Research in Science Education
https://doi.org/10.1007/s11165-022-10061-x

Learning Analytics for Assessing Hands-on Laboratory Skills in Science Classrooms Using Bayesian Network Analysis

Shiyan Jiang[1] · **Xudong Huang**[2] · **Shannon H. Sung**[2] · **Charles Xie**[2]

Accepted: 19 June 2022

Abstract
Learning analytics, referring to the measurement, collection, analysis, and reporting of data about learners and their contexts in order to optimize learning and the environments in which it occurs, is proving to be a powerful approach for understanding and improving science learning. However, few studies focused on leveraging learning analytics to assess hands-on laboratory skills in K-12 science classrooms. This study demonstrated the feasibility of gauging laboratory skills based on students' process data logged by a mobile augmented reality (AR) application for conducting science experiments. Students can use the mobile AR technology to investigate a variety of science phenomena that involve concepts central to physics understanding. Seventy-two students from a suburban middle school in the Northeastern United States participated in this study. They conducted experiments in pairs. Mining process data using Bayesian networks showed that most students who participated in this study demonstrated some degree of proficiency in laboratory skills. Also, findings indicated a positive correlation between laboratory skills and conceptual learning. The results suggested that learning analytics provides a possible solution to measure hands-on laboratory learning in real-time and at scale.

Keywords Laboratory skills · Bayesian networks · Science education · Learning analytics · Mobile application

Introduction

Improving and assessing hands-on laboratory skills is one of the top priorities in science education (Hensiek et al., 2016; Hofstein, 2017; NRC, 2006; Zhang et al., 2020). The laboratory skill is students' ability to perform scientific practices in science labs. Students could minimize the errors in measurement and data collection, and successfully observe and analyze scientific phenomena when mastering the essential laboratory skills (Prichard, 2003). In addition, obtaining hands-on laboratory skills can help secondary

✉ Shiyan Jiang
sjiang24@ncsu.edu

[1] Department of Teacher Education and Learning Sciences, North Carolina State University, Poe Hall, 208, 2310 Stinson Dr, Raleigh, NC 27695, USA

[2] Institute for Future Intelligence, Natick, MA, USA

school students be well prepared for college-level science courses (Leggett et al., 2004). Therefore, it is vital to explore *alternative (*Doran et al., 1993) ways of assessing and supporting the development of laboratory skills in K-12 science education.

Learning analytics, the core idea of which is that decision-making regarding the administration of learning should be guided by data, is proving to be a powerful approach for understanding and improving science education (Jiang et al., 2022; Geden et al., 2021; Zhai et al., 2020). For instance, researchers have developed Inq-ITS to assess scientific inquiry skills through mining log data of students' interactions with science simulations in the system (Gobert et al., 2013). Most log data in science learning settings, in particular, laboratories, were obtained from virtual environments. Such virtual laboratories offer a low-cost opportunity for student inquiry (Li et al., 2018). However, simulated experiments cannot replace the role of hands-on experiments. Its nature of virtuality sometimes led to low responsibility, as some students viewed learning in a virtual environment as playing a video game that nothing wrong could happen (Potkonjak et al., 2016). To our knowledge, few studies have focused on assessing hands-on laboratory skills using learning analytics in physical labs, in particular, in K-12 learning settings.

Adopting mobile devices in the physical lab provides educators an opportunity to mine educational data and apply learning analytics to capture dynamic scientific practices. We could investigate students' hands-on laboratory learning leveraging log data captured by the sensing technology in mobile devices. Such log data has great research potential as it could be used for fine-grained analysis. This approach is similar to how log data generated by a virtual learning system could be utilized to enhance learning (Park & Jo, 2017). The logs collected by mobile devices have been used to identify meaningful learning patterns. For example, Chiang et al. (2014) designed a mobile application for elementary school students' ecology field trip. During the trip, the application recorded students' actions on tablets such as capturing images, adding comments, sharing data, and posting questions. Then the researchers identified students who had a high-level knowledge construction through mining the log data of student-application interactions. Collectively, the literature shows that mining mobile log data provides a unique opportunity to automate the analysis of physical learning activities.

However, few studies involving mobile log data were carried out in K-12 laboratories; a laboratory is commonplace in schools to help students link theories to practice and build practical skills though (Beaumont-Walters & Soyibo, 2001). In particular, using mobile log data to understand hands-on laboratory learning was not thoroughly investigated in the literature (Chang et al., 2020). This is an underexplored area as it is difficult to develop applications to capture and record scientific activities in physical laboratories. It is even more challenging to design automated analysis tools to analyze student interactions with mobile devices because of the variety of scientific practices (e.g., manipulating laboratory equipment and taking measurements) that could be captured. This study serves as a first step in developing and studying automated analysis tools for physical laboratory learning by leveraging learning analytics to automatically assess hands-on laboratory skills. Specifically, we examine the following research questions:

- How could mobile log data be used to automatically measure students' hands-on laboratory skills?
- What is the relationship between students' laboratory skills and conceptual understanding?

Literature Review

In science labs, practices related to making observation and gathering evidence are recommended as the priority of science teaching (National Science Teachers Association, 2007). The typical ones include correct setup of equipment, appropriate use of tools, accurate measurement of variables, actualization of experimental operations, and complete gathering of experimental data (Gobaw & Atagana, 2016; Jou & Wang, 2013). Developing hands-on laboratory skills are crucial to prepare secondary students for advanced study of science in colleges and for employment in careers related to science, technology, engineering, and mathematics in the future.

The indirect way of assessing laboratory skills is to score written tests or lab reports and the direct way is to rate how students actually perform in the laboratory (e.g., Hunt et al., 2012). Researchers showed that the direct rating was a better index of laboratory skills to supplement or replace the traditional writing tests (Lunetta et al., 2007). Specifically, direct rating can be conducted in two levels: the holistic scoring for overall performance and the analytic scoring for specific skills (Chabalengula et al., 2009). The holistic scoring treats the activity as a whole and gives one single score (i.e., the completion of an experiment), while the analytic approach focuses on various skills and assigns scores for each skill (e.g., tool setup, data collection). Compared to assigning either holistic or analytic scores, it is better to use both (Harsch & Martin, 2013). Following this approach, our study assessed both individual skills and overall performance.

Even though direct ratings are "authentic" and "sensitive" (Hofstein & Lunetta, 2003, p. 43–44), they are not widely adopted in the classroom due to several challenges, such as challenges in observing and rating each experimenter in a large-size class. Therefore, many teachers tend to use summative lab reports to understand students' laboratory experience, and the research about direct ratings becomes less visible in the twenty-first century (Lunetta et al., 2007). Meanwhile, many students were not prepared well for hands-on laboratory learning. For instance, manipulative errors in laboratory learning were common among secondary and undergraduate students (Gobaw & Atagana, 2016; Minalisa, 2019). A major reason for the low performance in laboratory learning is the lack of real-time feedback (Beaumont-Walters & Soyibo, 2001). Real-time feedback at scale requires automating the analysis of hands-on laboratory lab skills (Zhang et al., 2020).

