Capturing modeling pathways using the Modeling Assessment for Physics Laboratory Experiments

Michael F. J. Fox[1,2], Benjamin Pollard[1,2], Laura Ríos[3], and H. J. Lewandowski[1,2]
[1]*JILA, National Institute of Standards and Technology and the University of Colorado, Boulder, CO 80309, USA*
[2]*Department of Physics, University of Colorado Boulder, Boulder, CO 80309, USA and*
[3]*Physics Department, California Polytechnic State University, San Luis Obispo, CA 93407, USA*

A choose-your-own-adventure online assessment has been developed to measure the process of modeling undertaken by students when asked to measure the Earth's gravitational constant, g, using a simple pendulum. This activity forms part of the Modeling Assessment for Physics Laboratory Experiments (MAPLE), which is being developed to assess upper-division students' proficiency in modeling. The role of the pendulum activity is to serve as a pre-test assessment with apparatus that students are likely to be familiar. Using an initial sample of student data from a development phase of the assessment, we show that the pendulum activity is able to discriminate between a range of student processes that are relevant to understanding student engagement with modeling as a scientific tool.

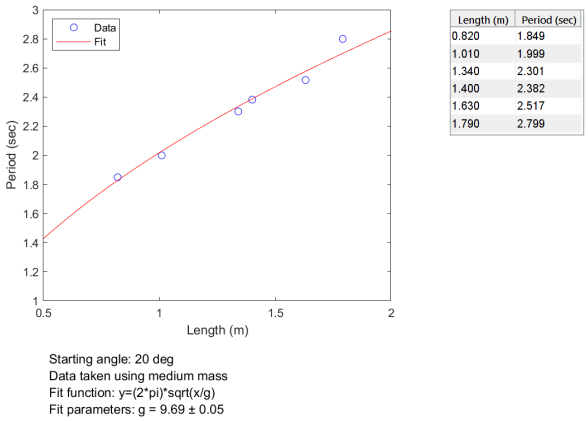
2020 PERC Proceedings edited by Wolf, Bennett, and Frank; Peer-reviewed, doi.org/10.1119/perc.2020.pr.Fox

I. INTRODUCTION

Knowing how to construct, test, and refine models is considered an important learning goal for many undergraduate laboratory courses [1]. Therefore, being able to assess both student proficiency in the skill of modeling and the impact a course has on learning that skill is valuable for both students and instructors. The Modeling Assessment for Physics Laboratory Experiments (MAPLE) is being developed in a 4-phase process to address this need for a validated instrument for assessing model-based reasoning in lab courses, specifically at the upper-division level [2, 3].

The guiding theoretical framework for the four phases of development is the Modeling Framework for Experimental Physics [4–6]. The Modeling Framework describes modeling through five sub-tasks that are connected in a cyclical and not necessarily linear manner [7]. The five sub-tasks are: making measurements; constructing models - of both measurement and physical systems; making comparisons between data and predictions; proposing causes for disagreements; and enacting revisions - to either models or apparatus. One result from the initial two phases of development of the MAPLE was the identification of a "need for a process-oriented assessment that allows us to observe the approaches to modeling students take" [2, 3]. In phase 3, the assessment items that compose the MAPLE have been constructed, resulting in a pre-test and two post-tests. Each test is composed of two parts: Part 1 is a choose-your-own-adventure activity that has been designed to measure the process a student takes when modeling — i.e. which sub-tasks they decide to undertake and in what order those decisions are made. Part 2 is a series of multiple-choice items each coupled with an open-response question asking students to explain their reasoning, designed to assess student competency in specific sub-tasks [8]. The MAPLE is an assessment designed for either an electronics or an optics lab course, hence the requirement for distinct post-tests. The pre-test is the same for both electronics and optics courses, and has been designed around a typical intro-level lab in classical mechanics, in order not to require specific domain knowledge that may hinder student capacity to engage in modeling [9].

