Using Log Data to Validate Performance Assessments of Mathematical Modeling Practices

Joe Olsen[1(✉)], Amy Adair[1], Janice Gobert[1,2], Michael Sao Pedro[2], and Mariel O'Brien[1]

[1] Rutgers University, New Brunswick, NJ 08901, USA
joseph.olsen@rutgers.edu
[2] Apprendis, Berlin, MA 01503, USA

Abstract. Many national science frameworks (e.g., Next Generation Science Standards) argue that developing mathematical modeling competencies is critical for students' deep understanding of science. However, science teachers may be unprepared to assess these competencies. We are addressing this need by developing virtual lab performance assessments that assess these competencies in science inquiry contexts. Through our design processes, we developed a method for validating the assessments that takes advantage of the unique opportunities afforded by collecting log data. Here, we describe this method and demonstrate its utility by analyzing students' competencies with one example sub-practice of mathematical modeling, *plotting controlled data generated from a simulation*.

Keywords: Intelligent tutoring system · Mathematical modeling · Log data

1 Introduction

To help promote students' deep understanding of science and mathematics necessary for future college and career readiness in STEM [1], standards like the Next Generation Science Standards (NGSS) [2] emphasize the integration of disciplinary ideas and concepts with science and engineering practices, including *using mathematics and computational thinking* (NGSS Practice 5) and *developing and using models* (NGSS Practice 2). These practices, though, can be difficult for teachers to assess without resources that can capture students' competencies in real time [3]. To address this need, we are developing virtual lab performance-based formative assessments within the Inquiry Intelligent Tutoring System (Inq-ITS) environment [4]. These assessments automatically measure students' competencies at building mathematical models within science inquiry contexts using knowledge-engineered algorithms [5, 6]. Part of the development process of assessments and algorithms entails ensuring that they validly and reliably capture the broad range of competencies students may demonstrate. In this paper, we present a method in which we triangulate the virtual lab evaluations with students' actions in the lab and their multiple-choice responses to collect evidence about specific interpretations

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13356, pp. 488–491, 2022.
https://doi.org/10.1007/978-3-031-11647-6_99

of assessment scores on the mathematical modeling task [6]. This method can be useful for rigorously validating the logs and assessment data yielded by intelligent tutoring systems that scaffold and assess competencies for similar complex domains.

2 Method

2.1 Participants and Procedure

US High school students (N = 107) completed an online multiple-choice assessment, followed by an Inq-ITS virtual lab on a physical science topic (i.e., momentum, gravity, or friction) chosen by their teacher. In the virtual lab, students collected quantitative data using an interactive simulation, and developed a mathematical model to fit the trend in their data. This paper focuses on students' responses and actions related to one of the many sub-practices assessed within the system, *plotting controlled data.* The data related to this sub-practice include students' responses to the *Selecting Controlled Data* multiple-choice item (Fig. 1) as well as students' actions on the *Plotting Data* stage, where students must label axes and choose data among the trials they have collected to plot on a graph (Fig. 2).

A student used a computer simulation to study the motion of a ball being dropped from different heights.

Which trials should she use to construct a graph the explore the relationship between the *Height of the Drop* and the *Speed of the Ball*?

a) Trials 4 and 5
b) Trials 2, 3, and 4
c) Trials 1 and 5
d) All the trials
e) Trials 1, 2, and 3

Trial Number	Mass of Ball (kg)	Height of Drop (m)	Speed of Ball (m/s)
1	5	10	10.10
2	10	30	24.21
3	10	50	31.26
4	10	70	40.41
5	15	70	45.48

Fig. 1. Multiple-choice item for *Selecting Controlled Data*

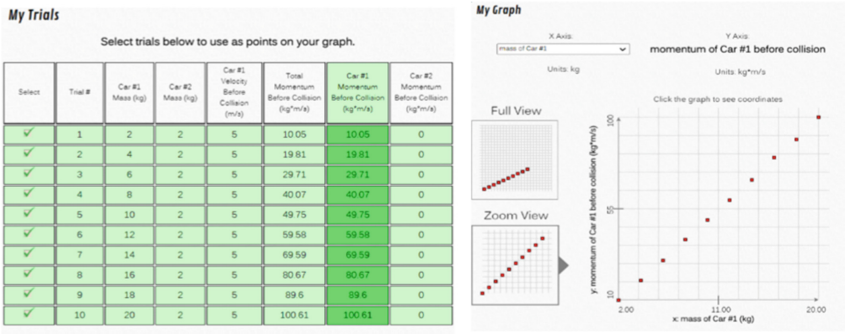

Fig. 2. Screenshots of the *Plotting Data* stage of the mathematical modeling task

