

Preliminary development and validation of a diagnostic of critical thinking for introductory physics labs

N.G. Holmes¹, and Carl Wieman²

¹*Physics Department, Stanford University, 382 Via Pueblo Mall, Stanford, CA, 94041*

²*Graduate School of Education and Physics Department,
Stanford University, 382 Via Pueblo Mall, Stanford, CA, 94041*

Recent work has demonstrated the affordances of novel instructional design in physics labs to develop critical thinking skills. To guide broad institutional change, however, there is a need for efficient and validated assessment methods. In this paper, we present preliminary work on developing and validating a closed-response assessment of critical thinking skills, which focuses on how students reason with experimental data and test the validity of scientific models. The assessment asks students to reason about the methods, data, and analysis of two hypothetical case studies of students experimentally testing a model of Hooke's law. We describe the development and refinement of the cases and questions, and present preliminary results of validation studies.

I. INTRODUCTION

More than 400,000 undergraduate students enroll in introductory physics courses at post-secondary institutions in the United States each year [1]. In almost all of these courses, students spend time learning in lectures, recitations, and instructional laboratories ("labs"). Labs are often the most resource-intensive components, since they require specialized equipment, space, facilities, and occupy a significant amount of student and instructional staff time [2].

While much discipline-based education research has evaluated student learning outcomes in lectures and recitations, there is far less research on the learning outcomes from stand-alone, introductory instructional labs [2–6]. There exist many open research questions regarding what students are or could be learning from lab courses, what sorts of pedagogies support that learning, and how to measure that learning.

The goals of lab courses are diverse and without consensus [2, 3, 5–7]. Recent work has suggested that physics labs offer little added value to support conceptual mastery [2, 3, 7–10]. Labs offer unique opportunities, however, to engage students in experimentation skills, abilities, and habits of mind [11, 13]. There have been recent calls to shift the attention of laboratory instruction towards these skills and practices, such as from the American Association of Physics Teachers [11] and the Framework for K-12 Science Education and the Next Generation Science Standards [12]. This shift in instructional targets provides renewed impetus to develop and evaluate instructional strategies for teaching such skills and practices in labs [14]. Efforts to reform physics instruction have been greatly facilitated by shared assessments [15–17].

II. DEVELOPING THE SURVEY

In this paper, we describe preliminary development, refinement, and validation of an assessment of critical thinking for introductory physics lab courses.

A. Motivation

We define critical thinking in the context of physics labs as the ability to: "critique data, to identify whether or not conclusions are supported by evidence, and to distinguish a significant effect from random noise and variability." [18, p.1]. For students to do this successfully, they must also have a facility with (or understanding of) experimental design, measurement and uncertainty, and argumentation from data. Critical thinking can, thus, be seen as a high-level goal that encompasses many of the scientific practices and skills desired from effective science curricula [11, 12].

Our recent work has used rigorous coding of students' lab notebooks to evaluate students' critical thinking in different lab curricula [18, 19]. The Physics Lab Inventory of Critical thinking (PLIC) is designed to provide a more efficient assessment method that emulated this notebook coding.

The PLIC uses case studies of fictional student groups conducting a mass on a spring experiment to evaluate the Hooke's law model for simple harmonic motion: $T = 2\pi\sqrt{\frac{m}{k}}$. This Hooke's law experiment was chosen as one whose physics content, experimental methods, and apparatus would be relatively familiar to any introductory physics student. The experimental data used in the assessment were based on actual data collected by an expert conducting such an experiment. Questions were chosen to probe students' physical understanding of variability and systematic effects (rather than their procedural knowledge), their interpretation of collected data and associated methods, and their evaluation of a physical model in light of conflicting evidence.

B. Methods to validate the questions and overall structure

The general process of development and validation has followed recommendations from [16, 20] (see Fig. 1). An early version of the PLIC was designed with questions based on the progression of self-questions by an expert conducting the

experiment. The expert process involved collecting measurements of the period for a range of masses, linearizing and graphing the results, and revising the model based on the need for an additional parameter to adequately fit the data. Test questions included reflecting on the choice of methods, the collected data, to interpret the results in the light of the model, and to make decisions about how to proceed.

