Creating a coupled multiple response assessment for modeling in lab courses

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

Research-based assessment instruments (RBAIs) are essential tools to measure aspects of student learning and improve pedagogical practice. RBAIs are designed to measure constructs related to a well-defined learning goal. However, relatively few RBAIs exist that are suitable for the specific learning goals of upper-division physics lab courses. One such learning goal is modeling, the process of constructing, testing, and refining models of physical and measurement systems. Here, we describe the creation of one component of an RBAI to measure proficiency with modeling. The RBAI is called the Modeling Assessment for Physics Laboratory Experiments (MAPLE). For use with large numbers of students, MAPLE must be scalable, which includes not requiring impractical amounts of labor to analyze its data as is often the case with large free-response assessments. We, therefore, use the coupled multiple response (CMR) format, from which data can be analyzed by a computer, to create items for measuring student reasoning in this component of MAPLE. We describe the process we used to create a set of CMR items for MAPLE, provide an example of this process for an item, and lay out an argument for construct validity of the resulting items based on our process.

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

I. INTRODUCTION

Whether one's goal is to improve pedagogical practice or to understand learning on a fundamental level, the ability to measure student learning outcomes is an integral part of education research. Researchers often use research-based assessment instruments (RBAIs) to study learning outcomes, either as a measure of the effectiveness of a pedagogy, or to identify trends and relationships that support learning [1]. As such, RBAIs focused on particular learning outcomes have been a significant focus in education research [2, 3].

Physics lab courses, in particular upper-division labs, come with a unique set of learning goals that are distinct from other courses [4]. However, there are relatively few RBAIs designed to measure learning outcomes that are unique to upper-division labs. Ref. [3] mentions three: the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) [5], the Concise Data Processing Assessment (CDPA) [6], and the Physics Lab Inventory of Critical Thinking (PLIC) [7]. Nonetheless, the PLIC was designed for introductory labs.

A common challenge in designing RBAIs is developing questions or prompts and their respective response options, together referred to as *items*, that can measure student reasoning in a scalable format, that is, practical with large numbers of respondents. We use the term *reasoning* in this work generally, defined as the logical connections between concepts. Such reasoning is especially relevant for typical learning goals of upper-division physics labs, which often involve constructing an argument or explicitly connecting one idea to another [4]. This focus on connections between ideas is distinct from learning goals that center on conceptual knowledge in and of itself. Of the three RBAIs mentioned above, the CDPA and the PLIC relate most closely to measuring aspects of reasoning. They do so using standard multiple choice items (in the case of the CDPA), or a variety of item formats that combine Likert scales and multiple-choice, multiple-response questions (in the case of the PLIC). Each of these formats include only closed-form responses, in which the respondent chooses from a set of predefined options. Such data can be analyzed at scale without the need for interpretation of free-response data from individual respondents.

A general closed-form item format that probes the connection between ideas is the Coupled Multiple Response (CMR) format. In physics education RBAIs, the CMR format was described in work to convert the free-response Colorado Upper-division Electrostatics Diagnostic (CUE) into a multiple choice format for scalability [8, 9]. The free-response CUE items present a prompt describing a typical problem in upper-division electrostatics and ask students about the strategies and methods they would use to arrive at a solution. Analyzing CUE data was labor intensive and required a validated, complex rubric [10]. Thus, the CMR format was developed to capture similar information in a closed-form format. The CMR version of these items asks students to choose from a list of answer choices representing different methods, and then select from another list reasoning elements (REs) that support that choice. CMR items are scored automatically based only on the combination of which answer choice and which REs the respondent selected.

In this work, we describe the development of a set of CMR items in the context of a larger project to create a new triad of RBAIs for upper-division physics labs, called the Modelling Assessment for Physics Laboratory Experiments (MAPLE). MAPLE aims to measure students' proficiency with modeling, as described in Section II. Modeling is a distinct learning outcome than those measured by previous RBAIs for physics labs. The purpose of this work is not to present preliminary findings from MAPLE. Instead, it is to establish the construct validity of the CMR items in the assessment by describing the process we employed to create them.