The fast adoption of smartphones to science laboratories and recent emergence of learning analytics in assessment provide a unique opportunity to automate the analysis of mobile log data to measure lab skills (Zhai et al., 2020). One challenge that researchers face when approaching automatic behavior modeling is the high level of uncertainty (Conati et al., 2002). This challenge is even more prominent in a real-world lab where uncertainty abounds. Various learning analytics approaches have emerged to address uncertainties in the learning settings, and among those, the Bayesian network modeling is an important one (Fan et al., 2021). The Bayesian network approach could be utilized to model the interdependencies of variables in a real-world scenario and deal with uncertainties. Also, it allows the incorporation of expert judgments to improve inference accuracy (Zhou et al., 2014). Such customization can help a Bayesian network achieve a high prediction accuracy even with small data samples (Zhou et al., 2014). Furthermore, Bayesian networks' probabilistic inferences can happen in real-time. Therefore, educators and students can receive dynamic feedback based on results from Bayesian networks (Tadlaoui et al., 2016).

One more advantage that makes Bayesian networks especially suitable to measure laboratory skill is its ability to follow the scoring method recommended by the

literature—combining analytic scores into a holistic score (Harsch & Martin, 2013). Bayesian networks can realize this approach by using a standard format—the tree structure. The tree's root nodes are the observable features (e.g., taking images). A level above them is the nodes representing individual skills (e.g., complete data gathering). Sitting above these two layers is the top-level node presenting the overall lab performance. In this way, we can understand the lab practices from both sub-skill and overall perspectives. In this study, we utilized a Bayesian network to analyze students' process data logged by a mobile application in science experiments.

Method

Learning Technology and Context

In this study, students used Infrared Explorer, a mobile AR (augmented reality) application developed by the research team, together with an infrared (IR) camera attached to a smartphone to explore thermal phenomena in the science laboratory (Fig. 1). Infrared Explorer is an application that facilitates both lab investigation and educational research. In the application, users could add thermometers on the points of interest to measure instant temperature, turn on the temperature–time T(t) graph to analyze temperature changes over time, and capture screenshots of thermal phenomena. On the research side, the application

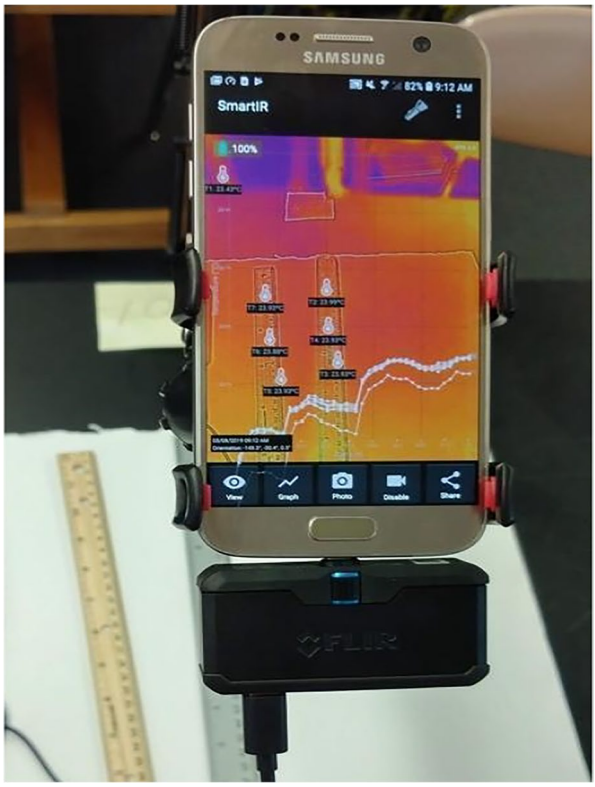

Fig. 1 The experiment set up: a smartphone, the mobile application, a plug-in IR camera, and experimental materials

logs students' interactions with it and records MP4 files of what students see through the infrared camera. The MP4 recordings provide a nuanced view of students' processes of conducting experiments, similar to video recordings for classroom activities but with a particular angle.

Before project implementation, we offered a professional development workshop for Eric (all names are pseudonyms), the science teacher, to learn a five-day curricular unit. The experiments covered concepts of radiation, natural convection, forced convection, conduction, and latent heat (Sung et al., 2021; Xie, 2011). These scientific concepts were covered before the inception of the lab. Students had one science class (approximately one hour) per day. The instructor only spent the first 10 min of the class introducing the main concepts of the lab and the rest of the class time was for students to conduct lab experiments with Infrared Explorer. When implementing the project, Eric placed students in pairs to conduct five science experiments in the curriculum and he circulated among the lab groups to monitor the progress of each group, provided feedback, and answered students' questions. The heat conduction experiment was selected to measure laboratory skills because it contained the most steps compared to other experiments in the curricular unit. Specifically, it included nine steps and the main activity was to compare the conductivity of metal and wood (Fig. 2). While Fig. 2 shows a linear sequence of steps, students went through these steps recursively in the lab.

Participants

A total of 72 seventh graders (female 32, male 37, prefer not to answer 3; American Indian or Alaskan Native 3; Asian/Pacific Islander 4; Latinx 1; Multiple ethnicity 6; White 53;

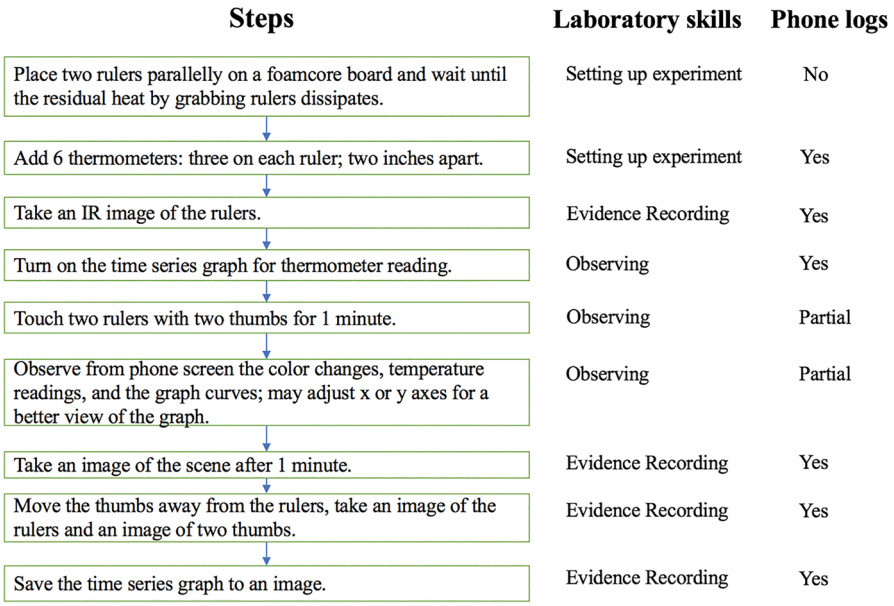

Fig. 2 The heat conduction experiment: steps, laboratory skills, and whether the mobile application captures logs or not

Prefer not to answer 5) from four classes ($n = 16 \sim 20$ in each class) who returned parental consents to participate in this study. They were from a suburban middle school in the Northeastern United States and taught by the same science instructor, Eric. Students conducted experiments in pairs (or a team of three when there was one left) based on whether students provided full consents (pairing those who missed their consent forms in the same groups) and formed 34 teams. After removing students missing parental consent or log files, the data of 30 teams were used for analysis.

