α were .993, .994, .938 and the intraclass correlations were .986, .988, .882 for claim, evidence, and reasoning, respectively. Then two raters discussed the disagreements and generated agreement scores. The agreement scores were used to compute the total scores of explanation skills.

2.3 Statistical Analysis

The analyses adopted the standardized total scores of inquiry skills and explanation skills. The relationship between inquiry skills and explanation skills was examined using linear regression. We performed *K*-means cluster analyses ($K = 2$) on the scores for inquiry skills and explanation skills, respectively, and classified students into low versus high for both inquiry and explanation. We performed the Chi-square analysis on inquiry group and explanation group to examine the distribution of students among these four quadrants. Finally, a multivariate general linear model (GLM) was performed to examine the extent to which the performance on inquiry skills and explanation skills differed among these four groups. The two dependent variables were inquiry skills and explanation skills. The independent variable was the four groups: High-High, High-Low, Low-High, and Low-Low with inquiry before explanation.

3 Results and Discussion

3.1 Results of Linear Regression

Results of linear regression showed that the total scores of inquiry skills significantly predicted the total scores of explanation skills, $B = .53$, $t(291) = 10.63$, $p < .001$. Results suggest that inquiry skills could explain 28% of the variance in explanation skills, $R^2 = .28$, $F(1, 291) = 112.99$, $p < .001$. However, the majority of variance (about 72%) in explanation skills could not be explained by inquiry skills. These findings imply that these two types of skills possess unique characteristics that represent unique constructs. The shared variance may represent the shared content knowledge (the relationship between the shape of the container and the density) that students apply when they do science and write an explanation. During experimentation,

however, knowledge about doing experiments is needed, such as how to formulate a hypothesis, how to test the hypothesis by designing a controlled experiment, how to collect appropriate and sufficient data, how to generate a claim based on the collected data, and how to warrant a claim. The process of doing experiments involved procedural knowledge, which is unlikely to be captured in a written explanation.

On the other hand, constructing an explanation requires knowledge about logic and writing coherently. For example, students must understand what components should be included in a good claim. Most students did not specify how they controlled the target IV (e.g., The shape did not change density.); thus, generated an incomplete claim. Students needed to report the specific data in the evidence, but they only repeated their conclusion in this section. In reasoning, students needed to specify theory and connect data to theory to further support the claim. In fact, most students were confused by claim, evidence, and reasoning and repeated the same contents in each section. Therefore, writing an explanation requires writing skills, especially in terms of how to generate a coherent and complete claim, sufficient and appropriate evidence, and a theory that links to data to support and validate a claim.

3.2 Results of Chi-Squire

Results of Chi-square showed that inquiry skills and explanation skills were not independent (see Table 2), $\chi^2(1, N = 293) = 6.18$, $p = .013$. Specifically, 46.5% of students with high inquiry skills and 27.1% of students with low inquiry skills wrote a high quality scientific explanation. Moreover, 53.5% of students with high inquiry skills and 72.9% of students with low inquiry skills wrote a low quality scientific explanation. In addition, results showed a subset of the explanation group whose column proportions did not differ significantly from each other at the .05 level. Specifically, 89.8% of students with high explanation skills had high inquiry skills, whereas 10.2% had low inquiry skills. Conversely, 78.9% of students with low explanation skills had high inquiry skills, whereas 21.1% had low inquiry skills. These findings imply that 49.1% of the total students showed "middle" knowledge. Among them, 44.7% had high inquiry skills, but low explanation skills and 4.4% had low inquiry skills, but high explanation skills. 50.9% of the total students showed consistent knowledge: 38.9% achieved both high inquiry and explanation skills and only 11.9% had both poor inquiry and explanation skills.

Table 2. Inquiry group and explanation group ($N = 293$)

		Explanation Skill		Total	χ^2 ($df = 1$)
		High	Low		
Inquiry skill	High	114 (47.5)	131 (53.5)	245 (100)	6.18*
	Low	13 (27.1)	35 (72.9)	48 (100)	
Total		127 (43.3)	166 (56.7)	293	

Note. * $p < .05$. Numbers in parentheses are the percentage in each category.