II. BACKGROUND

The Experimental Modeling Framework (EMF) describes a central aspect of laboratory learning. Represented as an iterative flowchart of connected subtasks, it describes the process of constructing, testing, and refining models of physical and measurement systems. For a broader overview and complete description of the EMF, see ref. [11]. For this work, we highlight two aspects of reasoning in the EMF that are particularly apparent in the format of MAPLE: reasoning within and between particular subtasks, and reasoning involved in moving through the overall EMF.

The goal of the MAPLE project is to create a set of RBAIs to measure students' proficiency modeling as defined by the EMF. MAPLE consists of three surveys contextualized in two upper-division lab course subject areas [12], and one area typically covered in intro courses: op-amps in electronics [13], photodiodes in optics [14], and the simple pendulum in introductory mechanics. The pendulum survey is to be used as a pretest, before students have learned necessary content knowledge around photodiodes or op-amps. In this work, we focus on the pendulum survey.

All three surveys have a similar structure comprising two parts, each corresponding to an aspect of the EMF mentioned above. Part 1 concerns the iterative, process-oriented aspect of the EMF, and thus prompts reasoning involved in moving through the framework overall. It is a choose-your-own-adventure-style sequence of actions concerning an apparatus and an associated measurement task, with each action affecting the data that is collected and the results of subsequent analysis. In Part 1, students become familiar with the apparatus, measurements, and analysis before doing part two. More information about Part 1 is available in another PERC Proceedings publication from this year (Fox et al.).

Part 2, the focus of this work, concerns reasoning within and between subtasks of the EMF. This part consists of a series of eight CMR items. Each item is associated with one or two subtasks, typically with one subtask corresponding to selecting an answer choice and a connected subtask in the EMF associated with selecting an RE. The content of the prompts and answer choices resulted from a set of hands-on interviews conducted by LR with students and a pendulum apparatus at

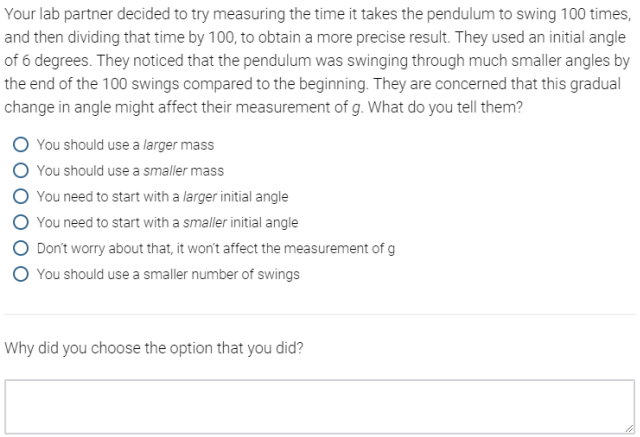

FIG. 1. An example item from the prototype MAPLE pendulum survey, with reasoning prompted as a free-response question.

the University of Colorado Boulder (CU), similar to those described in ref. [15] for the electronics survey.

BP, LR, and HJL initially created a prototype survey containing the prompts and answer choices that resulted from these interviews. The prototype includes a free-response question, "Why did you choose the option that you did?" for every item. This prototype version (implemented in Qualtrics [16]) was tested by BP and another researcher with CU students in a think-aloud interview format in which students responded to the survey in-person on a laptop computer. These interviews allowed us to refine the wording and formatting of the prototype items. Lastly, after REs were created using the process described below, MFJF and BP conducted eight similar think-aloud interviews with CU students using the closed-form CMR items to verify that the REs were interpreted as intended, and made minor changes to their wording as needed.

