Automating the Assessment of Problem-solving Practices Using Log Data and Data Mining Techniques

Karen D. Wang
Graduate School of Education
Stanford University
kdwang@stanford.edu

Shima Salehi
Graduate School of Education
Stanford University
salehi@stanford.edu

Max Arseneault
Stanford University
marsenea@stanford.edu

Krishnan Nair
Stanford University
aknair@stanford.edu

Carl Wieman
Department of Physics
Graduate School of Education
Stanford University
cwieman@stanford.edu

ABSTRACT
Interactive simulations provide an exciting opportunity to assess and teach students the practices used by scientists and engineers to solve real-world problems. This study examines how the logged interaction data from a simulation-based task could be used to automate the assessment of complex problem-solving practices. A total of 73 college students worked on an interactive circuit puzzle embedded in a science simulation in an interview setting. Their problem-solving processes were videotaped and logged in the backend of the simulation. We extracted different sets of features from the log data and evaluated their effectiveness as predictors of students' problem-solving success and evidence for specific problem-solving practices. Our results indicate that the application of data mining techniques guided by knowledge gained from qualitative observation was instrumental in the discovery of semantically meaningful features from the raw log data. These knowledge-grounded features were significant predictors of students' overall problem-solving success and provided evidence on how well they adopted specific problem-solving practices, including decomposition, data collection, and data recording. The results point to promising directions for how scaffolding/feedback could be provided in educational simulations to enhance student learning in problem-solving skills.

Author Keywords
science simulation, log data, problem solving, assessment, educational data mining

L@S '21, June 22–25, 2021, Virtual Event, Germany.

ACM ISBN 978-1-4503-8215-1/21/06 ...$15.00.
http://dx.doi.org/10.1145/3430895.3460127

CCS Concepts
•**Applied computing** → **Interactive learning environments;**
•**Information systems** → **Data mining;**

INTRODUCTION
Advances in artificial intelligence are reshaping the landscape of the human workforce. Routine tasks in manufacturing, transportation, and information processing have been increasingly replaced with technology, while the ability to solve unstructured, novel problems remains a uniquely human endeavor with growing demand [14]. An engineer needs to troubleshoot a faulty robotic arm. A doctor needs to diagnose a patient with complex clinical presentations. Preparing students to solve such novel problems in workplaces and everyday lives is an important goal of education. Accordingly, there is increasing recognition that problem-solving skills should be explicitly taught and measured in our schools [19]. The National Research Council (NRC) puts practices related to problem solving at the core of the Next Generation Science Standards [22]. The Accreditation Board for Engineering and Technology (ABET) lists "an ability to identify, formulate, and solve complex engineering problems" as a primary student outcome to consider when accrediting engineering programs [7]. In contrast to this clear vision, there is less consensus around how to measure and teach these practices. Interactive simulations provide an exciting opportunity to capture how well students learn the problem-solving practices using the wealth of logged interaction data. This study explores how data mining techniques could be applied to such data sets in order to evaluate student problem-solving practices as well as identify opportunities for adaptive support/feedback. Specifically, two research questions are addressed here:

- What features can be derived from the logged interaction data of students' problem-solving process?
- How useful are these features in predicting problem-solving success/failure and evaluating specific problem-solving practices?

RELATED WORK

Educational technology allows for assessment of learning beyond what is possible with traditional paper-and-pencil tests [5, 6, 20, 25]. One key affordance is the large amounts of data generated in computerized learning environments [24]. These data may consist of thousands of learners' responses in a large-scale computerized adaptive test, or a few students' detailed interaction log data captured in an interactive learning task. Researchers in the learning analytics and educational data mining communities have begun to capitalize on large data sets and advanced statistical models to assess students' knowledge state and predict their performance [4, 11, 16, 18]. While deep learning models have reached high accuracy in predicting student performance in multiple-choice test items [18], the same promise is yet to be fully realized for the assessments of complex skills such as problem-solving and scientific inquiry. How to parse the large volumes of unstructured interaction data to reveal meaningful evidence of student learning behaviors remains a challenge [6, 9]. Our research takes a step in addressing this challenge by exploring how data mining techniques could be used to process the log data of college students working through an interactive problem and yield insights about the practices adopted to solve the problem.