Measures

Conceptual Understanding Students' conceptual understanding was measured by coding their written scientific explanations in the lab report. Following the POE pedagogy (Prediction-Observation-Explanation; Gunstone, 1990), students described their understanding of the concept (conductivity in this case) before and after conducting the laboratory experiment by answering open-ended questions (e.g., "when you touch the two rulers, which one will feel cooler? Explain why."). Their answers were scored by two independent researchers. Both coders were researchers (one female and one male) on the project in their first and second year, the second author led the training process until they reached 0.84 inter-rater reliability in the second round of coding, and the scores range from 1 to 3. Specifically, the answer was coded as 1, 2, and 3 when it contains an error or is off-topic, does not contain an error but has no or partial explanation, and is correct and has a detailed explanation respectively. The intercoder reliability was 0.84.

Learning Gain Learning gain refers to the score difference between post-lab and prior-lab conceptual understanding. As the prediction phase had three questions, and the explanation phase had four, we normalized the total scores before making the pre-post comparison (i.e., normalized total score = total score/maximum total score). For example, one student's total score in prediction is six (e.g., $2+2+2$), and the maximum total score is nine (i.e., $3+3+3$), and then her normalized total score in prediction is 0.67, $(2+2+2)/(3+3+3)$. Following the same conversion, her normalized total score in explanation is 0.83, $(2+3+3+2)/(3+3+3+3)$. Then, this student's learning gain is 0.16. Note that each pair submitted one report, so the learning gain is considered a team-level outcome.

Laboratory Skill Analysis Using Bayesian Network Modeling

We used GeNle v3.0.6 to construct and fit the Bayesian network due to its robustness on Bayesian modeling and flexibility on switching between a graphical user interface and programming interface (BayesFusion, 2017). The graph-structure Bayesian network to model laboratory skills is illustrated in Fig. 3. The figure illustrates the initial probabilities of different nodes. For example, before the model was fit, the initial probability of taking a milestone image was 0.25. Likewise, students, in general, were modeled to have a probability of 0.3068 to demonstrate laboratory skills before model fit (calculated with expert-defined conditional probabilities and dependencies' joint probabilities). Details of how initial probabilities were selected can be found in the next few paragraphs.

In this model, laboratory skills contain three major dimensions: experiment setup, observation, and evidence recording (National Science Teachers Association, 2007). The first one is experiment setup, which captures if students prepare thermometers correctly to aid observation and keep devices and thermometers stable to guarantee desired

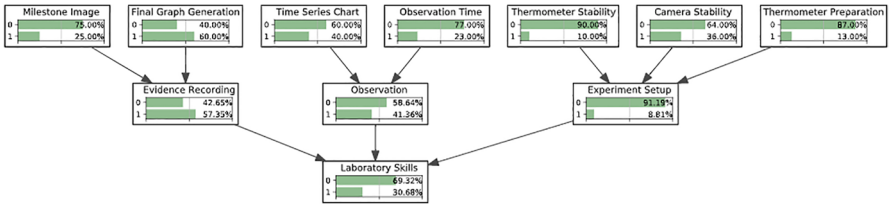

Fig. 3 Initially constructed Bayesian network to measure laboratory skills

temperature trends to be observed. The second one is observation, which indicates if students manage to use the required feature, temperature–time T(t) graph, and wait enough time to observe the intended phenomena. The observation step is operationally defined as no additional actions are logged when the sensor orientation is stabilized for at least 60 s. The last one is evidence recording, which registers if students record milestone screencasts (see Fig. 8 in the "Results" section) along the experiment process and generate a final T(t) graph to assist future analysis in the lab report. The hierarchies of the model allow us to understand both the sub-skills and the overall skill performance.

From the perspective of Bayesian statistics, the probability of a joint distribution with a set of random variables could be represented using the chain rule, as shown in Eq. (1).

$$P\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} P(x_i | x_{i-1} \ldots, x_i) \tag{1}$$

A Bayesian network model satisfies the conditional independence assumption (Pearl, 1988); Eq. (1) could then be simplified as Eq. (2).

$$P\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} P(x_i | Parents(x_i)) \tag{2}$$

The first step to run a Bayesian network is setting the initial probabilities of all the nodes in the network. In our case, the initial probability distributions for root nodes (i.e., observable features such as adding how many thermometers) were calculated from real data. For example, since among all the groups, the proportion of whom took a milestone image was 25%, then we set the discrete probability distribution (e.g., a Bernoulli distribution) of Milestone Image as $P(x_{milestone\_image} = k) = p^k(1-p)^{1-k}, where\ p = 0.25, k \in \{0, 1\}$. Note that the initial probability will be adjusted as the Bayesian network learns to fit the evidence.

For the other nodes (i.e., the internal nodes representing individual skills and the top-level node representing the overall skill level), their conditional probability distributions were initially specified by a subject matter expert from our research group. The values were agreed upon by other research members in the team to ensure the current model well represents the reality for laboratory-skill inference. This process of determining initial probabilities is also known as informative prior setup, which is a valuable feature of Bayesian models to incorporate individual perspectives and has been widely adopted (Levy, 2016). Specifically, the subject matter expert was assigned to fill in conditional probability tables for each internal and top-level node. In the case of evidence recording node, for example, if a table entry had both taken a milestone image and generated a final graph, the expert assigned a relatively high probability (e.g., 0.8) that the entry had shown a desirable skill level of evidence recording. This entry can be recorded as "1 (whether a milestone image was taken), 1 (whether a final graph was generated), 1

(whether a student showed evidence recording skill), and 0.8 (a probability value to be filled by the expert)".

Then, the Bayesian network could perform reasoning tasks by feeding the nodes with new evidence. In our case, students' skill levels could be inferred via feeding the Bayesian network with extracted lab behaviors and reading the updated posterior probability (i.e., the probability of an event happening after all evidence is taken into account). Then, we used the Bayesian network algorithm to calculate the performance of individuals and inspected the overall pattern across all individuals. The calculated results were presented in the "Results" section.

