Introductory physics laboratory practical exam development: Investigation design, explanation, and argument

Steven F. Wolf and Mark W. Sprague
Department of Physics, East Carolina University, C-209 Howell Science Complex, 10th Street, Greenville, NC 27858

Feng Li, Annalisa Smith-Joyner, and Joi P. Walker
Department of Chemistry, East Carolina University, 300 Science & Technology Building, 10th Street, Greenville, NC 27858

This study reports the development, validation, and implementation of a practical exam to assess science practices in an introductory physics laboratory. The exam asks students to design and conduct an investigation, perform data analysis, and write an argument. The exam was validated with advanced physics undergraduate students and undergraduate students in introductory physics lecture courses. Face validity has been established by administering the practical in 65 laboratory sections over the course of three semesters. We found that the greatest source of variability in this exam was due to instructor grading issues and discuss the implications of this result for our ongoing assessment efforts.

I. INTRODUCTION

As laboratory educators, we want all of our students to be able to design experiments to test a scientific idea, gather evidence, and then use that evidence to craft an argument to explain a phenomenon. In a nutshell, we want them to engage in science practices. Science practices have been increasingly emphasized in science education to promote students' scientific habits of mind, capabilities to conduct scientific inquiries, and engage in scientific reasoning [1–3]. Science laboratory courses are essential in undergraduate science curricula to improve students' science practices. At East Carolina University, we have been spending the past 2 years transforming the introductory physics laboratories to focus the curriculum on science practices.

The instructional approach that frames the transformation of East Carolina University's General Physics I laboratory curriculum is Argument-Driven Inquiry (ADI). ADI based science laboratory curricula emphasize the essential role of argumentation in science learning while promoting scientific inquiry [4]. ADI provides students opportunities to represent and communicate science with peers and receive feedback from peers as well. Students participating in ADI-based science laboratory courses were found to have better engagement in science learning and develop better scientific arguments [5, 6]. In General Chemistry laboratory courses, student performance on an end-of-course practical exam that required students to engage in investigation design, data collection, observation, inference, and argument construction revealed not only a significant positive difference for students in the ADI sections, but also a closing of the achievement gap for students from underrepresented groups in the ADI sections [7].

In our physics labs, we have been working to replicate the successes noted above using the ADI model. In order to do that, we have developed a new lab curriculum, which was first implemented in Spring 2018, and has undergone incremental changes each semester. We used the AAPT lab guidelines to inform our choices around lab content and appropriate assessment boundaries [8]. However, assessment of science practices is complicated in terms of its logistics and implementation. Over the same timeline we have been developing a practical exam that aligns to the assessment of science practices. This practical exam has been implemented at scale in all of our laboratory courses. In this paper, we outline the structure of this practical exam, describe the development process, and discuss the results of these assessments, focusing on how we used these results to address faculty concerns.

II. ASSESSMENT FRAMEWORK AND DESIGN

Scientific Practice describes an educational goal that students learn how to reason and act scientifically [9, 10]. As we designed the assessment, we utilized Ford's description of the nature of science practice as a basis for describing the introductory laboratory curriculum in terms of practices [11]. Ford described the "material practices" of science as having two distinct but complementary components:

Empirical Practices: practices related to manipulating nature to study aspects of it

Representative Practices: practices related to "making nature's behavior apparent" to peers [11, p. 408].

This dual nature of the practices of science is important to recognize, as it places a premium on the role of both the natural world and the community in the enterprise of science. Table I lists the empirical and representative practices described by Ford. We designed each item to assess at least one of these empirical or representative practices.

	Empirical practices:
EP1	Locate information relevant to a scientific problem.
EP2	Construct a relevant/appropriate scientific question for a given problem.
EP3	Design an experiment to test a scientific question.
EP4	Apply (or know when to apply) appropriate analytical methods to examine a scientific problem.
EP5	Appraise an experimental design to identify elements and limitations and how they impact scientific findings/conclusions.
EP6	Troubleshoot technical issues.
EP7	Evaluate evidence and critique experimental designs.
EP8	Interpret basic statistics (e.g., average and SD).
	Representative practices:
RP1	Generate a hypothesis or make a prediction based on a scientific model.
RP2	Construct an argument based on evidence.
RP3	Identify additional information needed to support an argument.
RP4	Provide alternative explanations for results that may have many causes.
RP5	Integrate and apply knowledge across sub-disciplines.
RP6	Represent data in a visual form.
RP7	Interpret visual representations of data.
RP8	Construct a Data table.
RP9	Data Analysis.