As an example, we show a prototype item from the pendulum MAPLE survey in Fig. 1. The item concerns measuring 100 swings of the pendulum at once, as opposed to measuring one swing at a time, and the effect of the swing amplitude decreasing over time. A reasonable answer choice is, "Don't worry about that, it won't affect the measurement of g," since the small angle approximation holds that the period of the pendulum is independent of the amplitude. However, the answer choice, "You need to start with a *larger* initial angle" could also be reasonable if one is concerned about the small amplitudes being undetectable by measurement apparatus. As opposed to a standard multiple choice item, the CMR format allows both of these choices to be rated as "correct" if the reasoning is consistent. This item is associated with the Enact Revision and Propose Cause subtasks, with the answer choices representing potential revisions and the reasoning representing the cause that leads to that revision.

III. METHODS

To create REs (reasoning elements) for the CMR items in MAPLE, we distributed the prototype versions of the surveys to instructors, which they then administered to their students. Via email lists and in-person communication, we asked instructors of upper-division labs that involve electronics or optics to fill out an online form with information about their course. Instructors were also provided a link to a version of the prototype survey that provided space for feedback on particular items and the survey overall. While responses in these spaces were minimal, the suggestions resulted in minor changes to the wording of prompts for increased clarity.

We sent each instructor who filled out the form a link to the prototype version of the pendulum survey to administer at the start of their course. We asked the instructors to send the link to their students within two weeks of the start of their course, and suggested that students take the assessment on their own time, outside of class. We also stated that it is best to offer a negligible amount of course credit to students for completing the assessment in order to increase response rates. At the end of each course, we sent the instructor a list of names and student IDs of the students who responded to the survey, as well as the anonymized survey data from their students.

Fifteen instructors responded in Fall 2019, the semester in which the data analyzed here were collected. They spanned colleges across the US, both private and public, from large universities to small, undergraduate-focused institutions. Seven instructors had students who responded to the prototype pendulum survey, for a total of 107 student responses to the complete survey. Demographic and major information from optional questions in the survey is shown in Table I. We include this information as a best practice for many reasons [17], including to provide context for our research findings, as well as to enable meta-studies that combat normative whiteness and highlight inequities in research [18].

TABLE I. Self-reported gender, race, ethnicity, and major of students. "Engineering" excludes the major Engineering Physics, which is included in "Physics."

Female	23.4%
Male	72.9%
Other gender	1.9%
American Indian or Alaska Native	0%
Asian American	14.0%
Black or African American	2.8%
Hispanic/Latino	14.0%
Native Hawaiian or other Pacific Islander	0%
White	59.7%
Other race/ethnicity	3.7%
Physics	92.6%
Engineering	0.9%
Other STEM	3.7%
Other disciplines	2.8%

To create REs from the free-response data, BP and MFJF performed an emergent coding analysis to identify common lines of reasoning. We created a separate set of codes for each item. To start, we selected 20 students from the data set at random, as a representative subset, and independently created codes that represented all lines of reasoning within the 20 responses to each item. BP and MFJF then discussed the codes that emerged, and found that most of the codes that we created independently were similar and could be matched

one-to-one. After discussing any differences in wording used to describe each code, we came to a shared understanding of what each code represented and created a unified set of codes for each item. We also included in the unified set the few codes that did not match one-to-one. As the purpose of the coding was not to make quantitative or generalizable claims about the prevalence of various student reasoning, we did need not perform any further tests of inter-rater reliability.

Then, using this unified set of codes as a starting point, we independently assigned codes to all of the responses, splitting the items between us so that each researcher coded only half of the items. We refined and added codes as needed, aiming to represent all of the various lines of reasoning observed in the data. We also noted that some of the codes pertaining to different items actually represented very similar lines of reasoning; nonetheless, we treated them as separate during this phase of the analysis.

After creating and assigning codes to the entire data set, we analyzed the codes themselves to create CMR REs for each item. This was done as a series of iterative discussions among BP, MFJF, HJL, and other researchers familiar with the project. The following guiding principles informed these discussions. We describe these general principles here, and then provide an example to illustrate them in Section IV. We label each principle with a letter here so that we can explicitly connect back to them in the next section.