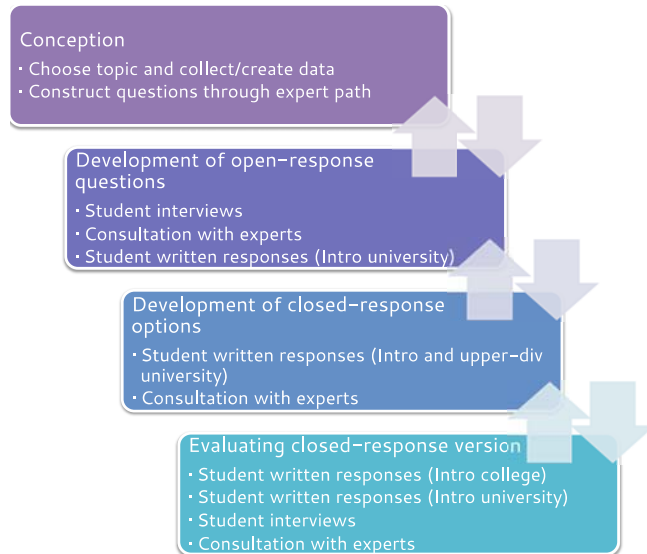


FIG. 1. Diagram summarizing the overall development process. Dual direction arrows represent the iterative nature of the process.

Think-aloud interviews were carried out with introductory and upper-division physics students using an open-response set of questions. Students’ interpretations, answers, and questions motivated significant revisions of the structure. For example, many interviewed students thought that “evaluating a model” meant finding the relevant constants (in this case, the spring constant, k). Due to this interpretation, a second fictional case study group was added. This group measures many repeated trials of the period for only two masses, then calculates and compares the spring constant, k , for each mass.

A revised open-response version of the PLIC was distributed at the start and end of an introductory physics lab course. Further revisions were informed by student responses. For example, a question about possible sources of uncertainty exposed that students conflate uncertainty, systematic effects, and measurement mistakes [21]. This question was revised to ask for sources of random variability and systematic effects separately.

The written responses also exposed issues with the data used by the two student groups. For example, the second student group measure the time for 50 periods for each mass and attach a timing uncertainty of 0.1s for each measured time (0.02s per period). Several students thought this uncertainty was unrealistically small and others said that no student would ever measure the time for 50 periods. In the revised version, both groups measure the time for five periods.

C. Constructing closed-response options

Open-response versions of the PLIC have been distributed to an introductory university physics course (pre- and post-instruction), an introductory community college physics course (pre- and post-instruction), and two upper-division physics lab courses (mid-course in both cases). Closed response items were constructed from responses by students in the university physics course and the two upper-division courses ($n=86$ test responses).

The open responses from a random selection of 19 students (pre and post) in the introductory community college course were used to test the closed-response options. The majority of the open responses were captured by the options created. Students wrote, on average, 18 responses on the whole test, with only one response per student being categorized as ‘other’. Half of the ‘other’ items were uninterpretable or irrelevant.

D. Developing the assessment format

While it was relatively straight forward to categorize students’ ideas into closed-response options, it became clear from students’ open-responses that a traditional “choose one multiple choice” format would be insufficient for this assessment. There were three key characteristics of student responses that motivated this decision. First, students listed multiple different ideas in response to each question (for example, multiple sources of uncertainty). Second, different students would describe the same features of the methods or data in either positive or negative terms. For example, some students would say that it was good that Group 1 measured multiple masses while other students would say that they did not measure enough. Third, there were paired aspects of responses. For example, a question first asks students how well Group 1’s k values agree, and then ask for the justification, which relates to their original evaluation of agreement.

A choose-many multiple choice format addresses the first issue of multiple ideas per student. Multiple choice options are listed in neutral contexts to address the second issue of positive and negative terms (e.g. “the number of masses”, rather than “many masses” and “few masses”). Finally, many questions are paired such that the first question asks a “what” question and the second asks for reasoning (e.g. “How well do you think Group 2 evaluated the model?” is followed by “Which items below best support your choice?”).

The analysis of the community college students, however, suggested that significant information was lost in choosing the ‘neutral’ options. For example, in a summary question where students are asked to compare the methods of the two different groups, students would respond with both pros and cons of the group selected (e.g. “They measured many masses but did not take enough repeated trials”). Having students select “number of masses” and “number of repeated trials” would not provide this distinction that one element was done well while the other was done poorly. Questions are now re-

structured for students to drag and drop the neutral reasoning elements into categories. An example of this structure is shown in Figure 2. Critical to this structure is to limit the number of items students can select. This is to require students to prioritize their reasoning and to also reduce time spent categorizing all of the options listed. This drag-and-drop structure will be evaluated in future interviews.