Next, we discussed representative student pairs based on their average skill performance and performance in each dimension of laboratory skills. These student pairs were selected as they instantiate maximum variation (Flyvbjerg, 2006) in terms of laboratory skills. The selection was also what Flyvbjerg (2006) called an informed-oriented selection: From review of their lab reports and recordings of what students saw through the infrared camera, we expected these student pairs to contain rich examples of scaling and assembling comparisons in laboratory learning. In the "Results" section, we also presented how log analysis of laboratory skills was related to students' conceptual understanding using a Pearson correlation analysis.

Results

Average Performance on Laboratory Skills

Table 1 shows the laboratory skill level across all students in terms of probabilities. An overall value of 71.0% is reported. A higher probability indicates a higher certainty of demonstrating desired laboratory skills. This probability is posterior probability and is calculated after taking into account all skill-related behaviors. In other words, the better a student acted in conformity with lab procedures (according to the observable evidence), the higher the probability reasoned by the model to indicate her or his ability to demonstrate the skill. This may be explained that practical lab abilities such as effectively executing lab procedures can be an important indicator on students' lab skills (see the review of Hofstein & Lunetta, 1982; Hofstein, 2017). Table 1 also lists the overall values of three sub-categories: experiment setup, observation, and evidence recording. The probabilities of 79.3%, 66.7%, and 71.1% indicate a satisfactory level of competency in these skills, where students are more likely to demonstrate these skills than not showing these skills. This table could serve as a reference to compare individual performance with the sample average. As the experiment was conducted by a pair of students, individual data comes from two participants.

Table 1 The average performance on laboratory skills and three sub-categories as reasoned by a Bayesian network in terms of posterior probabilities

Overall skill level	Sub-categories		
	Experiment setup	Observation	Evidence recording
71.0%	79.3%	66.7%	71.1%

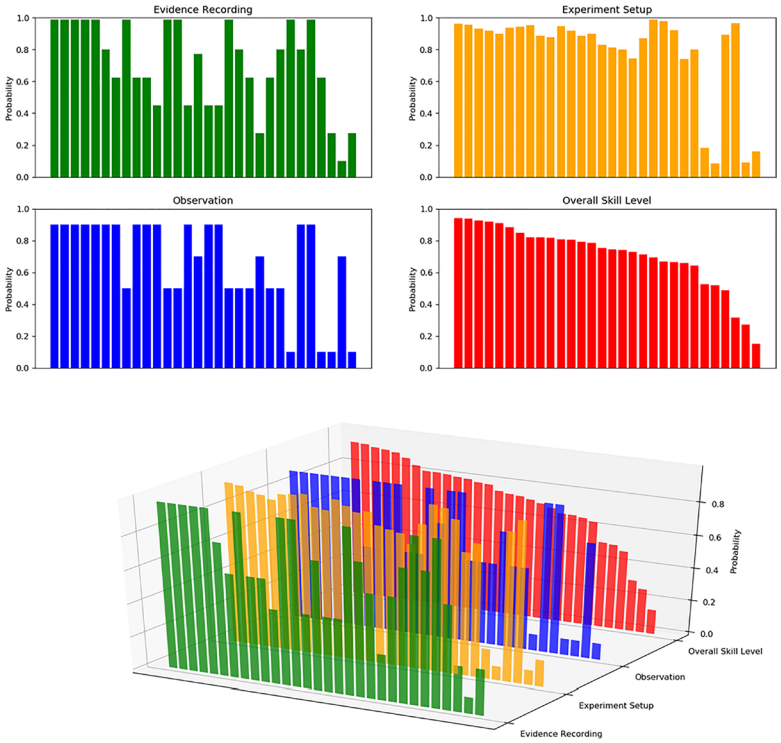

Fig. 4 A comprehensive view of individuals' performance on the overall skill level and the sub-categories

Table 2 A case of deficient experiment setup and its skill levels compared with the sample average

	Overall skill level	Sub-categories		
		Experiment setup	Observation	Evidence recording
Deficient setup	52.5%	18.2%	90.0%	80.0%
Average	71.0%	79.3%	66.7%	71.1%

We also provide a comprehensive view of how individuals performed in each sub-category, together with their overall skill levels in Fig. 4. Each bar in the graphs represents the result derived from one pair of lab partners. In the following sections, we will elaborate on the sub-categories of laboratory skills using representative cases.

Experiment Setup

Table 2 shows a case with a deficient experiment-setup skill. This pair of students performed lower-than-average in terms of overall laboratory skills (i.e., 52.5%) despite the greater-than-average level achieved on observation (90.0%) and evidence recording

(80.0%). The reason why lab performance was underachieved is that the students did not finish the experiment setup. Screenshots during its lab process (Fig. 5) reveal that instead of placing a total of six thermometers on both rulers with three on each, they only positioned three thermometers on one ruler. Besides, the students added a thermometer on the top left corner of the screen to compare the ambient temperature with the thermometer readings on the right ruler. Such a problematic experiment setup due to likely fallacious reasoning would adversely affect the experiment process. The collected evidence has a gap as necessary data was missing for a quantitative comparison between two rulers. Admittedly, the students did keep the camera steady, as well as the position of the thermometers fixed. They also turned on the T(t) graph to observe temperature change over time (right screenshot of Fig. 5) while watching heat conduction for at least 60 s. They recorded milestone images along the observation to complete the lab report. However, observations and the evidence recording based on a defective experiment setup made the collected data a futile effort to achieve expected experiment results.

A deficient experiment setup may happen in various forms, and the Bayesian network can easily handle such discrepancy. Figure 6 shows a worse example captured by the Bayesian network, as indicated by a mere 9.05% probability of competency on the experiment setup. The students deployed an insufficient number of thermometers (two with one on each ruler) and changed the positions of both thermometers during the experiment process. Such behavior would cause inconsistency in temperature readings and hence obscure the intended trend to be observed and reported.