TABLE I. Ford's Emprical and Representative practices [11]. These provide a basis for describing the scope of our lab practical assessment.

On the exam day, students are given a set of masses, a tube and a stopper connected to a string, which slides freely through the tube. Students are also told:

> You notice that for a given rotation radius R, the stopper (mass m) travels faster as the hanging mass M increases. You want to determine the relationship between hanging mass (M) and period (T) for a given radius (R).

TAs demonstrate the use of this apparatus by holding the tube

Practical Item	Scientific Practices Assessed
1 - Procedure	EP3, EP4, EP5
2 - Data Table	RP8
3 - Claim	RP2
4 - Plot	RP6, RP9
5 - Argument	EP5, EP8, RP2, RP5, RP9

TABLE II. Practical Exam items and their assessment of empirical practices (EP) and representative practices (RP).

upright while swinging the stopper through the air in horizontal circles at (nearly) constant speed and radius. Students are also reminded of the format of the power law ($T \propto M^p$) and asked: ***Which power best describes the relationship between rotational period*** (T) ***and hanging mass*** (M) ***for a fixed radius rotation*** (R)***?*** Then the practical exam asks students to design an investigation, collect and analyze data, and write an argument to support a claim. Students design and carry out a procedure to measure the rotational period of the given system. This includes determining how many different hanging masses to use, and deciding how to estimate the uncertainty of the period measurement. Students also decide what to plot (such as T vs. M, T vs. $\frac{1}{M}$, or $\log T$ vs. $\log M$) and how to use that plot to generate and support their claim.

Specifically, our practical exam consists of 5 items that the students include in their report:

1. An experimental procedure
2. A data table
3. Their claim
4. A plot of T vs. M that best supports their claim
5. An argument to justify why the evidence they provided best supports the claim they made

Table II shows which scientific practices are assessed by each item. As a part of the first item we evaluate the methods that students use to measure the hanging mass M. For example, if students simply read off the mass values engraved on the hanging mass sets, they do not receive full credit as that is not the best tool that they have access to with which they can make mass measurements. (There is an electronic balance in the lab.) This assesses EP4 related to choosing the appropriate analytical method (tool) for examining a scientific problem. A detailed rubric has been developed describing how each of the items should be graded. We will not fully describe the content of the rubric further than this in order to protect exam security. However, the rubric is distributed in document form to all of the Graduate Teaching Assistants (GTAs) and faculty involved in the lab, and also embedded in the (Course Management System) CMS so that the GTAs use it to grade all of the practical exams.

III. DEVELOPMENT PROCESS

The practical exam described above went through a development process before its initial deployment. This included two stages: (1) running the alpha version of the practical exam with undergraduate students in the senior-level advanced physics laboratory course ("alpha testers") followed by interviews with participants, (2) running the beta version with undergraduate students in a second-semester calculus-based physics course. After the second stage, the GTA of the Physics II course was also interviewed after completing and administering the practical exam. The practical was also reviewed by physics faculty during the entire validation process to address their concerns (especially concerns about exam security and cheating).

The alpha test was composed of two parts. In the first part, alpha testers were asked to complete the practical exam given the same equipment and resources as we would make available to the introductory laboratory students. In the second part, individuals who had completed the practical exam, were recruited for interviews. We (FL and SFW) asked the participants how they figured out answers to each question in the practical exam and their suggestions about how to improve the practical exam. The practical exam developers (MWS and SFW) made revisions to the exam based on the performance on and feedback from the advanced laboratory students. Major revisions included removing a section due to time constraints, and providing suggested relationships between the dependent and independent variables that students are asked to test in the investigation. The relationships were provided to address a deficiency in the participants' data collection and analysis as well as feedback from the interviews with the participants. The students stated that they were confused about what to test in the investigation without suggested relationships. Some minor revisions were also made related to the structure and wording of the practical exam.