2.2 Measures

We triangulated three data sources: students' responses to multiple-choice items (MCs), evaluation logs (ELs), and action logs (ALs). Specifically, the MC item for *plotting controlled data* prompts the student to identify which data from a table should be plotted according to a given goal (Fig. 1). The response is scored as correct (1) or incorrect (0) depending on whether the student selected the answer option with all controlled trials (i.e., choice B). We also collected students' EL scores generated from students' performances within the virtual lab. Scores were generated using knowledge-engineered algorithms based on the axes chosen and points plotted when constructing their graph [5, 6]. The EL score for the sub-practice of *plotting controlled data* was logged as correct (1) or incorrect (0) depending on whether the student selected all controlled trials to plot (Fig. 2). Finally, we gathered the sequential, timestamped actions taken by the students within the virtual lab through the ALs, which detailed what students clicked (e.g., which axes they chose, which points they selected and de-selected) and when.

2.3 Approach to Validating the Virtual Lab Performance Assessment

We compared students' scores on the virtual lab and multiple-choice assessment (i.e., the EL and MC scores, respectively) using 2×2 contingency tables. These represent the frequency distribution of student scores within each task type for *plotting controlled data.* We then selected a random subsample of students and analyzed their ALs to generate hypotheses that could explain any discrepancies found between MC and EL scores. Next, we determined which features of the ALs substantiated our hypotheses and distilled the remaining log data into summary reports of those features. From this data, we were able to generate arguments for or against the intended interpretation of the assessment scores.

3 Results: Applying the Method

Relationship between Multiple-Choice and Virtual Lab Performance for Plotting Controlled Data. Table 1 shows that, of the 85 students who received EL = 1 for this sub-practice, 60% (51/85) answered the related MC item incorrectly (MC = 0). We then analyzed the ALs to understand this pattern.

Table 1. 2×2 contingency table for *plotting controlled data* sub-practice.

		Virtual Lab Evaluation Log (EL) score		
		0	1	Total
Multiple Choice Item (MC) score	0	15 (14%)	51 (48%)	66 (62%)
	1	7 (7%)	34 (32%)	41 (38%)
	Total	22 (21%)	85 (79%)	107

From the ALs, we generated two hypotheses: (1) students who have mastered collecting controlled data with a simulation may not have mastered selecting a subset of

controlled data from a larger set of uncontrolled data, and (2) students may operate under the misconception that they should utilize (i.e., plot) all data points available to them when constructing mathematical models. In relation to the first hypothesis, we found that 94% (48/51) of the students who received EL = 1 and MC = 0 had collected only controlled data; thus, it was impossible for these students to plot uncontrolled data since they did not have uncontrolled data available. Given that these students had also incorrectly answered the MC item (Fig. 2), we suspect that these students do not fully have this competency; thus, future designs of the virtual lab assessment should be able to discriminate between students with partial competencies with this sub-practice. In relation to the second hypothesis, 83% (89/107) plotted all data points that they collected within the system. Of these students, 45% (40/89) selected the "all trials" option for the MC item. These two pieces of evidence together suggest that selecting all data points that are available might represent a misconception among students, and future design iterations should include scaffolds to help students address this misconception.

4 Discussion

For educators to be confident that systems correctly measure students' competencies, we suggest more effort be spent on developing and utilizing validation methods which leverage log data, such as the method described in this paper. This method can be generalized to other domains and systems that make use of an external measure (e.g., multiple-choice items) that align with fine-grained constructs, performance assessments that can evaluate the same constructs, and additional human-interpretable log data of students' behavior. Such methods like ours not only provide evidence about validity, but also highlight ways to improve the design of performance assessment tasks.

Acknowledgements. This material is based upon work supported by the U.S. Department of Education Institute of Education Sciences (Award Numbers: R305A210432 & 91990019C0037; Janice Gobert & Mike Sao Pedro) and an NSF Graduate Research Fellowship (DGE-1842213; Amy Adair). Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not necessarily reflect the views of either organization.

References

1. National Science Board: Science and Engineering Indicators Digest 2016 (NSB-2016-2). National Science Foundation, Arlington, VA (2016)
2. NGSS Lead States: Next Generation Science Standards: for States, by States. The National Academies Press, Washington, DC (2013)
3. Hernández, M.L., Levy, R., Felton-Koestler, M.D., Zbiek, R.M.: Mathematical modeling in the high school curriculum. Math. Teach. **110**(5), 336–342 (2016)
4. Gobert, J.D., Sao Pedro, M., Raziuddin, J., Baker, R.S.: From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. J. Learn. Sci. **22**(4), 521–563 (2013)
5. Dickler, R.: An intelligent tutoring system and teacher dashboard to support students on mathematics in science inquiry. ProQuest Dissertations Publishing (2021)
6. Dickler, R., et al.: Supporting students remotely: integrating mathematics and sciences in virtual labs. In: International Conference of Learning Sciences, pp. 1013–1014. ISLS (2021)