Our overall goal was to (a) have the REs represent a broad range of common reasoning observed in student responses, and also ultimately (b) include the reasoning that an expert experimental physicist would use in responding to the item. We also needed to (c) phrase the REs in such a way that it would be possible to unambiguously identify different lines of reasoning based only on which answer choice and REs were selected. We also made sure to (d) phrase each RE in a way that confined it to one subtask in the EMF. Additionally, we aimed to (e) minimize the cognitive overhead involved in reading the full list of REs for an item. Thus we aimed to limit the number of REs in the list and maintained consistent phrasing among them. With that aim in mind, we identified the lines of reasoning that were common between different items, and (f) represented them by similarly phrased REs across the items. Lastly, we (g) struck a balance between phrasing REs in a way that was too specific, effectively trivializing the choice between REs, and having them be too general, limiting our ability to interpret the choice as representative of meaningful differences in reasoning. With that balance in mind, we ensured that every answer choice had at least one RE that plausibly matched it, and allowed for REs that could be matched to only one answer choice.

IV. AN EXAMPLE ITEM

In this section, we present the codes and corresponding REs for a prototype item, calling back to the relevant motivating principles described in the previous section by letter. The prototype item is shown in Fig. 1, and the codes and REs are shown in Table II. The first three columns represent the results of the initial emergent coding of this item, showing the name and description of each code and the number of responses to which it was assigned. In these names and descriptions, we use the word "friction" to be a catch-all for a number of words in students' responses, which also includes "damping," "drag," and "air resistance." The final column represents the result of the subsequent discussions that converted the codes into REs.

In the case of the top six codes, we created REs that corresponded directly to the codes themselves, representing the range of reasoning in our data (a, b). We changed the wording of the codes to have similar phrasing between REs (f), and to have them represent proposed causes to motivate a revision (d). The seventh code in the table, "To reduce friction," represented responses that were distinct from "Friction reduces amplitude" or "Friction reduces period" in that they did not specify the effect of reducing friction, merely stating that reducing friction was the intended outcome. However, when creating REs, we decided against creating an RE representing this response alongside REs representing the more complete responses, in part to lessen the number of responses (e). We also recognized that in a closed-response format, it would be impossible to have REs for the more complete responses without prompting the reasoning behind "To reduce friction," so that if a student selected that RE, we would gain no further insight into their reasoning (c, g).

There were three codes that we decided to not represent as REs, largely to lessen the number of responses (e). These are the ones with "—" in the final column in Table II. They had relatively small numbers of responses, or revealed less about the level of modeling proficiency than the other codes (c).

Lastly, we added one RE that was not present in the coding process, referring to the equation not depending on mass. We added it to mirror the RE about the equation not depending on angle (f). This RE serves a purpose similar to a distractor option in a multiple choice format; it is a true statement, but it does not logically connect to the item prompt and answer choices. A student selecting this response would be aware of model parameters and assumptions generally, but would not be making a connection between the model assumption and the comparison presented in this item (c).

V. DISCUSSION AND CONCLUSION

We believe that the process we undertook for creating CMR items is an integral part of establishing the validity of MAPLE. In addition to statistical metrics for establishing reliability, sensitivity, and dimensionality (such as measures of item-total correlation, test-retest reliability, and factor analyses [19], all of which will be available after MAPLE has been deployed), the validity of RBAIs also rests on their construct validity [20–23]. Construct validity answers the questions, "Does the survey measure what it purports to measure?" and "Is that thing being measured relevant and well-defined?" For MAPLE, the thing being measured is proficiency with the EMF, and its construct validity centers around how closely aligned the surveys are with the theoretical foundations of the EMF. These questions cannot be answered by statistics; they are instead addressed by theoretical considerations and the

TABLE II. Codes and REs for the item shown in Fig. 1. N refers to the number of responses to which that code was assigned.