In the beta test, the revised practical exam was given to undergraduate students in second-semester calculus-based Physics II as a part of the recitation section. We reviewed the results of this group to ensure that the confusion experienced by the alpha test group was not present. We found that students created appropriate plots and text related to the question and were able to make a claim and show their evidence and reasoning supporting that claim. This established that students reliably understood what we were asking of them on this practical. The GTA who supervised the beta test was also interviewed for the feedback about suggested revisions based on the GTA's experience with teaching and assessing undergraduate students. The GTA was also interviewed about concerns on administering the practical exam in a classroom context. This led to changes such as our directing the TAs to demonstrate the use of the apparatus.

IV. IMPLEMENTATION AND RESULTS

This laboratory practical exam was given in the general physics laboratory, a laboratory course designed for students taking either the calculus-based introductory physics lecture or the algebra-based introductory physics lecture. In the analysis below, we do not attempt to tease out differences between these two groups as we are interested in the reliability and validity of this assessment with our students. Students in this course have generally come from science majors or mathematics as part of their natural science general education credit. Engineering students have not been required to take this lab. All general physics laboratory students in the Spring 2018, Summer 2018 (I and II), Fall 2018, and Spring 2019 semesters took this practical exam. In the analysis below we did not discuss results from the summer terms. These classes were much smaller than the typical class, and these students were generally not from the same population as the students in the Fall and Spring terms. Furthermore, we have excluded the Spring 2018 term as we piloted the transformed lab curriculum that semester in only one section. This paper does not discuss the relative performance of students in the traditional and transformed treatment, and we did not want the influence of this curricular change to interfere with the interpretation of these results. In the Fall 2018 term, the practical was given on all weekdays in 26 sections with a total of $N = 498$ students and 10 graders. In the Spring 2019 term, the practical was given Monday through Thursday in 18 sections with a total of $N = 358$ students and 7 graders.

Before the exam, students were given a generic prompt about the topic of the practical exam, in this case circular motion, so that they could review relevant terminology and equations. Also, students were allowed to bring their own laptop, as we did not have sufficient computers in the lab for each student to have one machine. Students took measurements in self-determined groups of two and were not allowed to speak to other groups. By necessity, students had to come up with a common procedure and share data (Items 1 and 2 above), but could not collaborate on other items. Students submitted their lab reports individually. All of this is consistent with how all lab reports had been generated and evaluated for the entire semester. Furthermore, students were aware that their submission would be checked for plagiarism and that plagiarized work would receive a reduced grade (down to a zero).

In order to ensure that all departmental stakeholders saw value in this new practical exam, we used the results below to answer the following questions/concerns brought forth by the faculty.

- Did the practical "get out" during the exam week (and beyond)?
- What are the biggest sources of variation?

The first question is reflective of a concern that students would cheat on the practical exam. Previously, the department went to great lengths to give the labs unique finals that changed every semester and were different between sections taught by different faculty/TAs. Collecting data to address this concern has been critical for changing this part of the departmental culture. To address this faculty concern during the lab practical, we took the added measure of keeping all copies of the practical exam in the laboratory. Given that a significant part of the task was collecting, analyzing, and representing data, we felt that "cheating" was extremely difficult by design.

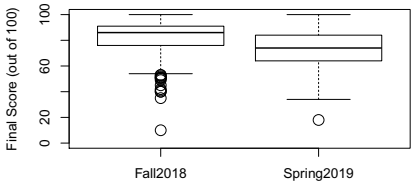

FIG. 1. Exam results by semester. Note that the mean score decreased from the fall to the spring.