Approximately half of the total students exhibited "middle" knowledge. These students showed intermediary knowledge in terms of inquiry and explanation skills. From the perspective of assessment, if they are assessed based on only one of these skills, they will be mistakenly evaluated. This is true for students who are good at doing science, but not skilled at writing explanations; as well as for students who are good at writing, but not skillful at doing science. If the former group of students is encouraged and trained in writing (or the latter in doing science), then students may have greater opportunity to excel as scientists in the future. However, if they are inaccurately reported as students who are poor at science based on their writing or doing science skills, we may not recognize the potential of a number of promising scientists. Hence, it is very important to assess science inquiry comprehensively with both doing science and writing about science.

3.3 Results of GLM

Table 3 displays the descriptive statistics of inquiry skills and explanation skills among four groups: High-High, High-Low, Low-High, and Low-Low with inquiry before explanation. Results of multivariate general linear model revealed a statistically significant difference in inquiry skills based on group, $F(6, 578) = 230.35$, $p < .001$; $\eta^2 = .705$. Tests of between-subjects effects indicated that group had a statistically significant effect on both inquiry scores ($F(3, 289) = 311.06$; $p < .001$; $\eta^2 = .764$) and explanation scores ($F(3, 289) = 226.64$; $p < .001$; $\eta^2 = .702$). The table below shows that mean scores for inquiry skills were significantly different between any two groups ($p < .001$). Mean explanation scores were also statistically different between any two groups ($p < .001$), except between High-High and Low-High ($p = 1.000$). The pattern of performance of inquiry skills is displayed: High-High > High-Low > Low-High > Low-Low. The pattern of performance of explanation is listed: High-High = Low-High > High-Low > Low-Low.

Table 3. Descriptive statistics

Group	N	Inquiry skills		Explanation skills	
		Mean	SD	Mean	SD
High-High	114	0.63	0.53	0.95	0.56
High-Low	131	0.13	0.40	−0.62	0.52
Low-High	13	−1.24	0.36	0.86	0.37
Low-Low	35	−2.10	0.67	−1.09	0.66
Total	293	0.00	1.00	0.00	1.00

Note. Group displays inquiry skills first, followed by explanation skills.

These findings further indicate that inquiry and explanation skills are differently represented in each group. Specifically, students with high explanation skills could consistently write good explanations irrespective of their inquiry skills. Conversely, students with high inquiry skills could do science better when explanation skills were

high than when explanation skills were low. This pattern exists among students whose inquiry skills were low: when their explanation skills were high, they could do better science than when their explanation skills were low (even though their absolute scores remained lower relative to students with high inquiry skills). For students whose explanation skills were low: when inquiry skills were high, they wrote better explanations than when inquiry skills were low. To sum up, if students are good at conducting experiments, these skills are likely to help them yield better performance on writing. Conversely, if students are good at writing scientific explanations, these skills are less likely to help them do better science as writing is the final step and would not impact their inquiry.

4 General Discussion and Implications

In this study, we explored whether what students wrote in their constructed responses reflected what they did during science inquiry in the Inq-ITS system. Results indicated that students' skills at doing science only explained 28% variance in writing an explanation. The 72% of unexplained variance is probably explained by the unique skills involved in writing. Similarly, inquiry skills involved a series of procedural knowledge while doing science. Chi-square analysis demonstrated that nearly half of the students' writing did not match with their "doing". Findings revealed two types of the *messy middle*, which further illustrated that approximately half of the total students were good at doing science, but not good at writing explanations (44.7%). However, there were few students who were good at writing explanations, but not good at doing science (4.4%). Students who were good at both accounted for 38.9%, whereas those who performed poorly on both skills accounted for 11.9%. Multivariate analysis further indicated that each group performed differently on inquiry skills and explanation skills, except for High-High and Low-High groups on explanation skills. Our study dusts off the messy middle knowledge between inquiry skills and explanation skills, unfolds the complex middle knowledge between doing and writing in science inquiry practices, and provides implications for teachers and researchers when they design instruction and assessment for science inquiry.

