Assessment of critical thinking in physics labs: concurrent validity

Cole Walsh, Katherine N. Quinn, and N. G. Holmes
Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY, 14853

Despite the significant amount of time undergraduate students spend in introductory physics labs, there is little consensus on instructional goals and accepted diagnostic assessments for these labs. In response to these issues, we have developed the Physics Lab Inventory of Critical thinking (PLIC) to assess students' proficiency with critical thinking in a physics lab context. Specifically, the PLIC aims to evaluate students' skills in making sense of data, variability, models, and experimental methods and to assess the effectiveness of lab courses at developing these skills. Here, we discuss two parts of the validation process using 3374 student responses collected from 12 institutions during the 2017-2018 academic year. As a part of our validation process, we evaluate the concurrent validity of the instrument, namely, the impact of physics maturity and lab design on student performance.

I. INTRODUCTION

The goals of instruction in undergraduate physics labs typically involve reinforcing student conceptual knowledge of topics introduced in lecture and, to a lesser extent, teaching students how to work with experimental equipment [1]. Though research has shown that the instructional goals of labs are highly debated and vary across disciplines and institutions [2], the value of traditional labs in student learning of physics content has recently come into question [3]. There are, however, many important skills that a lab setting uniquely allows students to learn [4]. Particularly, proficiency in making sense of data, variability, models, and experimental methods are all skills that can be developed in physics labs [5]. These skills make up a recently endorsed set of recommended learning goals for undergraduate physics lab courses by the American Association of Physics Teachers [6].

With ongoing laboratory course transformations now occurring at multiple institutions in an effort to meet these new instructional goals, there is an increasing need for validated ways to measure student acquisition of these skills and few validated assessments exist. In response to this, we have developed the Physics Lab Inventory of Critical thinking (PLIC). Here, we demonstrate the concurrent validity [7]—that is, how consistent performance is with certain expected results—of the PLIC. We expect that either from instruction or selection effects, performance on the PLIC should be higher with greater physics maturity of the respondent. We define physics maturity by the level of the lab course that students were enrolled in when they took the PLIC. We also address the impact of lab courses that have undergone specific transformations to meet the goals outlined above, which we refer to as Critical Thinking labs (CTlabs) [4], on PLIC performance. This work is part of the continual validation assessment of the PLIC following steps laid out in Refs. [8] and [9]. Other validation studies of the PLIC have been published here [10] and here [11] with work ongoing.

II. THE PLIC

The PLIC is an online survey that presents respondents with a hypothetical scenario where two groups of physicists are completing a mass-on-a-spring experiment to test a model where the period of oscillation of the bouncing mass is $T = 2\pi\sqrt{\frac{m}{k}}$. The first group conducts 10 repeated trials for the period of oscillation for two different masses and uses the given equation to find *k* in each case. Group 2 conducts two repeated trials for the period of oscillation for 10 different masses, plots T^2 versus *m* and fits to a straight line with the intercept fixed at the origin (one-parameter fit). Finally, Group 2 attempts to fit a straight line with a free intercept (two-parameter fit).

Respondents answer questions on four pages: one page for group 1, two pages for group 2 (one with the one-parameter fit data, and the other adding the intercept to their fit), and one page comparing the two groups. The questions use a combination of Likert scale, traditional multiple choice, and *multiple response* questions. The Likert scale questions ask respondents to evaluate the data (how well the two spring constant values agree or how well data agrees with the fit line) and evaluate the methods (how well the groups' methods investigated the model). The traditional multiple choice questions ask respondents to decide which fit group 2 should use and which group did a better job evaluating the model. The *multiple response* questions ask respondents to elaborate on their reasoning to the Likert scale and traditional multiple choice questions (*reasoning* questions) and to suggest what the group should do next (*what to do next* questions). For each of these *reasoning* and *what to do next* questions, there are between 5-10 items to choose from, and students are limited to selecting no more than three items.

In developing a scoring scheme for the PLIC, we collected responses from 24 physics experts (post-docs, lecturers, and faculty) and used these responses to classify items into 3 categories: expert, partially-expert, and novice. There were 1-2 items per question that were selected by at least 50% of experts, which we labeled as expert (E) items. Furthermore, 1-2 items per question that were selected by more than 30% of experts (but less than 50%) were labelled as partially-expert (P). Finally, 2-3 items per question that were picked by less than 10% of experts were identified as being novice (N).

We then sought to create scores based on these item classifications that corresponded uniquely to well-defined levels:

- Respondents who select *at least one* E item on a question receive 1 point.

- Respondents who fail to select any E items, but select *at least one* P item, are awarded 0.5 points.
- Respondents selecting *at least one* N item have 0.25 points deducted from their score.
- All other items are considered neutral and have no impact on a respondent's score on that question.
- All scores are floored at zero.