In this work, we describe the pre-test choose-your-own-adventure activity, which asks students to determine the Earth's gravitational constant, g, using a pendulum — henceforth referred to as the pendulum activity. We report student results from the phase-3 version of the pendulum activity, which will be used to inform future development of the assessment. We use these data to address the question: *Is the MAPLE pendulum activity able to discriminate between different paths students take through the Modeling Framework?* By *discriminate,* we mean the ability to produce measurably different data from students who take fundamentally distinct approaches to modeling. While related to the idea of the item-total or point-biserial correlation in classical item statistics [10] or item discrimination in two-parameter Item Response Theory models [11], the absence of distinct items in our activity precludes the use of these traditional measures of discrimination. Nonetheless, we aim to illustrate discrimination in our assessment by showing that distinct approaches to modeling can be detected in our data.



Length (m)	Period (sec)
0.820	1.849
1.010	1.999
1.340	2.301
1.400	2.382
1.630	2.517
1.790	2.799

FIG. 1. An example of a portion of the screen a student would see after: selecting an initial angle to release the pendulum; choosing to use 6 strings; measuring those strings to the top of the hanging mass when mounted; measuring the period of the pendulum swing; and performing a fit. In this example, both the measurement method used for the string lengths and the initial angle for release lead to the value for g from the fit to be smaller than the expected value.

To be able to answer this question, we have to categorize student paths by their shared characteristics [12]. We note that the path a single student takes may belong to multiple categories. The utility of performing these categorizations is that it allows us to measure the prevalence and tendency for relationships between the different modeling sub-tasks in student populations (in a similar vein as the work of Ĉanĉula *et al.* [13]).

For this work, we focus on two defining features of a path: the number of revisions and the end point. In the Modeling Framework, students perform a comparison between the prediction from a physical model and their data before making a decision whether to: stop if their comparison is "good enough"; or to continue by either enacting revisions or taking more data. Previous work has highlighted that students struggle with this part of the Modeling Framework [7, 9], hence by showing that the MAPLE is able to discriminate between the different pathways taken by students we demonstrate the assessment's relevance to problems of interest, as well as its scope as a measurement instrument.

A. The pendulum activity

The pendulum activity is administered online using Qualtrics [14]. Students are initially provided information on the period, T, of a pendulum, with a step-by-step derivation of the equation: $T^2 = 4\pi^2 L/g$, using the small-angle approximation, where L is the length of the pendulum. They are then presented with a diagram and list of the apparatus available. Then they are given the objective - to perform measurements to determine a value for Earth's gravitational constant, g. This is followed by the statement: *"Previous measurements have found* $g = 9.80 \pm 0.05$ m/s^2. *You must make sure that the apparatus can measure the same value for the class."*[15]

After the introduction, students are given a list of choices

TABLE I. Choices presented to students in the pendulum activity and the options within those choices regarding changes to the apparatus or to the physical model. These choices are shown to the student below the portion of the screen illustrated in Fig. 1, and are selected using radio buttons. A second screen is then displayed offering the parameter choices as appropriate. The final choice, 11, is displayed only after one of the other choices has been selected.

#	Choice	Parameter options
1	Revisit the information from the start	-
2	Mount protractor to set the angle at which you will measure the period of the pendulum	$5°$, $10°$, $20°$, $45°$, other
3	Measure the length of the pendulum for each of the strings you are using	on a table; to the top, middle or bottom of the pendulum mass when mounted
4	Use longer or shorter strings	3, 6, or 9 different string lengths
5	Change the mass to one with a different weight	small, medium, or large
6	Measure the period of the pendulum for each of the strings you are using	-
7	Change whether you square the period values before plotting them	squared, not squared
8	Set the model function used to fit your data in the software	squared function and/or include y-intercept
9	Look up information about the detector	-
10	Propose a reason for what is going on	-
11	Decide that I have completed the task and am ready to finish this activity	-

TABLE II. Demographic information for participants. Not all of the students ($N = 93$) responded to these optional questions, some reported in more than one category. Categories with no responses are not shown. Physics includes the major Engineering Physics.