TABLE I. Number of student responses to open-response versions evaluated to date

Process	Course	Time	N
Think-aloud interviews	Intro university physics majors		2
	Upper-division university physics majors		4
Constructing closed-response options	Intro physics lab (University)	Pre	31
	Junior electronics lab	Post	30
	Advanced physics lab	Mid	13
Testing closed-response options	Intro Physics Lab (Community College)	Pre	19
		Post	19
		<i>Total</i>	<i>124</i>

Several delivery modes were evaluated including creative use of traditional multiple choice cards, automatic scoring of photocopied paper surveys (e.g. through Remark Office [25]), or online survey tools. Several instructors were polled for their preference and online survey tools (currently through Qualtrics [26]) were chosen to provide the flexibility needed.

E. Scoring the assessment

Figure 3 shows the fraction of students in each group who wrote a particular response to the question “What should the Group 1 students do next?” The figure is included here to demonstrate the vastness of students’ ideas in this regime. It is clear, however, that a number of ideas are raised by only a small number of students. Future student interviews and testing of closed-response versions will further inform whether these options should be discarded.

Responses from experts were used to characterize the most expert-like reasoning (indicated with a star). There are expert items that students do and do not discuss (measuring different numbers of masses and changing the number of bounces, respectively). There are some shifts between pre- and post-tests, as well as discrimination between introductory and more advanced students. While we do not aim to draw any conclusions from this preliminary data set, these data provide some insight to validity of the test in terms of distinguishing novices from experts and the impacts of instruction. Further analysis with all test questions will better evaluate these and other elements of test validity and reliability.

A preliminary scoring structure is currently being tested. The aim of the scoring will be to distinguish expert-like responses from novice responses, taking into account coupled

Which group do you think did a better job of investigating the model?

Group 1
 Group 2
 Both the same

Which items below best support your choice in 5a? Choose and categorize no more than 3 important ideas.

Items	Group 1 did it better	Group 2 did it better	Both groups did equally well
- The number of masses used			
- The size of the uncertainties (in k or between data)			
- The quality of the data analysis			
- How well they recorded data			
- How well the k values agree			
- The number of repeated trials for each mass			
- The methods to incorporate or reduce uncertainty			
- The attempts to change the model equation or try different fits			
- How well they measured or took into account different variables (e.g. amplitude, temperature)			
- The number of bounces of the spring per trial			
- How well the data agrees with the predicted model			
- Other (Please describe)			

FIG. 2. Sample of the new drag and drop structured for paired questions.

responses, and mimicking a partial-marks scoring rubric that would be applied to the open-response questions [22]. Work on traditional choose-one [24] and choose-many [22, 23] closed-response tests have generally shown high correlations with free-response rubric scoring or evaluations from student interviews. Choose-many scoring was even shown to provide better alignment with free-response coding than choose-one scoring [23], though this used ‘all or nothing’ scoring, where students would get zero on an item if they did not select all the correct responses. A more nuanced scoring scheme has been shown to align well with a validated partial-marks open-response scoring rubric [22].

Significant validation testing will need to be carried out to evaluate different scoring systems to, for example, evaluate test-retest reliability and understand effects of random guessing. We also need to ensure students can complete the assessment in a reasonable amount of time.

III. NEXT STEPS

We will soon evaluate the remainder of the open-response surveys collected from the university and college physics lab courses (over 200 remain). We will also conduct think-aloud interviews with students on the closed-response version. Fi-