Students do not get additional points for selecting more than one E or P item, nor do they get repeatedly deducted for selecting multiple N items. This scoring scheme allows respondents to obtain a maximum possible score of 1 on each question regardless of how many items they select; we allow students to select up to three items, but do not penalize for picking fewer. Students can also obtain full points by selecting different E items, reflecting the observed variability in expert responses. This scheme also provides credit for partial displays of critical thinking and differentiates students who answer in correct and partially correct ways from students who still have novice ideas about physics experimentation. On any question a student can receive a score between 0 and 1 in increments of 0.25. The PLIC's current format of 10 questions then allows for a maximum possible total score of 10 points.

The original 24 experts (upon whom the scoring scheme was based) obtained an average overall score on the PLIC of 8.61 ± 0.17. We subsequently received 48 additional responses from physics experts who scored 8.08 ± 0.18. Since these two sets of data are much closer to each other than to the student population (shown below), we combine them for later comparisons.

III. METHODS

Over the course of the 2017-2018 academic year, we collected PLIC responses before and after course instruction from 25 courses across 12 institutions. The institutions sampled are comprised of two four-year colleges, two Master's granting institutions, and eight Ph.D granting institutions that represent a combination of both public and private institutions. A total of 3374 responses were collected from students who completed the survey, consented to participate in the study, and indicated they were at least 18 years of age. Of these responses, pre- and post-responses were matched for individual students from the student ID or full name they provided at the end of the survey. In 2017-2018, we collected matched pre- and post-surveys from 1119 students (2238 total responses). However, at both pre- and post-test, 20% of students were randomly assigned open-response versions of the survey as part of the ongoing validation of the instrument. In the analyses that follow, we focus exclusively on students who completed a closed-response version of both the pre- and post-surveys, of which there were 816 students (1632 responses). Future work will examine the open-response version of the survey. The lab level and type (CTlabs or other) for each class were inferred from information provided by instructors about their course via a course information survey (CIS). The CIS is part of an automated system associated with the PLIC, which was adapted from Ref. [12].

To assess concurrent validity, we split our matched dataset by physics maturity and compared performance on the PLIC. This split dataset includes 672 students in first-year (FY) labs, 110 students in beyond-first-year (BFY) labs, and 34 students in graduate level labs. We also compared performance after splitting our matched dataset by lab type. This dataset includes 201 students who participated in FY CTlabs and 471 students who participated in some other type of FY physics lab. We examined students' responses to individual questions in detail to illuminate the differences in overall performance.

Pre- and post-survey scores for each group of students referenced in Table. I follow an approximately normal distribution with roughly equal variances across the groups. For this reason, we used parametric statistical tests to compare paired (paired *t*-test) and unpaired sample means (unpaired *t*-test), and a one-way Analysis of Covariance (ANCOVA) to evaluate the effect of lab treatment on post-scores with pre-scores as a covariate. We used Cohen's d for matched samples to calculate effect sizes between pre- and post-means, Cohen's d for independent samples to calculate effect sizes between pre-scores from different groups, and partial η^2 to calculate effect sizes of the independent variable and covariate on the dependent variable in ANCOVA.

IV. RESULTS AND DISCUSSION

A. Physics Maturity

We begin by comparing respondents' performance on the PLIC by physics maturity. In Table I, we report the average scores for students enrolled in different level physics lab courses, as well as for our 72 experts who only took the PLIC once. The significance level and effect sizes between pre- and post-mean scores within each group are also indicated.

Differences between pre- and post-instruction means are not statistically significant ($p > 0.05$) and effect sizes are negligible (Cohen's $d < 0.1$) across all three groups of students. Conversely, pre-instruction means are statistically different between all groups (unpaired *t*-test, $p < 0.01$) other than between students in BFY and graduate level labs ($p = 0.08$). The effect sizes ranged from small (Cohen's $d = 0.31$) between students in BFY and graduate labs to very large ($d = 1.6$) between students in FY labs and experts. The clear differences in means between groups of differing physics maturity, coupled with the lack of measurable increase in mean scores following instruction at any level, may imply that these differences arise from selection effects rather than cumulative instruction. This has been seen in other evaluations of students' lab sophistication as well [13].

We illustrate in Fig. 1 how the differences in physics maturity play out on one question from the pre-instruction PLIC. We use the pre-survey here since it allows us to examine students' thinking before instruction, and eliminates concern about students seeing the PLIC for a second time. We chose the question presented here because it clearly demonstrates the inherent differences in item selections by respondents from these different groups. In this question, respondents

TABLE I. Performance on the PLIC across different levels of physics maturity. N is the number of matched students within a dataset, except in the case of the expert surveys where respondents only filled out the survey once. Significance levels and effect sizes are reported for differences in pre- and post-means within each group of students.