In order to explore exam security, we first compare scores across semesters. Poor exam security is suspected if the mean score significantly increases. Figure 1 shows a box and whisker plot of exam scores. Visually, we can see that, in fact, exam scores decreased from the fall to the spring. There is a significant difference in scores from the fall 2018 semester ($\bar{x} = 82, s = 12$) and the spring 2019 semester ($\bar{x} = 73, s = 15$); $t(674.22) = 9.197, p < 0.001$. The source of this difference is the subject of future study. PER has a general notion that there may be differences between on-sequence and off-sequence cohorts, such as the one noted by Docktor and Heller [12], and often analyze these cohorts separately. This may be evidence supporting that notion, however, further study is clearly warranted as there may be other factors at play. For example, scores in the fall semester were high enough that some sections in the fall did not curve their practical scores as was the standard departmental practice. It is possible that some of the graders changed their application of the rubric in order to curb grade inflation. Regardless of the source of this variation between the terms, we analyze the fall and spring terms separately as we continue.

Another key concern is that students who took the practical early in the week might discuss the practical with their peers taking the practical later in the week, giving the later students an advantage. Figure 2 shows a box and whisker plot of exam scores split by day for each of these semesters. In Fall 2018 there is a significant effect of day administered on exam score at the $p < 0.05$ level for the 5 days $[F(4, 493) = 5.59, p < 0.001]$, with an adjusted $R^2 = 0.036$. However, we looked into this further, and decided to remove Friday's results from that part of the analysis. On Monday-Thursday, there were 6 lab sections/day, graded by (at minimum) 3 unique TAs. Friday had only 2 lab sections and was graded by the same TA. There is a significant difference in the scores for the set of all other graders ($\bar{x} = 81.2, s = 12.45$) and the grader that graded the Friday labs ($\bar{x} = 88.9, s = 8.5$); $t(74.9) = 5.76, p < 0.001$. We discuss grader variability more comprehensively in the next paragraph. With the Fri-

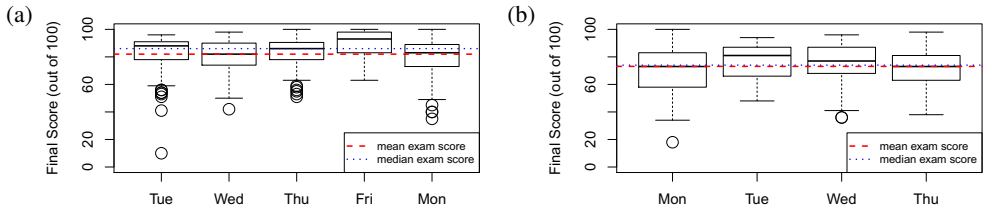

FIG. 2. Exam results by day for the (a) Fall 2018 semester and the (b) Spring 2019 semester respectively. The days are listed on the horizontal axis in the order that students took the practical. Note that in the Fall 2018 semester, students on Monday actually took the practical last as the semester ended on a Monday.

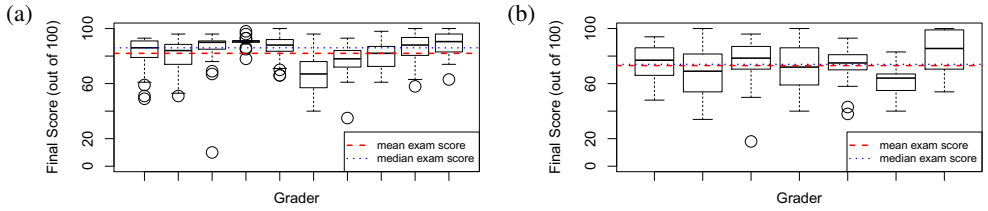

FIG. 3. Exam results by grader for the (a) Fall 2018 semester and the (b) Spring 2019 semester respectively.

day data removed there is no significant effect of day administered on exam score at the $p < 0.05$ level for the 4 days $[F(3, 462) = 2.18, p = 0.008]$, with an adjusted $R^2 = 0.016$. In the Spring 2019 semester, there is a significant effect of day administered on exam score at the $p < 0.05$ level for the 4 days $[F(3, 354) = 2.908, p < 0.035]$, with an adjusted $R^2 = 0.016$. A post-hoc analysis showed that students on Tuesday and Wednesday scored significantly higher than students on Monday. This difference is small enough that it could be due to calibration errors of the graders. In any case, these results do not support the idea that the practical was insecure as the students taking the exam on the last day did not have a significantly different score that the students taking the exam on the first day.