There has been an increasing number of research efforts that use log data to analyze student performance on problem-solving/scientific inquiry tasks [1, 2, 3, 10, 9, 12, 16, 23]. Techniques employed to process log data include text replay tagging [9], event sequencing [12, 16], and cluster analysis [1]. Gobert et al. (2013) [9] utilized human-labeled log data to train a detector for assessing the skill of designing controlled experiments in two simulation-based inquiry tasks. Kinnebrew et al. (2013) [12] employed the sequence mining technique to identify differentially frequent action patterns between high- and low-performing students in a computerized learning environment featuring a teachable agent. Teig et al. (2020) [23] incorporated metrics from the log data into latent class analyses to identify three distinct profiles of students' inquiry performance in PISA science assessment items. It is important to note that previous studies have largely focused on log data features related to the frequencies and durations of individual actions, while few studies explored whether more semantically meaningful features could be extracted from the log data by leveraging knowledge gained from qualitative observations of student performance. Meanwhile, the application of machine learning models in educational data mining to assess complex skills and practices also comes with challenges. First, obtaining a sufficiently large data set of students working on an open-ended task to train deep neural networks that extract features from raw data automatically is difficult. Zhai et al. (2020) [29] conducted a systematic review of machine learning-based science assessment and found that the largest training sample included 2978 students. This is in stark contrast to the massive amounts of data readily available in fields such as computer vision. Second, for machine learning models to serve as the basis for improving student learning outcomes, the interpretability of the model is as important as its predictive accuracy to be able to identify when and how to intervene to improve student learning [8]. Taking the above challenges into account, our goal in this study is to explore how to extract semantically meaningful features from the log data and use them in simple, robust regression models to measure problem solving.

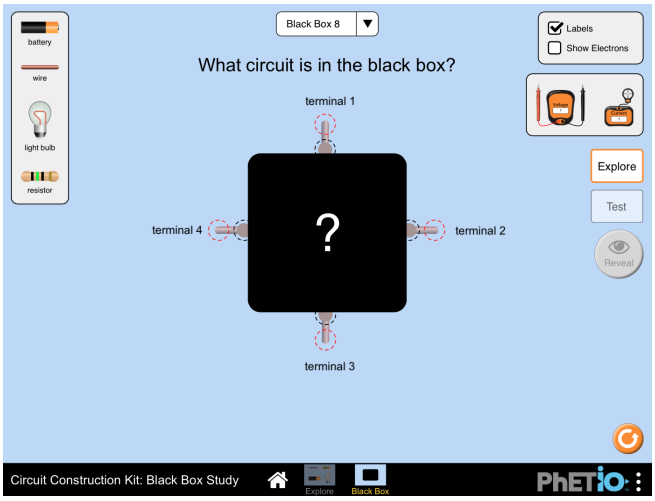

Figure 1. The black box problem user interface with the terminal labels added ©PhET Interactive Simulations

The Black Box Problem

The black box problem embedded in the Circuit Construction Kit (CCK) of PhET Interactive Simulations (https://phet.colorado.edu/) was used to study how college students solve problems in science and engineering domains [17]. The simulation provides an accurate model based on Ohm's law for every circuit constructed. The goal is to infer the circuit hidden behind a black box by connecting electrical components (wire, battery, resistor, lightbulb) and measurement tools (ammeter and voltmeter) to the wires ("terminals") protruding from the black box and interpreting the data obtained (Figure 1). Solving the hidden circuit requires basic knowledge of electric circuits and Ohm's law, yet the problem is significantly different from typical problems used to assess students on this topic. Typical end-of-chapter textbook problems provide all the quantities needed to solve the problem in the question and are conducive to the "plug and chug" approach [13], where students could simply plug the numbers in a formula to arrive at the correct solution. In contrast, the black box problem requires students to make decisions regarding what data to collect, how to collect the data, and how to interpret the data to reach a solution. Therefore, we can collect a much richer data set of practices that are relevant to solving a wide variety of real-world problems.

In a previous study from our lab, we employed qualitative methods to analyze video recordings of experts and novices solving the black box problem and identified two groups of practices that are key to effective problem solving: execution practices and reflection practices [21]. Execution practices encompass the actions taken by problem solvers when working toward a solution, including how they define the problem, decompose the problem, collect, record and interpret data in the task environment. Reflection practices describe how people

reflect on their solution process, including how they reflect on problem definition and assumptions, reflect on domain knowledge, reflect on strategies used to solve the problem, and reflect on a proposed solution. Many of these practices have been discussed under different labels (e.g., understand the problem, seek evidence, check and look back) by previous works examining problem solving in mathematics, science and engineering domains [15, 19, 26, 27]. Our earlier work includes a coding scheme that was iteratively developed to evaluate participants' effectiveness in using each of these practices based on video recordings of their problem-solving processes. We found that participants' problem-solving practice scores could significantly predict the probability of them correctly solving the problem [21].