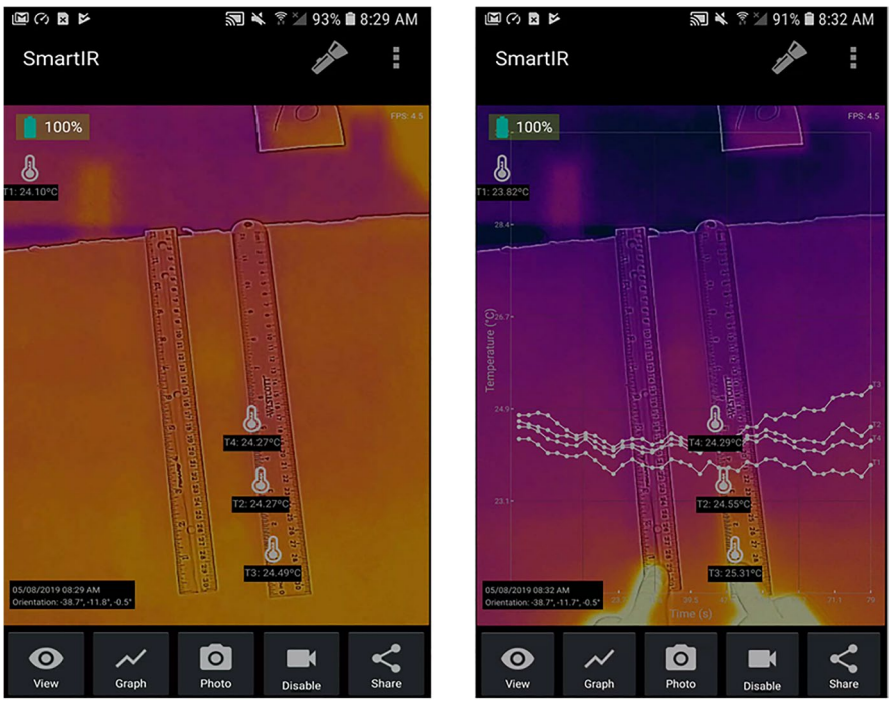

Fig. 5 Screenshots of the experiment setup (left) and the experiment process (right) of a case of deficient experiment setup

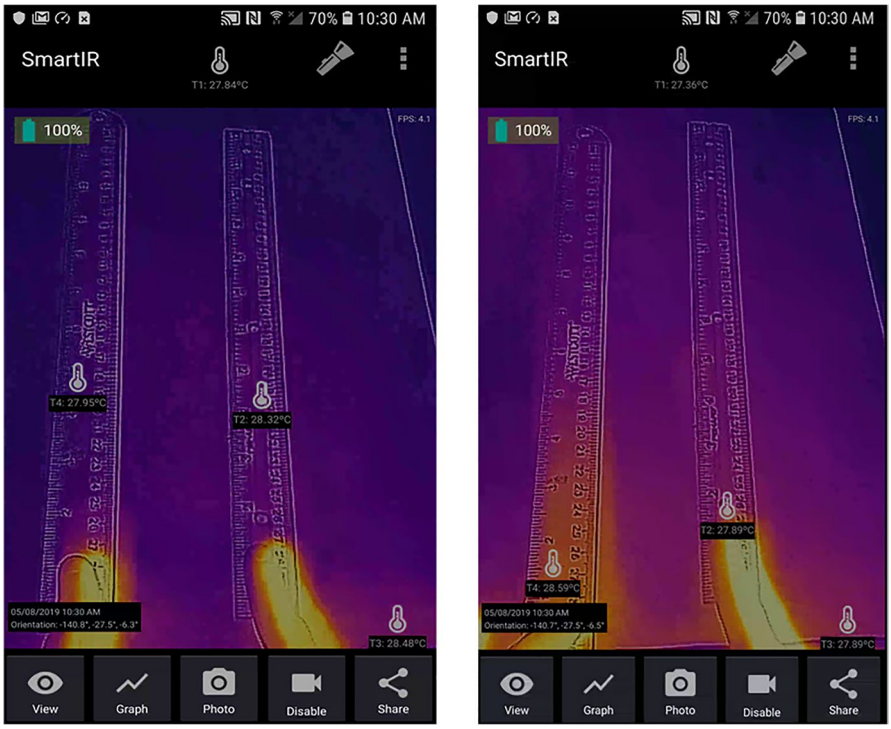

Fig. 6 Screenshots of the initial thermometer position (left) and their change during the experiment (right) of an even worse case of experiment setup

Observation

Table 3 shows the performance of an example who had a lower-than-average performance on observing. The students in this example did a relatively good job in setting up the experiment (87.0%) and recording evidence (80.0%), which indicates competency in utilizing instruments as well as recording required milestone images during the experiment. However, failure in turning on the graph showing temperature trend (Fig. 7 right) would greatly weaken the explanation on how fast and in what patterns heat conducts in different materials, while that explanation is required in the lab report. Their answers in the lab report

Table 3 A case of deficient observation and its skill levels compared with the sample average

	Overall skill level	Sub-categories		
		Experiment setup	Observation	Evidence recording
Deficient observation	71.2%	87.0%	50.0%	80.0%
Average	71.0%	79.3%	66.7%	71.1%

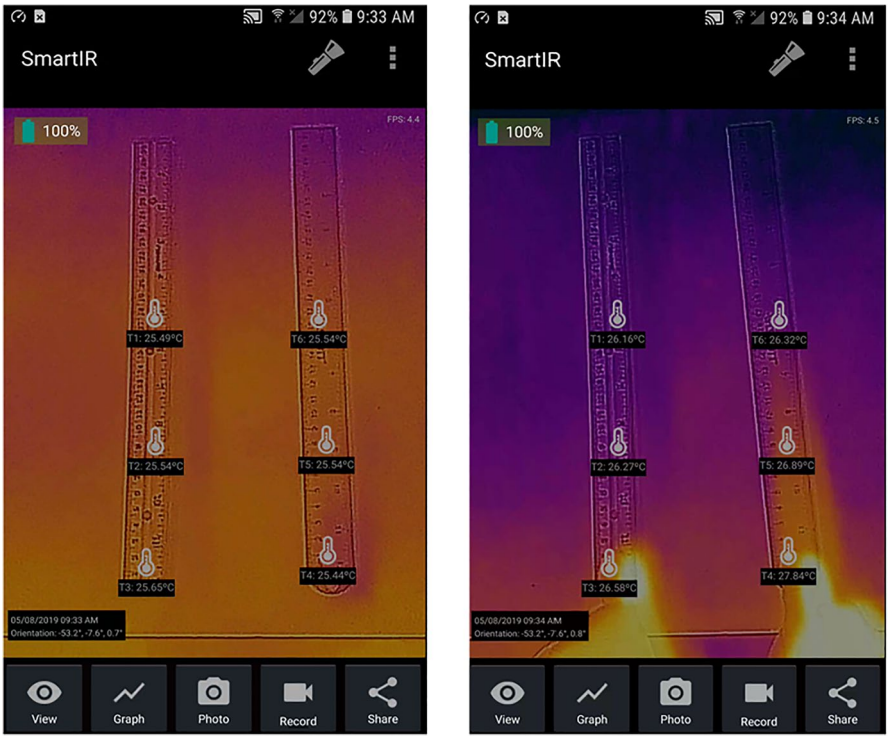

Fig. 7 Screenshots of the experiment setup (left) and the experiment process (right) of an example of deficient observation

confirmed this issue. To predict what would happen to the two rulers after the thumbs touched for 1 min, they did not mention there would be a difference between two materials and anticipated that the same pattern would happen to both rulers: "The part that your thumb is on will become warmer." Such a knowledge gap remains after the observation. They still did not realize the difference and described two rulers as a whole: "The thermometers near the bottom are warmer as that is the part of the ruler we are touching."