Our study provides empirical evidence that science inquiry assessment by either doing science or writing scientific explanations does not capture the students' overall inquiry skills. This study explicitly demonstrated that these two skills only shared a small portion of variance because they each involve unique constructs. Only about 40% students developed equivalent, high inquiry skills and explanation skills. Another 10% had equivalent but poor skills. Another half did not develop equivalent skills. Among them, about 45% students failed to write good explanations in their open responses, even though they had designed and conducted a good experiment to test their hypotheses. One possible explanation is that students did not know what information they should put in the claim, evidence, and reasoning in their open responses. Another explanation is that students had not reified what they knew into their mental model of the phenomena under investigation. In this situation, teachers or computer tutors in an ITS could provide scaffolding for students for claim, evidence, and reasoning:

(1) *Claim*. Prompt students that the written claim should be consistent with the experimentation process conducted. Specifically, the written claim should contain the same four components as displayed in the widget claim; (2) *Evidence*. Prompt students to observe how the data table presents data and that the written evidence also needs to display data with the values of mass, volume, and the corresponding density; and (3) *Reasoning*. Scaffold student to understand that reasoning should include a theory that supports the claim, data that supports the claim, and then how data connects to the theory. This scaffolding may lead students to construct deep mental models which reflect both their doing of science and their writing explanations about the phenomena under investigation.

There were a few students who were poor at inquiry skills, but skillful at writing explanations. It is possible that these students were parroting what they had "learned" in science class, but they were not clear about how to "do" science. For these students, it is necessary to scaffold them on procedural knowledge that is required for designing and conducting experiments, such as how to collect controlled trials for a specific research question and how to select appropriate and sufficient data to support a claim.

Students who were poor in both inquiry skills and explanation skills might not have mastered content knowledge or procedural knowledge for conducting a controlled experiment. This means that teachers or a computer tutor should not scaffold students based solely on either doing or writing, but from the inquiry phase where students showed difficulties. Thus, when students successfully complete experiments, they can continue on to their writing. It is better to remind students how information is displayed during experimentation and tell them they could use the same format when writing their explanations. Similarly, when it is time for them to write, they could be reminded of how claim and evidence is presented during experimentation. This scaffolding would enhance students' skills to build connections between doing and writing, and consequently write a good explanation.

This study reveals students' "messy middle knowledge" in science inquiry, which explicitly informs teachers and researchers of the students' complex learning patterns and helps them develop adaptive and individualized instruction, curriculum, or scaffolding. This study also suggests that science inquiry should be interactively assessed by evaluating both inquiry and explanation skills so as to avoid biased judgment. Even though the current study successfully uncovered unequal performance between inquiry and explanation skills, one limitation would be that we focused on the macro-level of inquiry and explanation skills by aggregating the scores of the subskills. In future work, we will further investigate whether the same phenomenon consistently exists by: (1) analyzing the subskills that co-occur in both inquiry and explanation processes, such as claim and evidence, and (2) adding more activities in the analyses. Understanding what, how, and why middle knowledge occurs facilitates adaptive feedback and scaffolding in an ITS.

Acknowledgement. The research reported here was supported by Institute of Education Sciences (R305A120778) to Janice Gobert.