In the current study, we first used the coding scheme to score a subset of the practices (problem definition, decomposition, data collection, data recording, and reflection on solution) using the video recordings and notes taken by students during the problem-solving process. Problem definition refers to the practice of defining a problem in one's own words to ensure that the goal of the problem and its underlying assumptions are well-understood. Decomposition is defined as the practice of breaking a complex problem into simpler sub-problems that are easier to solve. Data collection refers to the practice of systematically and effectively collecting the information/data needed to solve the problem. Data recording refers to the practice of keeping track of the data collected during the problem-solving process. Reflection on solution refers to the practice of checking one's solution in different ways to verify its accuracy. The human-rated scores of these practices provide an important benchmark for evaluating whether an automated scoring approach using the log data could effectively measure problem solving as a multidimensional construct.

METHODS

Participants

73 undergraduate students in the US participated in the study in a one-on-one interview setting with one of the co-authors of the paper. Participants worked on a computer throughout the interview and were provided with a calculator, pen and paper for calculations and notetaking. After informed consent, the researcher gave a brief tutorial to help participants navigate different features of the simulation and refresh their knowledge about Ohm's law (V = I*R) by instructing them to build a series circuit using different electrical components and take measurements using the ammeter and voltmeter. Participants were then given 15 minutes to solve the first black box problem and instructed to think out loud while solving it. The researchers interfered minimally during participants' problem-solving process, doing so only to remind them to think aloud or that they were running out of time. Participants drew a diagram of what they thought was hidden behind the black box on paper when they reached a solution or at the end of the 15 minutes. The full study involved an intervention followed by additional problem-solving tasks but we will be focusing on participants' baseline performance in solving the first black box problem in this paper. All study sessions were video- and audio-recorded. Additionally, participants' actions were logged in the backend of the simulation platform in the Javascript Object Notation (JSON) file format.

Measurements

Students' problem-solving outcomes were measured by a score assigned to their solutions. The score evaluates a student's proposed circuit in three dimensions: whether it contains 1) the correct electrical components (e.g., two resistors), 2) the correct values of the components (e.g., 20 ohms), and 3) the correct structure, or how the components were connected. Each dimension receives a score of zero for incorrect, one for partially correct and two for correct. The total solution score thus ranges from zero to six. Students' problem-solving practices were scored based on the video recordings and notes taken by students during the problem-solving process. Five practices, including problem definition, decomposition, data collection, data recording, and reflection on solution, were scored using a previously developed coding scheme [21]. Scores for each of these practices range from zero to three, reflecting four levels of increasing effectiveness in students' adoption of the practice. Two researchers coded 25% of the videos independently and reached an inter-rater reliability of 80%, and all discrepancies in scoring were resolved through discussion. These scores required approximately 45 minutes to an hour of human coding for each recorded problem-solving session and provided the key benchmark for comparison with the log data analysis of problem-solving practices.

Log Data Processing

The log files contain a vast amount of interaction data as students work to solve the problem, including user-initiated actions (e.g., adding a battery to the black box) and states of the simulation (e.g., how the components are connected), along with the respective timestamps in milliseconds. We wrote a Python script to parse the log data into discrete actions taken by problem solvers and counted the frequency of electric components added and measurement tools usage as the first set of features (Table 1). These action-based features could be obtained without any prior knowledge about the nature of the problem-solving task or the practices that we attempt to measure.

Feature	Description
Wire	How many times did a student add a wire?
Lightbulb	How many times did a student add a lightbulb?
Resistor	How many times did a student add a resistor?
Battery	How many times did a student add a battery?
Voltmeter	How many times did a student use Voltmeter?
Ammeter	How many times did a student use Ammeter?