Evidence Recording

Table 4 shows an example that performed excellently on evidence recording and other skills. The students set up the experiment well (96.5%), observed both the phenomenon and the temperature trend (90.0%), and performed a superior job in recording required evidence (99.0%). Such maneuvers were manifest in the milestone images extracted from the lab report (Fig. 8). As shown in those images, the students positioned thermometers according to the given diagram (Fig. 8a) and observed how heat conducts along a metal ruler as opposed to a wood ruler (Fig. 8b, c). They took pictures of the two thumbs with the visible pattern right after the experiment (Fig. 8d) and saved the T(t) graph registering the temperature trend over time (Fig. 8e).

Table 4 A case of excellent evidence recording and its skill levels compared with the sample average

	Overall skill level	Sub-categories		
		Experiment setup	Observation	Evidence recording
Excellent evidence recording	94.2%	96.5%	90.0%	99.0%
Average	71.0%	79.3%	66.7%	71.1%

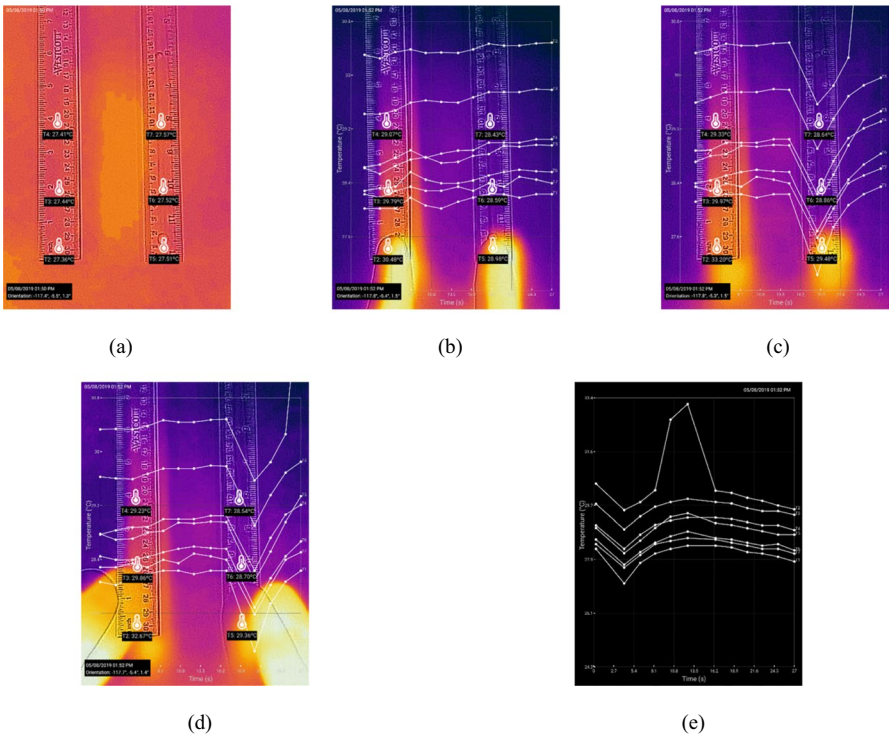

Fig. 8 An example of correctly preparing all required milestone images in the lab report. **a** Initial state. **b** State after 60 s. **c** State after thumbs removal. **d** State of thumbs. **e** Final T(t) graph

Laboratory Skills and Students' Conceptual Understanding

We explored how laboratory skills are related to students' conceptual understanding using a Pearson correlation analysis. The two variables involved are skill performance and conceptual learning gain. Fourteen of the 30 teams completed both the prediction and explanation questions (prediction score $M = 0.65$, $SD = 0.13$; explanation score $M = 0.70$, $SD = 0.18$). There was a lot of missing data from the POE lab report, and many groups only provided prediction data. The lack of explanation data was due to the fact that the POE lab report served as a self-paced guideline for students to navigate

their lab. Many groups did not complete the explanation question because they either ran out of time at the end of the lab or were not as motivated to provide their reasoning on the lab report as interacting with the Infrared Explorer technology. Even with small sample size, results demonstrated a significant positive correlation ($r = 0.582$, $p = 0.029$) between skill performance and learning gain (see Table 5). The analysis was not causal, and the sample size was small, but it implies the potential positive effect of laboratory skills on conceptual understanding: A better laboratory skill may be the reason for an improved conceptual understanding after the lab implementation, while poor skill may contribute to low-level scientific learning.

A closer look at student argument examples provided additional details. The team with the highest skill score (92.5%) improved their understanding of the concept after the lab (learning gain = 0.25). Their prediction on what would happen to two rulers was correct but abstract ("after touching them for 1 min, the metal ruler will lose the heat faster than the wooden one because the wooden one is a better insulator"). After observation, they expanded the explanation with a detailed description of the visual evidence they saw: "The heat was being conducted better in the metal ruler than in the wood ruler, so the metal ruler had the heat farther up on the ruler, and the wood ruler had the heat more centered on the thumb. The distance on the rulers shows the conductivity of each ruler."

As a comparison, the team with the lowest laboratory skills (31.7%) had a negative learning gain (− 0.33) and replaced their correct prediction with misconceptions in the post-lab explanation. They predicted correctly that "the metal will feel colder because your hands will transfer the heat to the metal ruler better." After the erroneous operations, they switched the answer to a popular misconception—believing that metals conduct coldness (rather than heat) better than other objects (e.g., Pathare & Pradhan, 2010). They wrote that "the metal ruler felt cooler to touch because the coolness transfers to your hand faster." Such decreased conceptual understanding might be caused by problematic observation and insufficient evidence recording, as this pair's individual lab skills—experiment setup, observation, and evidence recording—are 96.6%, 10.0%, and 27.5%, respectively. The last two scores are extremely low.

As a reference, we also list the scores of conceptual understandings for the three examples we introduced in previous sections (see Table 6). Because the team with excellent evidence recording did not complete their lab report, their scores are not available. For the rest two examples that have deficient lab operations, one team's scores only reached the group average, and the other had a negative learning gain. Their performance generally fits the trend we found from the correlation analysis.