	N	Pre Avg.	Post Avg.	p	d
FY	672	5.55 ± 0.07	5.70 ± 0.07	0.073	0.07
BFY	110	6.32 ± 0.17	6.36 ± 0.15	0.826	0.02
Grad	34	6.9 ± 0.2	6.8 ± 0.3	0.804	0.04
Experts	72	8.26 ± 0.14			

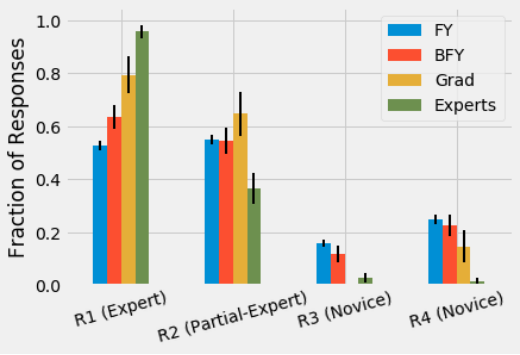

FIG. 1. Fraction of respondents that select a particular item in response to a question on the pre-instruction PLIC grouped by lab level. The question and responses R1-R4 are explained in the text. Error bars represent one standard error of the data.

were asked: "How similar or different do you think Group 1's spring constant (k) values are?" (Likert scale, not scored) and "what features were most important in comparing the two k values?" (*multiple response*, scored). The items that will affect a respondent's score, as well as their classifications, are:

- R1 (Expert) - the difference in the k-values compared to their uncertainties,
- R2 (Partial-Expert) - the size of the uncertainties (or the variability in the data),
- R3 (Novice) - the difference between the oscillation periods of the masses,
- R4 (Novice) - how they accounted for human error.

As expected, the expert answer to this question is picked increasingly more often with the physics maturity of the respondent. Additionally, the two novice answers to this question are picked less frequently with additional physics maturity of the respondent. The partial-expert item, however, is picked almost equally as often by all students. The fact that students equally value a partially correct answer at all levels could indicate that most intro to advanced physics lab courses have some focus on uncertainty.

Though we have described just one question in detail here, the performance differences are present across all of the questions on the PLIC. Experts score higher than students in BFY and graduate labs, and these students score higher than students in FY labs. As seen here, these performance differences are manifest in the expertness of responses.

TABLE II. Table of ANCOVA results for post-instruction scores.

	Effect	F	p	partial η^2
Post-scores	Pre-scores	25.53	< 0.001	0.037
	Lab Treatment	25.00	< 0.001	0.036

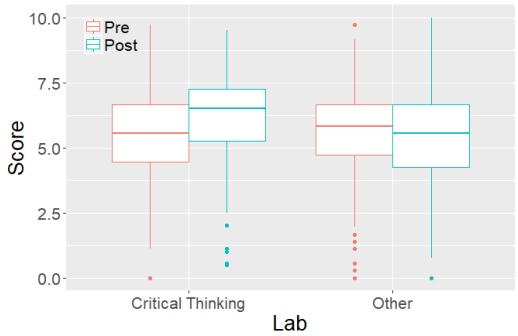

FIG. 2. Box plots of overall PLIC scores grouped by lab type.

B. Performance by Lab Type

We now examine how students participating in labs designed to meet the instructional goals of the PLIC performed in comparison to their counterparts in traditional lab settings. We limit our analysis to FY labs where we had three labs from two institutions that were best described as being of the CTlabs format, while the rest were of another format. The overall performance of these two groups of students is illustrated in Fig. 2.

We performed an ANCOVA comparing PLIC post-scores across lab treatments using pre-scores as a covariate (Table II). We see that, controlling for pre-instruction scores, lab treatment has a statistically significant impact on post-instruction scores with a small effect. For context, this effect is relatively larger than the effect size between students in BFY labs and graduate labs.

To see how these differences in overall scores arise, we looked at how students in different labs compared in their responses to one question. Fig. 3 shows the fraction of students

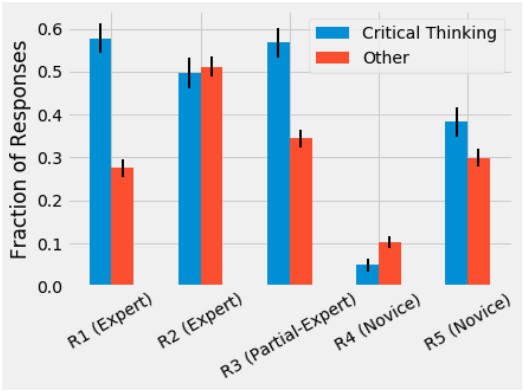

FIG. 3. Fraction of students that select a particular item in response to a question on the PLIC post-survey grouped by lab instruction. The question and responses R1-R5 are explained in the text. Error bars represent one standard error of the data.

who selected particular items in response to one of the questions on the PLIC post-survey grouped by lab type. We show only the post-instruction data since there was no statistically significant difference in pre-instruction scores on this question between the two groups. Again, we chose this question because it most clearly illustrates key differences in student thinking by respondents taught in different labs.