Table 1. Action-based features extracted from the log data

Through qualitative observation of students' problem-solving process, we noted that the type of circuits built by students were closely related to the practices used to solve the problem. This knowledge led us to extract a second set of knowledge-grounded features by parsing the log data into distinct episodes in which a problem solver engages in one of the following behavioral patterns: constructing circuits, taking measurements,

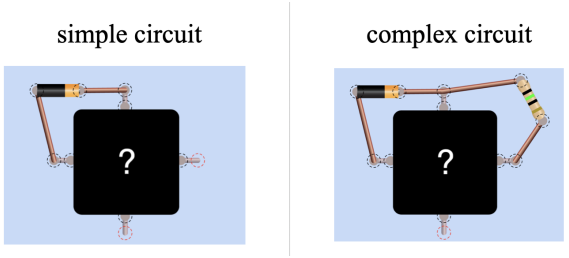

Figure 2. Examples of the two types of circuits built by students. Simple circuits connect two terminals of the black box at a time, complex circuits connect more than two terminals.

or pausing as indicated by inactivity (Table 2). These episodes may potentially reveal key evidence of the problem-solving practices adopted by students as they attempted to collect data, make sense of the data collected and/or reflect on their problem-solving process. All but two students built some types of circuits when solving the problem. Furthermore, most of the circuits built by problem solvers fell into two categories: simple circuits that connect two terminals of the black box, and complex circuits that connect more than two terminals at a time (Figure 2).

The action-based and knowledge-grounded features from the log data were used as two sets of input variables in multiple linear regression models to predict both the solution scores and the scores of specific problem-solving practices of individual students. An additional input variable used in both models was problem solvers' most advanced physics course taken as a proxy for their physics background knowledge ("Background"), which had three levels: high school introductory, high school Advanced Placement (AP), and college physics. Best subset selection was performed to identify the best-fitting models for solution score prediction. Comparison of model performance using these two sets of features allows us to assess the value of knowledge-grounded features.

RESULTS

The results are reported in the following order. First, we provide a brief summary of the descriptive statistics for the features from log data, students' solution scores, and scores of specific problem-solving practices. Next, we present the results of multiple linear regression models and logistic regression models using action-based and knowledge-grounded features to predict students' problem-solving outcomes as measured by the solution scores. Lastly, we present the results of multiple linear regression models using knowledge-grounded features to predict the scores of problem-solving practices as coded by human researchers.

Table 3 presents the descriptive statistics. The black box was a challenging problem for college students in our sample. Students' solution scores ranged from 0 to 6, with a mean of 2.59 and a standard deviation of 1.85. There was also considerable variance in students' scores of specific problem-solving practices. Overall, students were more effective at decomposition and less effective at reflecting on their solutions and verifying its accuracy.

Finding 1. Knowledge-grounded features are better predictors of solution scores than action-based features.
Table 4 & 5 present the results of the best-fitting linear regression models using action-based and knowledge-grounded features to predict solution scores. All variables except for the physics background were scaled to have zero mean and unit variance. There is a considerable difference in the model performance depending on which set of features is used. The model using knowledge-grounded features has an adj. R-squared of 0.60 ($F(5, 67) = 22.22$, $p < 0.001$), while the model using action-based features has an adj. R-squared of 0.16 ($F(3, 69) = 5.60$, $p = 0.002$). Furthermore, in the model using action-based features, none of the action-based features is a significant predictor of the solution score, and the only significant predictor is students' physics background. In contrast, in the model using knowledge-grounded features, a subset of the features are significant or marginally significant predictors of the solution score. Addition of students' physics background to this set of predictors does not improve the model fit (adj. R-squared without background = 0.60, adj. R-squared with background = 0.61, $F(2, 65) = 2.21$, $p = 0.12$). All these results demonstrate that the knowledge-grounded features are much better predictors of students' problem-solving outcomes than the action-based ones.

The knowledge-based features also have more pedagogical value, in that they can provide useful feedback to both students and teachers on what to do to improve problem-solving performance. A subset of the circuit-based features are significantly or marginally significantly associated with solution scores as indicated by the p-values of individual predictors (Table 5). Building simple circuits to connect pairs of terminals, obtaining non-zero ammeter readings, and pausing after using the ammeter are all associated with an increase in solution scores. On the other hand, building complex circuits and obtaining a reading of zero on the ammeter are both associated with a decrease in solution scores. Solving for the hidden circuit requires collecting relevant and interpretable data from the simulation, which can be best achieved through the construction of simple, rather than complex, circuits connecting pairs of terminals from the black box at a time. Proper use of the ammeter to measure currents in these circuits would result in non-zero readings, while most of the "zero" readings are results of using the ammeter on incomplete circuits and provide no useful information. Moreover, the collected data must be noted and recorded for effective interpretation, which requires pausing after taking measurements.