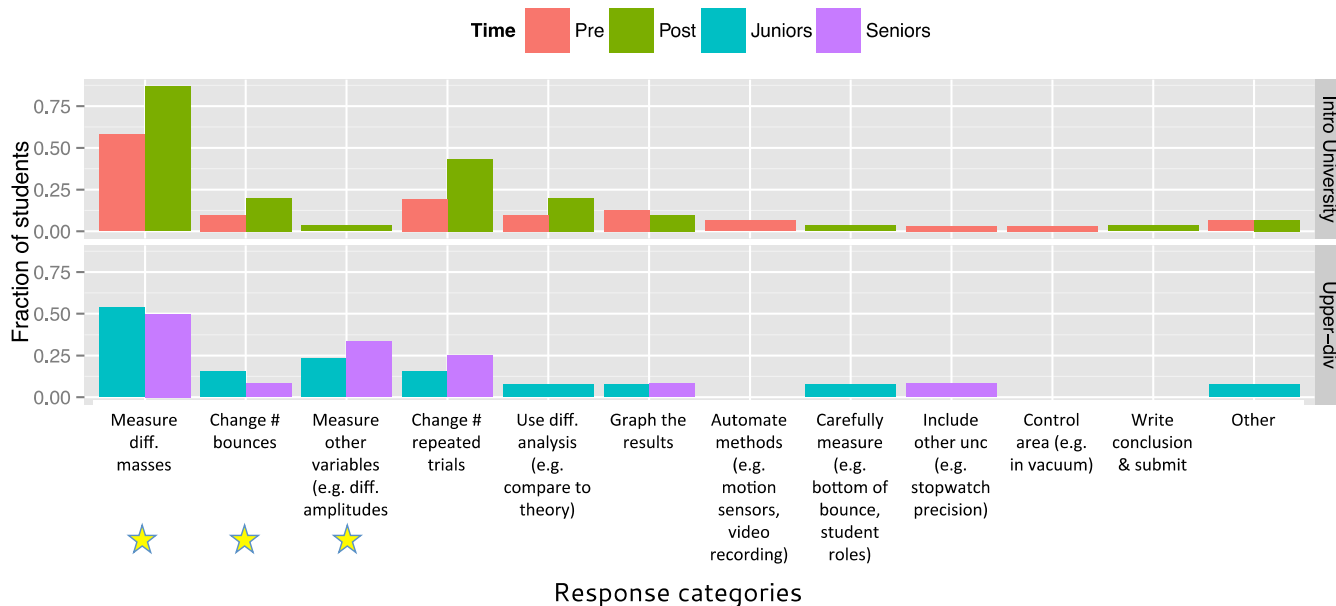


FIG. 3. Fraction of students selecting various responses to the question: “What do you think the group 1 students do next?” Expert responses are labeled with a star.

nally, we will elicit additional expert responses and refine the scoring system. We expect to have a closed-response beta version ready for validation in Fall 2016. We will collect data with revised closed- and free-response versions to engage in more thorough validity and reliability assessments, as briefly outlined above.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of P. LePage and D. Bonn in developing the survey, M. Stetzer and D. Marasco for testing the survey, and S. Senthilkumar for assisting in developing closed-response options.

-
- [1] P. J. Mulvey & S. Nicholson, www.aip.org/statistics. 2011.
 - [2] M. Séré, *Sci. Educ.* **86**, 5 (2002)
 - [3] S.R. Singer, M.L. Hilton, & H.A. Schweingruber, *America's Lab Report: Investigations in High School Science* (National Academies Press, Washington, D.C., 2005)
 - [4] S.R. Singer, N.R. Nielsen, & H.A. Schweingruber, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (National Academies Press, Washington, D.C., 2012)
 - [5] A. Hofstein & V.N. Lunetta, *Rev. Educ. Res.* **52**, 2 (1982)
 - [6] A. Hofstein & V.N. Lunetta, *Sci. Educ.* **88**, 1 (2004)
 - [7] R. Millar, *The role of practical work in the teaching and learning of science* (National Academy of Sciences, Washington, D.C., 2004)
 - [8] R. Trumper, *Sci. & Educ.* **12**, 7 (2003)
 - [9] C.E. Wieman & N.G. Holmes, *Am. J. Phys.* **83**, 11 (2015)
 - [10] C.E. Wieman & N.G. Holmes, in *PERC proceedings, College Park, MD, 2015*.
 - [11] American Association of Physics Teachers, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum*, aapt.org/Resources/ (2014)
 - [12] H. Quinn, H.A. Schweingruber, & T. Keller *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Academies Press, 2012)
 - [13] C.E. Wieman, *Phys. Teach.* **53**, 6 (2015)
 - [14] B.M. Zwickl, aps.org/units/fed/newsletters/spring2016 (2016)
 - [15] A. Madsen, S.B. McKagan, & E.C. Sayre, *Phys. Rev. ST-PER.* **11**, 1 (2015)
 - [16] A. Madsen, S.B. McKagan, & E.C. Sayre, Unpublished arxiv.org/abs/1605.02703 (2016)
 - [17] S. Freeman, S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, & M.P. Wenderoth, *PNAS.* **111**, 23 (2014)
 - [18] N.G. Holmes, C.E. Wieman, & D.A. Bonn, *PNAS.* **112**, 36 (2015)
 - [19] N.G. Holmes & D.A. Bonn, *Phys. Teach.* **53**, 6 (2015)
 - [20] W.K. Adams & C.E. Wieman, *Int. J. Sci. Educ.* **33**, 9 (2011)
 - [21] N.G. Holmes & C.E. Wieman, in *2015 BFY Proceedings, College Park, MD, 2015*.
 - [22] B.R. Wilcox & S.J. Pollock, *Phys. Rev. ST-PER.* **10**, 2 (2014)
 - [23] K.D. Kubinger, & C.H. Gottschall, *Psych. Sci.* **49**, 4 (2007)
 - [24] M. Scott, T. Stelzer, & G. Gladding, *Phys. Rev. ST-PER.* **2**, 2 (2006)
 - [25] <http://remarkssoftware.com/products/office/>
 - [26] <https://www.qualtrics.com/>