Gender and race/ethnicity	%	Major and year	%
Male	72.0	Physics	90.3
Female	22.6	Engineering	1.1
Other gender	5.4	Other STEM	4.3
Asian American	14.0	Other disciplines	3.2
Black or African American	3.2	Second	15.1
Hispanic/Latino	14.0	Third	45.2
White	61.3	Fourth	26.9
Other race/ethnicity	3.2	Fifth and beyond	11.8

(Table I). Each choice is designed to be associated with one sub-task from the Modeling Framework. On selecting some of the choices, students are presented with more information, while others allow them to set parameters that they will use for collecting data using the pendulum apparatus. We consider the first *measurement* to have been made once all the following choices have been selected: the length of the pendulum has been measured (choice 3), the period has been measured (choice 6), and the model function to fit the data has been chosen (choice 8). At this point, a complete graph is shown to the student (e.g. Fig. 1), which displays the data points, the fitted line, and the value for the fit parameters. One of the fit parameters is g. The list of choices available to the students is displayed below the graph. Subsequent *measurements* occur whenever one of the model or apparatus parameters changes, generating new data or a new fit, and causing a new evaluation of the fit parameter g. In total, there are 1920 different combinations of parameters available for students to explore. The pendulum activity ends when either choice 11 is selected, which prompts the student to enter the value they wish to report for g, or they have made 50 choices.

II. METHODOLOGY

The MAPLE pre-test was administered to 133 students in 7 courses at 6 institutions in the U.S. The size of these institutions is well distributed: 2 large, 2 medium, and 2 small or very small [16]. We report the demographic data of the students in Table II to provide context for this work. The assessment was issued in Fall 2019. We aggregate all students into one group for this analysis, as we are currently not interested in course-level analysis. We analyze the responses from only the 93 students who reported a value for g at the end of the pendulum activity. Of the 133 students that clicked the link to participate in the assessment, 29 did not select choice 11 to end the activity and report their final answer. The majority, 18, of these did not start the activity, with the remainder not reporting for a variety of technical reasons. This leaves 11 students who took part in the modeling process and did not report a value for g. These 11 have legitimate modeling pathways, which we do not include in the present analysis because of the inability to classify the nature of their end point. This highlights the fact that the present analysis is only on a subset of possible pathways that can exist, and so our claims can only be recognized in that context.

We now define how we use the student data to quantify the two defining features of a path that we are interested in, the number of revisions and the end point. To calculate the number of revisions each student made, we counted each time one of the parameters associated with choices 2, 3, 4, 5, 7, or 8 in Table I were changed — i.e. a new measurement had occurred. This manner of counting does not count any changes to parameters before the first value of g has been measured, and so is consistent with the goal of identifying how many times a student performed a revision. Note, choices 2, 3, 4, and 5 correspond to revisions to the apparatus, while choices 7 and 8 correspond to revisions to the model, and for the current analysis we do not distinguish these two. By analyzing the shape of the distribution of the number of revisions taken by a group of students, we characterize the modeling behavior of the group as a whole (discussed further in Section III A).

The end point of the pendulum activity is reached upon selection of choice 11, where a text box is provided with instructions to enter the value for g that the student wanted to report. There was no prompting to include an uncertainty with that value, though fit parameter results were always presented with accompanying uncertainties. We processed this input and recorded separately students' reported values for

Earth's gravitational constant g_r and associated uncertainty δg_r. We classify the reported values in two ways: first by agreement with the stated value of $g = 9.80 \pm 0.05\ \mathrm{m/s^2}$; and second by agreement with a measured value during the activity.

The first classification provides insight into the sub-task of making comparisons and student views on what is considered to be a "good enough" comparison to finish the activity. There are two conditions we have identified that might be used to make this comparison. The first we refer to as the *bull's-eye* condition: that g_r is within the expected range of $9.80 \pm 0.05\ \mathrm{m/s^2}$, as this is consistent with the explicit goal provided to the students at the start of the activity (see Section I A). This condition is not affected by whether the student reported an uncertainty. The second we refer to as the *overlap* condition: that the set of values $[g_r - \delta g_r, g_r + \delta g_r]$ has a non-empty intersection with the set $[9.75, 9.85]\ \mathrm{m/s^2}$, which, in taking account of the uncertainty on the measurement, may be considered to be a more expert-like comparison [17, 18].

The second classification allows us to determine whether a student used the data they gathered to report their final answer. This is especially important for the pendulum activity, because students have been told what the expected outcome is and they are able to report their final answer after having navigated through only one of the choices. We classify students' reported values into three groups: (A) g_r and δg_r match pairwise with a measured value and its uncertainty; (B) g_r matches with a measured value but the uncertainty does not match; and (C) g_r does not match with any measured value.