Finding 2. Knowledge-grounded features are effective at identifying low-performing problem solvers.
A robust model for predicting problem-solving outcomes would lay the foundation for providing timely and personalized feedback to improve student performance. It is especially important for the model to identify students who are at risk of failing to solve the problem. To this end, we fitted a multiple logistic regression model to predict the probability of receiving a low solution score using the knowledge-grounded features selected from the linear regression model. Problem solvers in the low-performing group (48% of all participants) have

Feature	Description
No. of circuits	How many circuits were built during the problem-solving process?
Pairs of terminals connected	How many pairs of terminals were connected by simple circuits?
Percentage complex	What percentage of circuits built were complex circuits?
Voltmeter (zero reading)	How many readings of zero were obtained using the Voltmeter?
Voltmeter (non-zero reading)	How many non-zero readings were obtained using the Voltmeter?
Ammeter (zero reading)	How many readings of zero were obtained using the Ammeter?
Ammeter (non-zero reading)	How many non-zero readings were obtained using the Ammeter?
Pause time post circuit construction (pause - circuit)	What was the average pause time after a circuit was built?
Pause time post Voltmeter (pause - V)	What was the average pause time after the Voltmeter was used?
Pause time post Ammeter (pause - A)	What was the average pause time after the Ammeter was used?

Table 2. Knowledge-grounded features extracted from the log data

a solution score of two or below, which indicates that they had little grasp of effective problem-solving strategies and made very limited progress in solving the black box. Problem solvers in the non-low-performing group have a solution score of three or above.

The performance of the classifier is evaluated by the test accuracy rate, Cohen's kappa, and confusion matrix on the test set averaged across 20 random 60-40 train-test splits. The model has a test accuracy rate of 81% (95% CI: 78% - 84%) and a test Cohen's kappa of 0.62 (95% CI: 0.55 – 0.68), a value considered as "substantial" (range: 0.61 - 0.80) for machine-human scoring agreements in the application of machine learning for science assessment [28]. Table 6 presents the confusion matrix for the test set as a measure of the model's class-specific performance. Of the 13 students who indeed received a low solution score, the model on average correctly identified 10.35, or 80% of the group. Of the 16 students who were not in the low solution score group, the model on average correctly identified 13.15, or 82% of the group. The results suggest that the model is effective at differentiating students' problem-solving outcomes using knowledge-grounded features derived from the log data.

Finding 3. Knowledge-grounded features provide important evidence for evaluating problem-solving practices. Our final analysis evaluates how the knowledge-grounded features correspond to the adoption of specific problem-solving practices, including problem definition, decomposition, data collection, data recording and reflection on solution. A strong correspondence between features from the log data and scores of specific problem-solving practices would make it possible to automate the assessment of these practices. More importantly, it would help teachers/students understand how to achieve better problem-solving performance by focusing their attention on specific practices that they need to improve the most.

Linear regression models were fitted to the human-rated practice scores using all knowledge-grounded features from Table 2 as predictor variables. The best subset selection method was used to identify best-fitting models. The models explained a large portion of the variances in the decomposition (adjusted R-squared = 0.77, $F(1, 71) = 240$, $p < 0.001$), data collection (adjusted R-squared = 0.64, $F(2, 70) = 64.27$, $p < 0.001$), and data recording scores (adjusted R-squared = 0.58, $F(5, 67) = 21.14$, $p < 0.001$). On the other hand, the models accounted for a smaller portion of the variances in the scores of problem definition and reflection on solution (adjusted R-squared = 0.30 for both). We conjecture that the 30% of variances explained in problem definition and reflection on solution are tied to a general construct of problem-solving effectiveness rather than constructs that are specific to the respective practices. The lack of strong associations between knowledge-based features and the practice of problem definition and reflection on solution is not surprising, as both of these practices were scored by human coders based primarily on students' think-aloud utterances during the problem-solving process, for instance, "what is this problem asking," or "does this answer make sense?"

Table 7 presents the best-fitting models for predicting the scores of three problem-solving practices: decomposition, data collection, and data recording. Scores of these practices were associated with different subsets of the knowledge-grounded features. In the following section, we will discuss what features were significantly associated with each of the practices and how these associations were in line with our qualitative observation.

The percentage of complex circuits is a significant predictor of decomposition practice scores in the regression model: a one unit increase in the percentage of complex circuit built is associated with a 0.88 unit drop in the decomposition score. In solving the black box problem, how well one adopts decomposition is closely related to the type of circuits built. Building simple circuits that connect two terminals at a time would allow for a section of the hidden circuit to be modularized and inferred. In contrast, building complex circuits that connect more than two terminals signals that a problem solver attempts to solve the whole problem at once without breaking it into small, manageable parts.