In this question, students are asked: "How similar or different do you think Group 2's data are from the new best-fit line?" (Likert scale, not scored) and "what features were most important in comparing the fit to the data?" (*multiple response*, scored). The items that will affect a respondent's score are:

- R1 (Expert) - the way points are scattered above and below the line,
- R2 (Expert) - how close the points are to the line compared to their uncertainties,
- R3 (Partial-Expert) - number of points with uncertainties crossing the line,
- R4 (Novice) - the number of outliers,
- R5 (Novice) - number of points above the line compared to the number below.

One of the expert items, R2, is commonly picked by students in both groups and is mostly unaffected by instruction. However, students in CTlabs favored the other expert item, R1, much more than students in other labs following instruction. Both R1 and R5 are concerned with the overall distribution of points about the best fit line. Thus, students taught in CTlabs appear to become more engaged with the importance of the distribution of residuals, albeit some students may oversimplify this to the sheer number of points above and below the line. Nonetheless, students taught in other labs maintain their interest in the closeness of the points to the best fit line and do not acknowledge the importance of the distribution of residuals any more than they did prior to instruction. Interestingly, students in both groups become less interested in the presence of outliers following instruction.

Again, though we have included only one question here for illustration purposes, students in CTlabs see, on average, increases in performance on all but one question, whereas the average scores for students in other labs increase on only 4 out of 10 questions.

V. LIMITATIONS AND FUTURE WORK

Our data contains only classes taught with a calculus-based curriculum. We are currently collecting data from several algebra-based courses at multiple institutions to explore differences in algebra-based and calculus-based instruction. We also intend to explore research questions related to the correlation of critical thinking skills, as measured by the PLIC, with other measures such as grades, concept knowledge, and attitudes towards science.

Future validation studies will use Classical Test Theory to further evaluate the validity and reliability of the instrument, including test-retest statistics and partial sample reliability. This reliability statistic will be useful in uncovering potential biases in our data, which may skew results and motivate the use of alternative data analysis methods such as imputation.

VI. CONCLUSIONS

With the need for large-scale instructional reform in physics lab instruction, there will be an equally important need for a method to evaluate these new instructional goals. Here, we have introduced one such method for measuring these goals, the PLIC. Though likely due to selection effects rather than instruction, respondents with a greater level of physics maturity perform consistently higher on the PLIC. Further, though we have seen no statistically significant shifts in performance following instruction for large cohorts of students, this is not the case for students enrolled in CTlabs designed to teach the skills that the PLIC is designed to measure. We have shown that CTlabs have a statistically significant and small effect on PLIC performance. These results establish one measure of validity of the PLIC and its usefulness in measuring the skills that we aim to teach in meeting the AAPT's new guidelines for physics labs in the future.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1611482. We would like to acknowledge Carl Wieman, members of CPERL for their useful feedback, and Heather Lewandowski and Bethany Wilcox for their support developing the administration system.

[1] R. Millar, *The role of practical work in the teaching and learning of science* (National Academy of Sciences, Washington, D.C., 2004).
[2] A. Hoffstein & V.N. Lunetta, *Sci. Educ.* **88**, 1 (2004).
[3] N. Holmes, *et al.*, *Phys. Rev. Phys. Educ. Res.* **13**, 010129 (2017).
[4] N.G. Holmes, *et al.*, *PNAS.* **112**, 36 (2015).
[5] E. Etkina, *et al.*, *Am. J. Phys.* **74**, 979 (2006).
[6] Kozminski, Joseph, *et al*. *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum.* (2014).
[7] B.R. Wilcox and H. Lewandowski. *Phys. Rev. Phys. Educ. Res.* **12**, 020132 (2016).
[8] W.K. Adams & C.E. Wieman, *Int. J. Sci. Educ.* **33**, 9 (2011).
[9] A. Madsen *et al. Am. J. Phys.* (2017).
[10] N.G. Holmes & C.E. Wieman, in *PERC proceedings, Sacramento, CA, 2016*.
[11] K.N. Quinn *et al.*, in *PERC proceedings, Cincinnati, OH, 2017*.
[12] B.R. Wilcox, *et al.*, *Phys. Rev. Phys. Educ. Res.* **12**, 010139 (2016).
[13] B.R. Wilcox & H. Lewandowski, *Phys. Rev. Phys. Educ. Res.* **13**, 023101 (2017).