References

1. Generation Science Standards Lead States: Next generation science standards: for states, by states. National Academies Press, Washington (2013)
2. National Research Council: A framework for K-12 science education: practices, crosscutting concepts, and core ideas. National Academies Press, Washington (2012)
3. Gobert, J.D., Sao Pedro, M., Raziuddin, J., Baker, R.S.: From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. J. Learn. Sci. **22**, 521–563 (2013). doi:10.1080/10508406.2013.837391
4. Shute, V.J.: Stealth assessment in computer-based games to support learning. In: Tobias, S., Fletcher, J.D. (eds.) Computer Games and Instruction, pp. 503–524. IAP (2011)
5. Sao Pedro, M.A., Gobert, J., Baker, R.S.: The development and transfer of data collection inquiry skills across physical science microworlds. In: Paper presented at the American Educational Research Association Conference (2012)
6. Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., Nakama, A.: Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. User Model. User Adap. **23**, 1–39 (2013)
7. Chen, Z., Klahr, D.: All other things being equal: acquisition and transfer of the control of variables strategy. Child Dev. **70**, 1098–1120 (1999)
8. Tschirgi, J.E.: Sensible reasoning: a hypothesis about hypotheses. Child Dev. **51**, 1–10 (1990)
9. Baker, R.S., Clarke-Midura, J., Ocumpaugh, J.: Towards general models of effective science inquiry in virtual performance assessments. J. Comp. Assist. Learn. **32**, 267–280 (2016)
10. Gotwals, A.W., Songer, N.B.: Reasoning up and down a food chain: using an assessment framework to investigate students' middle knowledge. Sci. Educ. **94**, 259–281 (2010)
11. Roberts, R., Gott, R., Glaesser, J.: Students' approaches to open-ended science investigation: the importance of substantive and procedural understanding. Res. Pap. Ed. **25**, 377–407 (2010). doi:10.1080/02671520902980680
12. Roberts, R., Johnson, P.: Understanding the quality of data: a concept map for 'the thinking behind the doing' in scientific practice. Curriculum J. **26**, 345–369 (2015)
13. Rupp, A.A., Gushta, M., Mislevy, R.J., Shaffer, D.W.: Evidence-centered design of epistemic games: measurement principles for complex learning environments. J. Technol. Learn. Assess. **8**, 1–48 (2010)
14. National Research Council: Knowing what students know: the science and design of educational assessment. In: Pellegrino, J.W., Chudowsky, N., Glaser, R. (eds.) National Academies Press, Washington (2001)
15. Quellmalz, E., Timms, M., Schneider, S.: Assessment of student learning in science simulations and games. National Research Council, Washington (2009)
16. McNeill, K., Lizotte, D.J., Krajcik, J., Marx, R.W.: Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. J. Learn. Sci. **15**, 153–191 (2006). doi:10.1207/s15327809jls1502_1
17. Liu, O.L., Lee, H.S., Linn, M.C.: Multifaceted assessment of inquiry-based science learning. Ed. Assess. **15**, 69–86 (2010). doi:10.1080/10627197.2010.491067
18. Martınez-Cortizas, A., Pontevedra-Pombal, X., Garcıa-Rodeja, E., Novoa-Munoz, J.C., Shotyk, W.: Mercury in a Spanish peat bog: archive of climate change and atmospheric metal deposition. Science **284**, 939–942 (1999)
19. Toulmin, S.: The Uses of Argument. Cambridge University Press, Cambridge (1958)

20. Sadler, T.D.: Informal reasoning regarding socioscientific issues: a critical review of research. J. Res. Sci. Teach. **41**, 513–536 (2004)
21. Sandoval, W.A., Millwood, K.A.: The quality of students' use of evidence in written scientific explanations. Cogn. Instruct. **23**, 23–55 (2005)
22. Hogan, K., Maglienti, M.: Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. J. Res. Sci. Teach. **38**, 663–687 (2001). doi:10.1002/tea.1025
23. Gobert, J.D., Pallant, A.R., Daniels, J.T.: Unpacking inquiry skills from content knowledge in geoscience: a research and development study with implications for assessment design. Int. J. Learn. Technol. **5**, 310–334 (2010)
24. Glaesser, J., Gott, R., Roberts, R., Cooper, B.: The roles of substantive and procedural understanding in open-ended science investigations: using fuzzy set qualitative comparative analysis to compare two different tasks. Res. Sci. Ed. **39**, 595–624 (2009). doi:10.1007/s11165-008-9108-7
25. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. **86**, 420–428 (1979). doi:10.1037/0033-2909.86.2.420