Code Name	**Code Description**	**N**	**Resulting RE**
Keep things consistent	Everything must be kept consistent for the model to apply.	19	To keep things consistent between measurements
Detector cannot measure	By the 100th swing the angle may be so small that the detector cannot measure a swing very accurately	3	So that the photogate can measure accurately
Small angle approximation	The equation requires that the angles are small	9	Because of the small angle approximation
Period is independent	The pendulum swings with the same period regardless of its starting angle.	24	The equation for period does not depend on the angle
Friction reduces amplitude	Friction/damping reduces the angle/amplitude/energy of the pendulum over time	26	Friction/air resistance cannot be ignored because it reduces the amplitude
Friction reduces period	Friction/damping reduces the period of the pendulum over time	14	Friction/air resistance cannot be ignored because it reduces the period
To reduce friction	This will result in less loss to friction/damping	24	*This code was encompassed by the two more specific REs immediately above.*
100 is a lot	100 swings seems like too many to account for	11	—
Friction is negligible	The effect of friction/damping/air resistance is minimal	2	—
Misc	Doesn't fit into any of the other codes	7	—
—	—	—	The equation for the period does not depend on the mass

process employed to move from theoretical foundations to a finalized instrument.

The relevance and theoretical basis of the EMF itself have been established [11]. To determine if the survey measures what it purports to measure in the context of its CMR items, we turn to the methodologies employed to create these items. The item prompts and answer choices emerged directly from the hands-on interviews with students, as referenced in Section II. This phase of the process ensured that the items represent approaches that are reasonable and familiar to students. Think-aloud interviews, also mentioned in Section II, ensured that the items themselves are being interpreted as intended.

Partnering with instructors across the US broadened the base of student perspectives, making our REs more representative of a broader range of common approaches than those observed in a single institution. Feedback from these instructors about the prototype survey further ensured that the survey was valid in a variety of instructional contexts.

Expert perspectives also ensured that the survey measures a relevant and well-defined construct in experimental physics and physics education. Feedback from our instructor partners incorporated the perspectives of expert physics educators. Nonetheless, think-aloud interviews with experts is a subject of future study. Feedback from colleagues in the iterative discussions to create REs incorporated the perspectives of expert physics education researchers, in addition to our own expertise. Lastly, the researchers directly involved in the development of MAPLE have extensive personal experience with laboratory physics via their graduate studies. Thus, the research team has collective, valuable personal insight into relevant and authentic physics laboratory practice.

Hand-in-hand with establishing validity of an RBAI is defining the scope of that validity. Thus, we note here some limitations to the validity argument we present in this work. First, the CMR items in MAPLE do not capture all the salient aspects of the EMF. In particular, an important aspect of the EMF lies in the iterative and process-focused nature of the framework. This aspect concerns the reasoning involved in moving through the EMF overall, and is primarily captured in Part 1 of the MAPLE surveys. The validity of Part 1 will be discussed in future publications. Secondly, as noted previously, reliability and discrimination metrics are also important aspects of validity; these will also be discussed in future publications after MAPLE has been deployed. And lastly, the external validity of MAPLE is limited by the range of perspectives, both student and expert, that contributed to its development process. The data collection discussed here relied on convenience sampling, so we do not claim that our samples of students and instructors are representative of the national population. Therefore, the use of MAPLE outside of the contexts in which it was designed should involve checks to see if its content is interpreted as intended and if its results yield meaningful discrimination and reliability.

In conclusion, we describe here the creation of a set of CMR REs for MAPLE, an RBAI measuring students' proficiency with the EMF. The development process, with input from students and experts, establishes the construct validity of the items. These CMR items probe reasoning within and between particular subtasks of the EMF in a scalable way. This reasoning is a relevant learning goal in upper-division labs, and an important aspect of physics laboratory learning overall. Thus, RBAIs such as MAPLE are essential to understanding and improving such courses, which in turn are an integral part of the physics undergraduate curriculum.