Table 5 The Pearson correlation between laboratory skills and conceptual learning gain

	Value range	*n*	*M*	*SD*	*r*	R2
Laboratory skills	0 to 1	14	0.75	0.16	0.582*	0.339
Conceptual learning gain	− 1 to 1	14	0.05	0.15		

* indicates $p < 0.05$

Table 6 The conceptual understanding of the three examples introduced in previous sections

	Overall skill level	Sub-categories			Prediction	Explanation	Conceptual learning gain
		Experiment setup	Observation	Evidence recording			
Deficient setup	52.5%	18.2%	90.0%	80.0%	0.78	0.83	0.06
Deficient observation	71.2%	87.0%	50.0%	80.0%	0.78	0.67	−0.11
Excellent evidence recording	94.2%	96.5%	90.0%	99.0%	N/A	N/A	N/A

Discussion and Implications

A unique and significant contribution of this study is that we demonstrated an automated analysis of laboratory skills leveraging learning analytics, specifically, mining smartphone logs using Bayesian network modeling. This study fills the literature gap of automatically assessing hands-on laboratory practices (Beaumont-Walters & Soyibo, 2001). Automated analysis of laboratory skills is a first step of developing and studying intelligent tools for supporting physical laboratory learning, which is one area that is relatively understudied. Findings from this study show the potential capacity that mobile logs have to inform on hands-on laboratory learning.

We analyzed the lab process with a Bayesian network. A Bayesian network is a "white-box" in which the prediction process is explicit, compared with a black-box model in which variables' contribution is unclear (Conati et al., 2002). Therefore, we can mimic the actual lab process in the Bayesian network by defining which actions represent a laboratory skill. This straightforward structure allows future researchers to easily understand the logic and quickly adapt the model to their study. Besides, using a Bayesian network provides an opportunity to develop real-time feedback systems, as it can instantly update results (in the form of probabilities) when being fed new evidence (Fan et al., 2021). As long as the log captures new data, the reasoning process of the Bayesian network is triggered immediately to provide updated posterior probabilities for the nodes representing the variables of interest. For example, in the thermal conduction activity of this study, there should be six virtual thermometers, with three on each ruler. The probability of the correct setup increases when more thermometers are added to the rulers. It peaks when all six thermometers are added during the experiment. But if the addition stops in the middle (e.g., only adding two thermometers), the probability will stay at a lower value. Investigating the learning effect of a Bayesian network-based real-time feedback system for hands-on lab activities is a fruitful area for future exploration.

To further investigate why students do not place the expected number of thermometers, teachers can orchestrate with Bayesian model's insights to intervene with students and qualitatively understand their rationales (e.g., unclear instructions, carelessness, trial-and-error). Bayesian models, including Bayesian network, can be dynamically updated (Marcot & Penman, 2019). The dynamic feature of Bayesian network means that we can rapidly iterate the existing model when more data has been collected to provide potentially more accurate inferences. Moreover, researchers who are interested in modeling students' scientific inquiry with Bayesian network can build their models based on our posterior probabilities, which can be helpful when no informative initial probabilities can be defined (e.g., by subject matter experts) (Levy, 2016).

Moreover, our results describe in detail the skill level of participating middle schoolers. In general, they could implement all desired experiment steps with a probability of 71%. Aligning with results in other studies (e.g., Kapici et al., 2020; Viegas et al., 2018), we found that individuals were diverse in their skill levels. Their probability of performing a skill ranged from as low as 10% to as high as 100%. Furthermore, some skills (e.g., proper observation) had more low-performers than the others (e.g., many students forgot to turn on the T(t) graph to observe the temperature change over time). Such findings imply a growing need for personalized intervention in laboratory learning, such as notifying these students of turning on the T(t) function during the observation stage. While the focus of this study is not using the Bayesian network to provide feedback to students, we expect that these customized interventions potentially can be provided by a teacher in no time using the feedback

from the Bayesian network. Potentially, such a feedback system can relieve teachers from the labor-intensive work of simultaneously monitoring a large number of students in the class, and grants them more time to focus on the task of implementing tailored instruction to address differential learning demands. However, automated feedback systems might create instructional challenges for teachers when they perceive feedback as misaligned to instruction (Wilson et al., 2021). Future studies should pay close attention to challenges and opportunities in bringing real-time feedback systems into laboratory teaching and learning.

Besides, we found that a well-performed lab was related to a gain of conceptual understanding, while a low performance was associated with a loss of understanding. This finding challenges the results of several other studies (e.g., Colorado DOHE, 2012), which showed that physics labs had limited influence on students' conceptual learning. In this study, we assessed lab skills and learning performance for teams, not individual students in each team. A fruitful area for future exploration is using learning analytics to mine lab skills for individual students in collaborative learning environments and investigating the impact of group dynamics in shaping the development of lab skills and conceptual learning.

Furthermore, we found that some low-skilled performers exhibited misconceptions that possibly came from observational errors. This finding again emphasizes the need to help the students with low laboratory skills, as such low performance may lead to worse conceptual learning. For instance, insufficient observation time may lead to an incomplete observation of the thermal process and thus lead to a problematic explanation. To meet such demands, an automatic approach supported by educational technology has unique advantages. It can monitor a large population simultaneously and promptly identify targets. Thus, it would be worthwhile to further investigate explicit and direct connections between lab performance and conceptual understanding towards supporting effective laboratory learning at scale.

In summary, the present study provides preliminary evidence for the feasibility of modeling students' performance in laboratory learning with a Bayesian network. To the best of our knowledge, this is the first study to automatically analyze and assess the students' laboratory performances in physical settings. We hope this study could set an exciting first step for developing and studying automated analysis tools for hands-on laboratory skills.

Limitations

There are limitations to this study. First, the findings for the second research question about the relationship between laboratory skills and learning performance can be biased due to the fact that many students did not have time to answer questions in the explanation phase. The interpretation of the positive correlation between laboratory skills and learning performance should be contextualized within the limitation of the small sample and the sample of students who had time or were willing to finish questions in the explanation phase. Second, although smartphone log data is enough to reconstruct the lab process, additional data can provide rich details about student learning. Future studies could apply computer vision to analyze the content of images and natural language processing techniques to mine student communications for enriching the evidence for understanding laboratory practices. Third, this study utilized specific devices to realize the logging in the physical space, such as various phone sensors and the plug-in IR camera. In the future, more phone sensors and plug-in devices (e.g., detecting students' gestures using Kinect sensors) could be integrated into science classrooms and harness such power to help both instruction and research. Meanwhile, we should be aware of ethics and bias issues in capturing data through sensors

and building models to automate the analysis of laboratory skills. Overall, more research utilizing the logs from mobile devices and sensors are needed to identify meaningful patterns in secondary laboratory learning.

Funding This work was supported by the National Science Foundation (NSF) grants DRL-2054079. However, any opinions, findings, conclusions, or recommendations are our own and do not reflect the views of the NSF.

Declarations

Conflict of Interest The authors declare no competing interests.

References

BayesFusion (2017). QGeNIe Modeler. *User Manual. Retrieved July 6, 2022, from* https://support.bayesfusion.com/docs/

Beaumont-Walters, Y., & Soyibo, K. (2001). An analysis of high school students' performance on five integrated science process skills. *Research in Science & Technological Education, 19*(2), 133–145.

Chabalengula, V. M., Mumba, F., Hunter, W. F., & Wilson, E. (2009). A model for assessing students' science process skills during science lab work. *Problems of Education in the 21st Century, 11*, 28–36.