III. RESULTS AND DISCUSSION

A. The distribution of the number of revisions

Before looking at the distribution of the number of revisions from our sample of students, let us discuss the expectations we have for the possible shapes this distribution may have. If students engage in the iterative aspect of modeling by making revisions to their apparatus or models with the aim of improving agreement, then one would expect that the distribution of the number of revisions should be peaked at a finite number of revisions. On the other hand, if students tend not to make revisions, then the peak of the distribution should occur at zero revisions. In Fig. 2(a), we see that the latter is the case for our sample of students, with 23 not performing any revisions. This then raises the question of why these students are not engaging in revisions: is the design of our instrument limiting our ability to measure? We answer this question in Section III B.

If we work on the assumption that, in order to get a "good enough" answer, students need to perform revisions and iterate through the Modeling Framework, then one would expect to see a difference in the distributions of the number of revisions between students whose reported results met a "good enough" condition (i.e. a peak at a non-zero number of revisions) and those that did not. In Fig. 2(b) and 2(c), we indeed see tentative evidence that such a difference occurs for the *bull's-eye* condition, with a non-zero peak occurring at 8 revisions. This has the caveat that we do not have a sample size large enough to perform statistical comparisons and, therefore, present this data as an preliminary indication of the kind of insight available in data from the MAPLE. Conversely, in Fig. 2(d) and 2(e), for the *overlap* condition, there appears to be less difference between the distributions. In total, 58 students met at least one of the two "good enough" conditions, the majority, 40, satisfied only one of the conditions, suggesting that there are two distinct groups of students. This, in part, explains why there is a difference between the shapes of the distributions between Fig. 2(b) and 2(d). However, at this moment, we can only speculate on why there is some association between which "good enough" condition students' reported answers met and the shape of the distribution of the number of revisions; which will be the topic of future investigations.

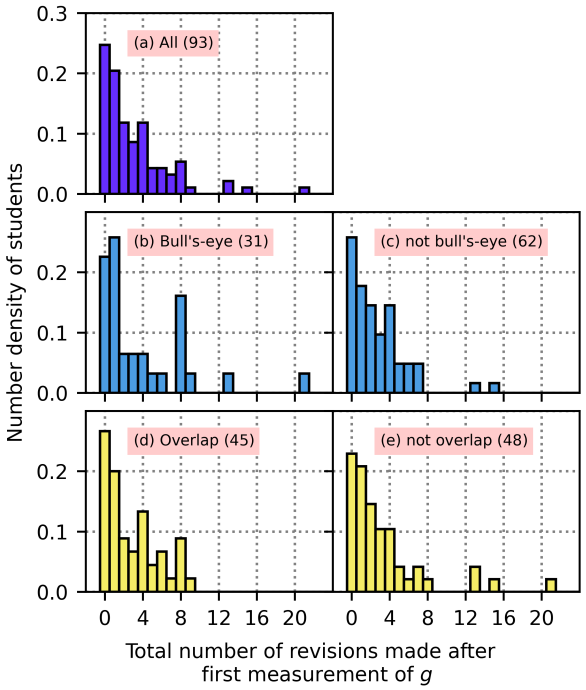


FIG. 2. Distributions of the number of students as a function of the number of revisions made, normalized by the total number of students, which is given in each legend. In (a) the distribution is for all students in our sample. In (b) and (c) the distributions are for students who either did or did not satisfy the *bull's-eye* condition, respectively. Similarly, (d) and (e) correspond to those students satisfying or not the *overlap* condition, respectively. This measure counts a change even if a student returns to a previous set of parameters.

Before returning to the question of why the most common number of revisions was zero before reporting a final answer, we digress to note more students reported an answer that satisfied the more expert-like (using uncertainties for reasoning) *overlap* condition (45) than the *bull's-eye* condition (31). In addition to demonstrating that the MAPLE is able to distinguish between nuanced aspects of the 'making comparisons' sub-task of the Modeling Framework, these numbers themselves are interesting, as they show students using a set-like paradigm for thinking about measurements [17].

TABLE III. Contingency table for students who did not perform any revisions. The first two rows indicate the meaning of groups A, B, and C as defined at the end of Section II.