ACKNOWLEDGMENTS

Dimitri Dounas-Frazer and Jacob Stanley contributed significantly to earlier phases of this project. Bethany Wilcox provided valuable input on our REs. Alexandra Werth also provided such input, and conducted interviews. We also gratefully acknowledge our instructor partners and their students. Support from NSF under Grant Nos. DUE-1611868 and PHYS-1734006.

[1] L. Ding, Theoretical perspectives of quantitative physics education research, Physical Review Physics Education Research **15**, 020101 (2019).

[2] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource Letter RBAI-1: Research-Based Assessment Instruments in Physics and Astronomy, American Journal of Physics **85**, 245 (2017).

[3] A. Madsen, S. B. McKagan, E. C. Sayre, and C. A. Paul, Resource Letter RBAI-2: Research-based assessment instruments: Beyond physics topics, American Journal of Physics **87**, 350 (2019).

[4] AAPT Committee on Laboratories, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (Am Assoc Phys Teach, 2014).

[5] B. R. Wilcox and H. J. Lewandowski, A summary of research-based assessment of students' beliefs about the nature of experimental physics, Am. J. Phys. **86**, 212 (2018).

[6] J. Day and D. Bonn, Development of the concise data processing assessment, Phys Rev Spec Top-PH **7**, 010114 (2011).

[7] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the Physics Lab Inventory of Critical thinking (PLIC), Physical Review Physics Education Research **15**, 10135 (2019).

[8] B. R. Wilcox and S. J. Pollock, Validation and analysis of the coupled multiple response Colorado upper-division electrostatics diagnostic, Physical Review Special Topics - Physics Education Research **11**, 1 (2015).

[9] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Physical Review Special Topics - Physics Education Research **10**, 1 (2014).

[10] S. V. Chasteen, R. E. Pepper, M. D. Caballero, S. J. Pollock, and K. K. Perkins, Colorado Upper-Division Electrostatics diagnostic: A conceptual assessment for the junior level, Physical Review Special Topics - Physics Education Research **8**, 020108 (2012).

[11] 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).

[12] 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, Physical Review Physics Education Research **14**, 020118 (2018).

[13] L. Ríos, B. Pollard, D. R. Dounas-Frazer, and H. J. Lewandowski, Pathways to proposing causes for unexpected experimental results, in *Physics Education Research Conference Proceedings*, Vol. 2018 (American Association of Physics Teachers, 2018).

[14] D. R. Dounas-Frazer, J. T. Stanley, and H. J. Lewandowski, Instructor perspectives on iteration during upper-division optics lab activities, in *Physics Education Research Conference Proceedings* (American Association of Physics Teachers, 2018).

[15] 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, Physical Review Physics Education Research **15**, 010140 (2019).

[16] Qualtrics, Qualtrics (2005).

[17] A. N. Parks and M. Schmeichel, Obstacles to Addressing Race and Ethnicity in the Mathematics Education Literature, Journal for Research in Mathematics Education **43**, 238 (2012).

[18] S. Kanim and X. Cid, Demographics of physics education research, Physical Review Physics Education Research **16**, 020106 (2020).

[19] M. Wilson, Chapter 7: Reliability, in *Constructing Measures: An Item Response Modeling Approach* (Lawrence Erlbaum Associates, 2004) Chap. 7, pp. 139–154.

[20] M. Wilson, Chapter 8: Validity, in *Constructing Measures: An Item Response Modeling Approach* (Lawrence Erlbaum Associates, 2004) Chap. 8, pp. 155–180.

[21] M. T. Kane, An argument-based approach to validity., Psychological Bulletin **112**, 527 (1992).

[22] D. Borsboom, G. J. Mellenbergh, and J. van Heerden, The Concept of Validity, Psychological Review **111**, 1061 (2004).

[23] A. Maul, Rethinking Traditional Methods of Survey Validation, Measurement: Interdisciplinary Research and Perspectives **15**, 51 (2017).