Chang, H. Y., Lin, T. J., Lee, M. H., Lee, S. W. Y., Lin, T. C., Tan, A. L., & Tsai, C. C. (2020). A systematic review of trends and findings in research employing drawing assessment in science education. *Studies in Science Education, 56*(1), 77–110.

Chiang, T. H., Yang, S. J., & Hwang, G. J. (2014). Students' online interactive patterns in augmented reality-based inquiry activities. *Computers & Education, 78*, 97–108.

Colorado DOHE (Department of Higher Education). (2012). *Online versus traditional learning: A comparison study of Colorado community college science classes*. Retrieved July 6, 2022, from https://wcet.wiche.edu/resources/online-versus-traditional-learning-a-comparison-study-of-colorado-community-college-science-classes/

Conati, C., Gertner, A., & Vanlehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction, 12*(4), 371–417.

Doran, R. L., Boorman, J., Chan, F., & Hejaily, N. (1993). Alternative assessment of high school laboratory skills. *Journal of Research in Science Teaching, 30*(9), 1121–1131.

Fan, Y., Zhang, J., Zu, D., & Zhang, H. (2021). An Automatic Optimal Course Recommendation Method for Online Math Education Platforms Based on Bayesian Model. *International Journal of Emerging Technologies in Learning (iJET), 16*(13), 95–107.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Research Practice*, 390–404. https://doi.org/10.4135/9781848608191.d33

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education, 31*(1), 1–23.

Gobaw, G. F., & Atagana, H. I. (2016). Assessing laboratory skills performance in undergraduate biology students. *Academic Journal of Interdisciplinary Studies, 5*(3), 113.

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. *Journal of the Learning Sciences, 22*(4), 521–563.

Gunstone, R. F. (1990). Children's science: A decade of developments in constructivist views of science teaching and learning. *The Australian Science Teachers Journal, 36*(4), 9–19.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281–307.

Hensiek, S., DeKorver, B. K., Harwood, C. J., Fish, J., O'Shea, K., & Towns, M. (2016). Improving and assessing student hands-on laboratory skills through digital badging. *Journal of Chemical Education, 93*(11), 1847–1854.

Hofstein, A. (2017). The role of laboratory in science teaching and learning. *Science Education*, 357–368. https://doi.org/10.1007/978-94-6300-749-8_26

Hofstein, A., & Lunetta, V. N. (1982). The role of the laboratory in science teaching: Neglected aspects of research. *Review of Educational Research, 52*(2), 201–217.

Hofstein, A., & Lunetta, V. N. (2003). The laboratory in science education: Foundations for the twenty-first century. *Science Education, 88*(1), 28–54.

Hunt, L., Koenders, A., & Gynnild, V. (2012). Assessing practical laboratory skills in undergraduate molecular biology courses. *Assessment & Evaluation in Higher Education, 37*(7), 861–874.

Jiang, S., Tatar, C., Huang, X., Sung, S. H., & Xie, C. (2022). Augmented Reality in Science Laboratories: Investigating High School Students' Navigation Patterns and Their Effects on Learning Performance. *Journal of Educational Computing Research, 60*(3), 777–803.

Jou, M., & Wang, J. (2013). Ubiquitous tutoring in laboratories based on wireless sensor networks. *Computers in Human Behavior, 29*(2), 439–444.

Kapici, H. O., Akcay, H., & de Jong, T. (2020). How do different laboratory environments influence students' attitudes toward science courses and laboratories? *Journal of Research on Technology in Education, 52*(4), 534–549.

Leggett, M., Kinnear, A., Boyce, M., & Bennett, I. (2004). Student and staff perceptions of the importance of generic skills in science. *Higher Education Research & Development, 23*(3), 295–312.

Levy, R. (2016). Advances in Bayesian modeling in educational research. *Educational Psychologist, 51*(3–4), 368–380.

Li, H., Gobert, J., Graesser, A., & Dickler, R. (2018). Advanced educational technology for science inquiry assessment. *Policy Insights from the Behavioral and Brain Sciences, 5*(2), 171–178.

Lunetta, V. N., Hofstein, A., & Clough, M. P. (2007). Learning and teaching in the school science laboratory: An analysis of research, theory, and practice. *Handbook of Research on Science Education, 2*, 393–441.

Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental Modelling & Software, 111*, 386–393.

Minalisa, M. (2019, April). The development of performance assessment of inquiry-based learning (IBL) to improve student's science process skill of class XI Senior High School 1 Bayang. In Ramli, Yohandri, Festiyed, Wurster, R. Jaafar, S. A. Bakar (Eds.), *Journal of Physics: Conference Series, 1185*(1), 012134. IOP Publishing.

National Science Teachers Association. (2007). *NSTA position statement: The integral role of laboratory investigations in science instruction*. Retrieved July 6, 2022, from https://www.nsta.org/about/positions/laboratory.aspx

NRC. (2006). *America's lab report: Investigations in high school science*. The National Academies Press.

Park, Y., & Jo, I. H. (2017). Using log variables in a learning management system to evaluate learning activity using the lens of activity theory. *Assessment & Evaluation in Higher Education, 42*(4), 531–547.

Pathare, S. R., & Pradhan, H. C. (2010). Students' misconceptions about heat transfer mechanisms and elementary kinetic theory. *Physics Education, 45*(6), 629.

Potkonjak, V., Gardner, M., Callaghan, V., Mattila, P., Guetl, C., Petrović, V. M., & Jovanović, K. (2016). Virtual laboratories for education in science, technology, and engineering: A review. *Computers & Education, 95*, 309–327.

Prichard, E. (2003). *Practical Laboratory Skills Training Guides (Complete Set)*. The Royal Society of Chemistry.

Sung, S. H., Li, C., Chen, G., Huang, X., Xie, C., Massicotte, J., & Shen, J. (2021). How does augmented observation facilitate multimodal representational thinking? Applying deep learning to decode complex student construct. *Journal of Science Education and Technology, 30*(2), 210–226.

Tadlaoui, M. A., Aammou, S., Khaldi, M., & Carvalho, R. N. (2016). Learner modeling in adaptive educational systems: A comparative study. *International Journal of Modern Education and Computer Science, 8*(3), 1.

Viegas, C., Pavani, A., Lima, N., Marques, A., Pozzo, I., Dobboletta, E., ... & Lima, D. (2018). Impact of a remote lab on teaching practices and student learning. *Computers & Education*, *126*, 201–216.

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education, 168*, 104208.

Xie, C. (2011). Visualizing chemistry with infrared imaging. *Journal of Chemical Education, 88*(7), 881–885.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151.
Zhou, Y., Fenton, N., & Neil, M. (2014). Bayesian network approach to multinomial parameter learning using data and expert judgments. *International Journal of Approximate Reasoning, 55*(5), 1252–1268.