Zero revisions	All	A	B	C
Any g_r and reported g_r match	-	✓	✓	✗
With δg_r and reported δg_r match	-	✓	✗	✗
Either *bull's-eye* or *overlap*	15	11	3	1
Neither *bull's-eye* nor *overlap*	8	3	1	4

B. Student behavior when not performing revisions

There are four possible reasons for students to report an answer after performing zero revisions. We list these reasons below, while counting the number of the 23 students' reported results that fall into these categories:

1. the parameters are set that would produce a value of g to meet the either the *bull's-eye* or the *overlap* condition before performing the first measurement. (14 results - *bull's-eye*: 6; *overlap*: 11)
2. The first value that is measured is reported without matching either the *bull's-eye* or the *overlap* condition. This suggests a barrier exists preventing iteration. (4 results)
3. Being aware of what the final answer should be, students skip the process of performing revisions and report $g_r = 9.80\ \mathrm{m/s^2}$ despite not measuring that value. (1 result)
4. The reported answer is unrelated to either their measured value or the expected value, which might, for instance, be due to a typographic error. (4 results)

In order to distinguish between these four reasons, we classify students' reported values as described in Section II and evaluated in Table III for just those students who did not perform any revisions. Most of this subset of students' reported values satisfied either of our two conditions and this value was the first value they measured (the sum of groups A and B in the first row of Table III). The relatively small number of students whose reported results met the *bull's-eye* condition indicates that the design of the pendulum activity is at an appropriate level for the students from the sample group, such that the vast majority do not solve the problem in the first step. Further supporting this, when inspecting students' reported results for the *overlap* group, only 3 students reported $\delta g_r \leq 0.1\ \mathrm{m/s^2}$, with a couple reporting uncertainties as large as $7\ \mathrm{m/s^2}$. Indeed, the number of students with zero revisions who satisfied both criteria was 3.

A small number of students' reported results (4) did not meet either of our "good enough" conditions, but they still reported their measured value (the sum of groups A and B in the second row of Table III). This suggests that either they did not know how to engage in performing revisions, were not motivated to do so, or were using a different "good enough" condition to either of the ones we have defined. Our data is not able to distinguish these reasons, however, future analysis of think-aloud interviews with students should help to illuminate this aspect of student behavior.

Only 1 student did not perform any revisions and reported $g_r = 9.8\ \mathrm{m/s^2}$ (group C, top row). Across all students in our study, for any number of revisions, 7 reported $g_r = 9.8\ \mathrm{m/s^2}$ or $g_r = 9.81\ \mathrm{m/s^2}$ without having measured that value during the pendulum activity. As the activity is a verification activity, it was expected that some students might resort to this option when reporting their result. It is reassuring that this number is relatively low (7.5%), and we have shown that we are able to identify this mode of engagement with modeling. Finally, we find 4 students with zero revisions report answers that do not meet either "good enough" conditions or were one of their measured values. Manual inspection of their measured values and processes did not provide any explanation for their reported results.

This study is limited by the relatively small number of students who took part in phase 3 development of the MAPLE. The full validation of the MAPLE will occur once the final instrument has been developed and deployed (phase 4). Furthermore, we have been careful when describing students' reported results satisfying either of our "good enough" conditions, to not imply that either of those were the conditions a student actually used. This is important, as our data shows what a student did, while the question of why they did it requires inferences to be made from those data. Ultimately, there still remain many more questions, such as: what is the effect of framing by the instructor on student interaction with the activity? What are the paths experts, such as physics faculty, would use to complete the pendulum activity? And what other modeling pathways can be captured by the MAPLE?

IV. CONCLUSIONS

We have presented the pre-test choose-your-own-adventure activity part of the online MAPLE assessment based on using a pendulum to measure Earth's gravitational constant — the pendulum activity. This part of the MAPLE is specifically designed to measure the process students take through the Modeling Framework. We have demonstrated, using data from an initial sample of students, that the pendulum activity is able to discriminate effectively between and within the two different modeling pathways presented, by showing that student responses exist in all paths we identified. There are more ways of characterizing paths in the Modeling Framework that we have not used for this analysis, and, therefore, we do not claim to have, as of yet, a complete classification of all the possible pathways that the MAPLE is able to capture. Our choice to look at the number of revisions and end point was motivated by observations that these were a part of the Modeling Framework that students found challenging. Our results suggest that student behaviors around making comparisons and performing revisions are complicated, for example, by showing that students' reported results can satisfy different "good enough" conditions; therefore motivating further investigation.