Two knowledge-based features, the number of pairs of terminals connected by simple circuits and the percentage of complex circuits, are associated with the data collection practice scores. A one unit increase in the pairs of terminals connected is associated with a 0.52 unit increase in the data collection score. On the other hand, for each unit increase in complex circuits built, the data collection score would go down by 0.4 unit. Based on our observation, data collection for solving the black box includes building circuits followed by using

Variable	Unit	Mean	SD	Min	Max
Action-based features					
Wire	count	9	8	0	43
Lightbulb	count	2	2	0	12
Resistor	count	1.5	2	0	7
Battery	count	2	2	0	10
Voltmeter (V)	count	23	26	0	169
Ammeter (A)	count	18	19	0	100
Knowledge-grounded features					
No. of circuits	count	17	11	0	45
Pairs connected	count	4	2	0	6
Complex circuits	%	27	34	0	92
V - zero	count	13	20	0	154
V - non-zero	count	10	11	0	39
A - zero	count	6.5	8	0	46
A - non-zero	count	11	15	0	85
Pause - circuit	sec	11	9	0	65
Pause - V	sec	10	10	0	55
Pause - A	sec	12	10	0	62
Problem-solving Practice Scores					
Problem definition		1.17	0.53	0	2.22
Decomposition		2.11	1.06	0	3
Data collection		1.27	0.93	0	3
Data recording		1.43	1.08	0	3
Reflection on solution		0.78	0.90	0	3
Solution score		2.59	1.85	0	6

Table 3. Descriptive statistics for features from the log data and outcome variables

the voltmeter/ammeter for voltage/current readings, and to a lesser extent, clicking on a battery to check its voltage. We have also observed students with limited physics knowledge relying on the brightness of a lightbulb to gather data. In order to collect all the data needed to solve the problem, at least six simple circuits need to be built, one across each pair of the four terminals. The results indicate that for data collection to be effective, the type of circuits built is more important than the usage of voltmeter/ammeter alone.

Scores of the data recording practice are positively related to the number of pairs of terminals connected by simple circuits, as the building of simple circuits allows for effective data collection. Data recording scores were also related to counts of ammeter readings: additional non-zero ammeter readings were associated with an increase in data recording scores, while additional zero ammeter readings were associated with a decrease. More interestingly, data recording scores were positively related to the pause time post ammeter and voltmeter usage. Students were provided a pen and paper to take notes as needed when solving the black box problem, and we observed them adopting the practice to varying degrees, from creating a diagram and recording all data collected in detail to not taking any notes. Results from the regression model confirm that effective data recording is preceded by effective data collection and occurs during the pauses in between on-screen activities as students recorded and organized the current and voltage readings on paper.

Predictor	Estimate	SE	p value
Intercept	-0.61	0.23	0.01
Background - AP	0.50	0.28	0.09
Background - college	1.09	0.29	< 0.001
Voltmeter	-0.13	0.11	0.22
Adjusted R-squared	**0.16**		

Table 4. Best-fitting linear regression models predicting solution scores using action-based features

Predictor	Estimate	SE	p value
Intercept	0	0.07	1
Pairs of terminals connected	0.50	0.09	< 0.001
Percentage complex	-0.26	0.09	0.007
Ammeter - zero reading	-0.27	0.08	0.001
Ammeter - non-zero reading	0.16	0.08	0.05
Mean pause post Ammeter	0.15	0.08	0.09
Adjusted R-squared	**0.60**		

Table 5. Best-fitting linear regression models predicting solution scores using knowledge-grounded features

	True performance	
Predicted	Low-performing	Non low-performing
Low-performing	10	3
Non low-performing	3	13
Total	**13**	**16**

Table 6. Confusion matrix for the test set averaged across 20 train-test splits (rounded to the nearest integer).

DISCUSSION & CONCLUSION

In this study, we explored how to automate the measurement of problem solving through log data generated in an open-ended task embedded in a science simulation. We first parsed the log files into user interface-level actions taken by individual students (action-based features), an approach commonly used in other studies. We then extracted a second set of features by segmenting sequences of actions into semantically meaningful episodes related to circuit-building. These features are grounded in our knowledge gained through qualitative observation of the problem-solving process; that the type of circuits built by students were closely related to the practices used to solve the problem. While none of the action-based features was a significant predictor of problem-solving outcomes, the knowledge-grounded features were significantly associated with students' solution scores. The substantial improvement in model performance when using knowledge-grounded vs. action-based features to predict student solution scores illustrates the value of using knowledge gained from qualitative analysis to define the type and unit of features to be extracted from the log data. Though the features identified in this study are specific to the circuit simulation, the workflow of using