Since we do not observe any issues with exam security, we now discuss the question of exploring the variability of the graders. Figure 3 shows a box and whisker plot of exam scores for each grader in the Fall 2018 and Spring 2019 semesters. In the Fall 2018 semester, there is a significant effect of grader on exam score at the $p < 0.05$ level for the 10 graders $[F(9, 488) = 22.69, p < 0.001]$, with an adjusted $R^2 = 0.282$. In the Spring 2019 semester, there is a significant effect of grader on exam score at the $p < 0.05$ level for the 7 graders $[F(6, 351) = 6.514, p < 0.001]$, with an adjusted $R^2 = 0.085$. Understanding and quantifying grader variability is the greatest challenge to establishing reliability we have yet to fully answer. However, we have already begun to address this concern. In the Fall 2018, we did not have example lab practicals for TAs to grade as a method of calibration. In the Spring 2019, we did have the TAs calibrate with a small number of practicals from the previous semester. Visually, we see fewer outliers on the box and whisker plots suggesting this modest effort was met with some success. Still, we are developing a method for enhancing grader calibration for future semesters and quantitatively report on the success of these efforts in the future.

V. CONCLUSIONS

We have developed, validated, and implemented a practical exam to assess science practices in an introductory physics laboratory. In particular, the practical assessed empirical practices such as, "Design an experiment to test a scientific question," and representative practices such as, "Represent data in a visual form," and "Construct an argument based on evidence." This practical exam was given in high-enrollment introductory lab courses designed for mostly science majors.

This practical exam has been functionally unchanged from Spring 2018 to Spring 2019 without compromising the integrity of the assessment. Exam scores went down from Fall 2018 to Spring 2019, and did not significantly increase over the course of the week. In fact, the largest source of variation in student scores is due to the instructor/grader. To study this largest source of variation, we are considering the effects of implementation differences by the Graduate Teaching Assistants as part of the overarching lab transformation effort [13]. Traditional assessments are designed to limit instructor subjectivity, but do not provide a true measure of student ability to engage in science practices. This work addresses barriers to authentic assessment of science practices.

ACKNOWLEDGMENTS

This work was supported by NSF Award # 1725655.

[1] National Research Council *et al.*, *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas* (National Academies Press, 2012).

[2] S. R. Singer, N. R. Nielsen, and H. A. Schweingruber, *Discipline-Based Education Research* (National Academies Press, Washington, D.C., 2012).

[3] G. E. DeBoer, *A History of Ideas in Science Education: Implications for Practice.* (ERIC, 1991).

[4] J. P. Walker, V. Sampson, S. Southerland, and P. J. Enderle, Using the laboratory to engage all students in science practices, Chem. Educ. Res. Pract. **17**, 1098 (2016).

[5] V. Sampson, J. Grooms, and J. P. Walker, Argument-driven inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study, Science Education **95**, 217 (2011).

[6] J. P. Walker and V. Sampson, Learning to argue and arguing to learn: Argument-driven inquiry as a way to help undergraduate chemistry students learn how to construct arguments and engage in argumentation during a laboratory course, Journal of Research in Science Teaching **50**, 561 (2013).

[7] J. Walker, V. Sampson, S. Southerland, and P. Enderle, Using the laboratory to engage all students in science practices, Chemistry Education Research and Practice **17**, 1098 (2016).

[8] 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, American Association of Physics Teachers , 29 (2014).

[9] M. J. Ford, Educational implications of choosing "practice" to describe science in the next generation science standards, Science Education **99**, 1041 (2015).

[10] N. L. States, Next generation science standards: For states, by states (2013).

[11] M. Ford, Disciplinary authority and accountability in scientific practice and learning, Science Education **92**, 404 (2008).

[12] J. Docktor and K. Heller, Gender differences in both force concept inventory and introductory physics performance, AIP Conference Proceedings **1064**, 15 (2008), https://aip.scitation.org/doi/pdf/10.1063/1.3021243.

[13] A. Smith-Joyner, J. P. Walker, H. Hundley, M. W. Sprague, and S. F. Wolf, Graduate teaching assistant fidelity of implementation in introductory physics laboratories. (2019), submitted to PERC 2019.