ACKNOWLEDGMENTS

We wish to thank the instructors and students who took the MAPLE for this work. We also wish to thank Alexandra Werth for conducting think-aloud interviews that led to pre-distribution revisions of the pendulum activity. This work was supported by NSF under Grant Nos. DUE-1611868 and PHYS-1734006.

[1] J. Kozminski, H. Lewandowski, N. Beverly, S. Lindaas, D. Deardorff, A. Reagan, R. Dietz, R. Tagg, J. Williams, R. Hobbs, *et al.*, AAPT recommendations for the undergraduate physics laboratory curriculum, (2014).

[2] D. R. Dounas-Frazer, L. Ríos, B. Pollard, J. T. Stanley, and H. J. Lewandowski, Characterizing lab instructors' self-reported learning goals to inform development of an experimental modeling skills assessment, Phys. Rev. Phys. Educ. Res. **14**, 020118 (2018).

[3] L. Ríos, B. Pollard, D. R. Dounas-Frazer, and H. J. Lewandowski, Using think-aloud interviews to characterize model-based reasoning in electronics for a laboratory course assessment, Phys. Rev. Phys. Educ. Res. **15**, 010140 (2019).

[4] B. M. Zwickl, N. Finkelstein, and H. J. Lewandowski, The process of transforming an advanced lab course: Goals, curriculum, and assessments, American Journal of Physics **81**, 63 (2013).

[5] B. M. Zwickl, N. Finkelstein, and H. J. Lewandowski, Incorporating learning goals about modeling into an upper-division physics laboratory experiment, American Journal of Physics **82**, 876 (2014).

[6] D. R. Dounas-Frazer and H. J. Lewandowski, The modelling framework for experimental physics: description, development, and applications, European Journal of Physics **39**, 064005 (2018).

[7] L. Ríos, B. Pollard, D. Dounas-Frazer, and H. J. Lewandowski, Pathways to proposing causes for unexpected experimental results, in *Physics Education Research Conference 2018*, PER Conference (Washington, DC, 2018).

[8] In phase 4, the open responses have been used to inform the creation of multiple-choice reasoning elements in a coupled-multiple-response format.

[9] B. M. Zwickl, D. Hu, N. Finkelstein, and H. J. Lewandowski, Model-based reasoning in the physics laboratory: Framework and initial results, Phys. Rev. ST Phys. Educ. Res. **11**, 020113 (2015).

[10] S. A. Livingston, Item analysis, in *Handbook of test development*, edited by T. M. Haladyna and S. M. Downing (Lawrence Erlbaum Associates, Mahwah, NJ, 2006) Chap. 19, pp. 400–21.

[11] R. M. Luecht, Designing Tests for Pass-Fail Decisions Using Item Response Theory, in *Handbook of test development*, edited by T. M. Haladyna and S. M. Downing (Lawrence Erlbaum Associates, Mahwah, NJ, 2006) Chap. 25, pp. 575–98.

[12] We use the terms 'process', 'path', and 'pathway' as synonyms in this work.

[13] M. P. Ĉanĉula, G. Planinŝiĉ, and E. Etkina, Analyzing patterns in experts' approaches to solving experimental problems, American Journal of Physics **83**, 366 (2015).

[14] https://www.qualtrics.com/.

[15] This phrasing is currently under review as part of the development process of the MAPLE.

[16] Indiana University Center for Postsecondary Research (n.d.), *The Carnegie Classification of Institutions of Higher Education* (Bloomington, IN, 2018).

[17] F. Lubben, B. Campbell, A. Buffler, and S. Allie, Point and set reasoning in practical science measurement by entering university freshmen, Science Education **85**, 311 (2001).

[18] B. Pollard, R. Hobbs, D. Dounas-Frazer, and H. J. Lewandowski, Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses, in *Physics Education Research Conference 2019*, PER Conference (Provo, UT, 2019).