	Decomposition	Data Collection	Data Recording
	Coefficient estimate (SE)		
Pairs of terminals connected		0.52(0.08) ***	0.59(0.08) ***
Percentage complex	-0.88(0.06) ***	-0.40(0.08) ***	
Ammeter - zero reading			-0.15(0.08) .
Ammeter - non-zero reading			0.26(0.08) **
Mean pause time post Voltmeter			0.16(0.08) .
Mean pause time post Ammeter			0.17(0.09) .
Adjusted R-squared	**0.77**	**0.64**	**0.58**

*** p<0.001; ** p<0.01; * p<0.05; . p<0.1

Table 7. Best-fitting linear regression models predicting decomposition, data collection and data recording practice scores (range: 0-3) using knowledge-grounded features

human knowledge to define the optimal level of feature granularity (discrete actions vs. meaningful episodes) can be generalized to other interactive learning environments aimed at assessing complex constructs such as problem solving using students' interaction traces.

In the context of the black box problem, we found that students' interaction patterns related to circuit building and ammeter usage could reliably predict their problem-solving outcomes. The building of complex circuits that connect more than two terminals of the black box at a time and obtaining readings of zero using the ammeter significantly decrease the probability of a student solving the problem. In contrast, the building of simple circuits that connect two terminals at a time, obtaining non-zero ammeter readings and longer pause times after using the ammeter increase the probability of solving the problem.

We also found that knowledge-grounded features derived from the log data correspond closely to the human-rated scores measuring students' effectiveness in adopting specific problem-solving practices. In particular, the practice of decomposition was measured by the percentage of circuits built that were complex. A higher percentage of complex circuits corresponds to a student being less effective at decomposing the problem. The practice of data collection was closely related to the pairs of terminals connected by simple circuits. Building simple circuits across all six pairs of terminals corresponds to a student being highly effective at collecting the data needed to solve the problem. Less expected was the association between the practice of data recording and features from the log data, as the practice was evaluated by human coders primarily based on the quality of notes taken by students on paper. Nonetheless, the quality of data recording could be inferred by how effective a student was at building simple circuits and using the Ammeter to collect data, as well as the duration of the pauses after voltmeter and ammeter usages. Longer pauses are associated with more effective data recording practices. Findings from this study will serve as the basis for real-time interventions aimed at improving students' problem-solving practices in future studies. For example, the task environment could release a prompt about decomposition when the log data shows that a student has repeatedly built complex circuits, and the absence of substantial pauses after measurement events (voltmeter and ammeter) would trigger the system to release a prompt about keeping track of the data collected.

We note that there are several limitations with this work. First, as the black box problem is only one instance of tasks designed to mimic authentic problems in science and engineering fields, the generalizability of the problem-solving practices identified in this task needs to be further established. Second, the performance of the classifier to predict students' problem-solving outcomes is limited by our small sample size to train and test the model. In future work, we plan to adopt the workflow of knowledge-grounded feature extraction to similar authentic problem-solving tasks in different domains. We are particularly interested in the extent to which the trials set up by students to collect data and the pause times correspond to students' effectiveness in problem solving. We also intend to run a larger scale study using the black box problem to refine and extract more features from the log data to measure problem solving and provide real-time interventions for students. The ultimate goal is to increase the pedagogical value of interactive simulations in assessing and teaching problem-solving practices useful for solving real-world challenges.

ACKNOWLEDGMENTS

We would like to thank Ana-Maria Istrate and Michael Kauzmann for their help with processing the log data. This research was supported by the Gordon and Betty Moore Foundation.

REFERENCES

[1] Saleema Amershi and Cristina Conati. 2009. Combining unsupervised and supervised classification to build user models for exploratory learning environments. *JEDM| Journal of Educational Data Mining* 1, 1 (2009), 18–71.

[2] Charoula Angeli, Sarah K Howard, Jun Ma, Jie Yang, and Paul A Kirschner. 2017. Data mining in educational technology classroom research: Can it make a contribution? *Computers & Education* 113 (2017), 226–242.

[3] Ryan S Baker, Jody Clarke-Midura, and Jaclyn Ocumpaugh. 2016. Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning* 32, 3 (2016), 267–280.

[4] Ryan Shaun Baker and Paul Salvador Inventado. 2014. Educational data mining and learning analytics. In *Learning analytics*. Springer, 61–75.

[5] Randy E Bennett. 2018. Educational assessment: What to watch in a rapidly changing world. *Educational measurement: issues and practice* 37, 4 (2018), 7–15.

[6] Jody Clarke-Midura and Chris Dede. 2010. Assessment, technology, and change. *Journal of Research on Technology in Education* 42, 3 (2010), 309–328.

[7] ABET Engineering Accreditation Commission. 2020. Criteria for Accrediting Engineering Programs 2020-2021. (Sept. 2020). Retrieved September 30, 2020 from https://www.abet.org/accreditation/accreditation-criteria/

[8] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. *arXiv preprint arXiv:1807.00154* (2018).

[9] Janice D Gobert, Michael Sao Pedro, Juelaila Raziuddin, and Ryan S Baker. 2013. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences* 22, 4 (2013), 521–563.

[10] Samuel Greiff, Sascha Wüstenberg, and Francesco Avvisati. 2015. Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education* 91 (2015), 92–105.

[11] Tanja Käser, Nicole R Hallinen, and Daniel L Schwartz. 2017. Modeling exploration strategies to predict student performance within a learning environment and beyond. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 31–40.

[12] John S Kinnebrew, Kirk M Loretz, and Gautam Biswas. 2013. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining* 5, 1 (2013), 190–219.

[13] Gerd Kortemeyer. 2016. The losing battle against plug-and-chug. *The Physics Teacher* 54, 1 (2016), 14–17.

[14] Frank Levy and Richard J Murnane. 2013. Dancing with robots: Human skills for computerized work. *Washington, DC: Third Way NEXT* (2013), 5–35.

[15] OECD. 2013. *PISA 2012 Assessment and Analytical Framework*. 264 pages. DOI:http://dx.doi.org/https://doi.org/https://doi.org/10.1787/9789264190511-en

[16] Sarah Perez, Jonathan Massey-Allard, Deborah Butler, Joss Ives, Doug Bonn, Nikki Yee, and Ido Roll. 2017. Identifying productive inquiry in virtual labs using sequence mining. In *International conference on artificial intelligence in education*. Springer, 287–298.

[17] Katherine Perkins, Wendy Adams, Michael Dubson, Noah Finkelstein, Sam Reid, Carl Wieman, and Ron LeMaster. 2006. PhET: Interactive simulations for teaching and learning physics. *The physics teacher* 44, 1 (2006), 18–23.

[18] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*. 505–513.

[19] Burkhard Priemer, Katja Eilerts, Andreas Filler, Niels Pinkwart, Bettina Rösken-Winter, Rüdiger Tiemann, and Annette Upmeier Zu Belzen. 2020. A framework to foster problem-solving in STEM and computing education. *Research in Science & Technological Education* 38, 1 (2020), 105–130.

[20] Edys S Quellmalz and James W Pellegrino. 2009. Technology and testing. *science* 323, 5910 (2009), 75–79.

[21] Shima Salehi. 2018. *Improving problem-solving through reflection*. Stanford University.

[22] NGSS Lead States. 2013. Next Generation Science Standards: For States, By States. (July 2013). Retrieved July 30, 2019 from http://www.nextgenscience.org/

[23] Nani Teig, Ronny Scherer, and Marit Kjærnsli. 2020. Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data. *Journal of Research in Science Teaching* 57, 9 (2020), 1400–1429.

[24] Candace Thille, Emily Schneider, René F Kizilcec, Christopher Piech, Sherif A Halawa, and Daniel K Greene. 2014. The future of data-enriched assessment. *Research & Practice in Assessment* 9 (2014), 5–16.

[25] Mary Webb, David Gibson, and Alona Forkosh-Baruch. 2013. Challenges for information technology supporting educational assessment. *Journal of Computer Assisted Learning* 29, 5 (2013), 451–462.

[26] Donald R Woods. 2000. An evidence-based strategy for problem solving. *Journal of Engineering Education* 89, 4 (2000), 443–459.

[27] Margaret Wu and Raymond Adams. 2006. Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics education research journal* 18, 2 (2006), 93–113.

[28] Xiaoming Zhai, Lehong Shi, and Ross H Nehm. 2020a. A Meta-Analysis of Machine Learning-Based Science Assessments: Factors Impacting Machine-Human Score Agreements. *Journal of Science Education and Technology* (2020), 1–19.

[29] Xiaoming Zhai, Yue Yin, James W Pellegrino, Kevin C Haudek, and Lehong Shi. 2020b. Applying machine learning in science assessment: a systematic review. *Studies in Science Education* 56, 1 (2020), 111–151.