Using a research-based assessment instrument to explore undergraduate students' proficiencies around measurement uncertainty in physics lab contexts

Gayle Geschwind,[1,2] Michael Vignal,[1,2,3] Marcos D. Caballero,[4,5,6] and H. J. Lewandowski[1,2]

[1]*JILA, National Institute of Standards and Technology and the University of Colorado, Boulder, Colorado 80309, USA*
[2]*Department of Physics, University of Colorado, 390 UCB, Boulder, Colorado 80309, USA*
[3]*Department of Physics, Willamette University, 900 State Street, Salem, Oregon 97301, USA*
[4]*Department of Physics and Astronomy and CREATE for STEM Institute, Michigan State University, East Lansing, Michigan 48824, USA*
[5]*Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, Michigan 48824, USA*
[6]*Department of Physics and Center for Computing in Science Education, University of Oslo, 0315 Oslo, Norway*

(Received 30 May 2024; accepted 8 July 2024; published 31 July 2024)

Concepts and practices surrounding measurement uncertainty are vital knowledge for physicists and are often emphasized in undergraduate physics laboratory courses. We have previously developed a research-based assessment instrument—the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE)—to examine student proficiency with measurement uncertainty along a variety of axes, including sources of uncertainty, handling of uncertainty, and distributions and repeated measurements. We present here initial results from the assessment representing over 1500 students from 20 institutions. We analyze students' performance pre- and postinstruction in lab courses and examine how instruction impacts students with different majors and gender. We find that students typically excel in certain areas, such as reporting the mean of a distribution as their result, while they struggle in other areas, such as comparing measurements with uncertainty and correctly propagating errors using formulas. Additionally, we find that the importance that an instructor places in certain areas of measurement uncertainty is uncorrelated with student performance in those areas.

DOI: 10.1103/PhysRevPhysEducRes.20.020105

I. INTRODUCTION

Measurement uncertainty is a core concept in physics experiments, as all measured quantities have associated uncertainties. Knowledge of uncertainty and how it affects the interpretation of the outcomes from an experiment is crucial for both presenting results from experiments and understanding others' work [1]. The importance of measurement uncertainty has led to recommendations for including this topic in introductory science laboratory courses [2–4]. However, instruction in this area could often be improved, with students frequently struggling to understand many of the important aspects of measurement uncertainty, including error propagation, taking several measurements to get a distribution of results, and comparing measurements with uncertainty, even after taking a course emphasizing these topics [5–13].

To facilitate improved learning of measurement uncertainty in laboratory courses, we previously developed a research-based assessment instrument (RBAI), the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) [14,15]. SPRUCE was developed to measure student proficiency with measurement uncertainty practices along ten dimensions that were identified as important to undergraduate physics laboratory instructors. Other assessments have included components of measurement uncertainty [6,13,16–19], but SPRUCE is the first assessment that focuses solely on this topic in detail. Specifically, we designed SPRUCE for first- and second-year lab courses and it is to be given in a pre-post instruction format via an online survey platform. The group at the University of Colorado analyzes the results of the surveys and presents them to the instructor in an easily interpretable report, where their course's data are shown in comparison to aggregate data from other courses. These data can be used by the instructor to improve the course as well as by the researcher to learn about student understanding of measurement uncertainty.

Previously, we explored the development of SPRUCE, including conducting instructor interviews [14], writing SPRUCE items, and examining the validity of these items [15]. In the work presented here, we produce the first research results from an analysis of students' responses to SPRUCE from many courses and institutions to provide a broad landscape of student understanding in this area by answering the following research questions:

1. How proficient are students at demonstrating the practices and concepts of measurement uncertainty as a general topic and on subtopics such as comparing measurements or propagating uncertainties? Where do they excel and where do they need additional support?
2. How does instruction impact student proficiency with the practices and concepts of measurement uncertainty?
    - How does instruction impact this proficiency for students with different majors and genders?
    - How does the importance an instructor places on specific learning objectives about measurement uncertainty impact this proficiency?

We choose to investigate both major and gender for several reasons. First, because students' major and gender are often intertwined, choosing to explore one of these variables while ignoring the other could lead to inaccurate results for the variable included in the analysis [20]. Additionally, prior studies have noted correlations with both of these variables on students' scores on physics assessments, leading to recommendations for removing gender bias from both instruction and assessment [19,21,22]. Because an important goal for the creation of SPRUCE is data literacy for all students, not just physicists, we are interested in examining the correlations between students' major and their performance on SPRUCE.

To answer these questions, we present results from 1576 students enrolled in 31 courses at 20 institutions in which SPRUCE was administered during the Spring 2023 and Fall 2023 semesters. We first provide an analysis of postinstruction responses, including a deeper look into student reasoning along several of the different areas of measurement uncertainty measured by SPRUCE. We then analyze pre-post shifts, including the impact of students' major(s) and gender on the results. Finally, we examine, in detail, three separate areas of measurement uncertainty—sources of uncertainty, handling of uncertainty, and distributions and repeated measurements—including both results from statistical analysis and example student reasoning provided during think-aloud interviews.

II. BACKGROUND

A. Previous work on student learning of measurement uncertainty

Previous studies have explored students' handling of measurement uncertainty in undergraduate physics laboratory courses. For example, students frequently have misconceptions about uncertainty, some of which do not improve even postinstruction [6,7,13,16–18,23–25]. Further research has explored introductory students' challenges with measurement uncertainty and found that students frequently fail to report uncertainty and often cannot identify the primary source of error in an experiment; this work also determined that students struggle with proper use of significant figures [26]. In addition to these studies, much has been learned about students' use of measurement uncertainty through previous RBAIs on this topic. These misconceptions and the use of RBAIs are discussed in detail below.

The first of these RBAIs is the Physics Measurement Questionnaire (PMQ) [6,16]. This RBAI aims to examine students' ability with measurement uncertainty, specifically looking at repeated measurements and measurement comparison with uncertainty. It consists of multiple-choice questions followed by open-response questions allowing students to provide their reasoning for their multiple-choice selections. Nominally, only the open response is coded for the analysis. This makes it difficult to perform a large-scale administration of this RBAI. Through use of this survey, researchers in South Africa found that they could separate student thinking into two categories: pointlike and setlike reasoning. Students who fall into the point paradigm believe there is a "true" experimental value, whereas students who use setlike reasoning understand that experiments provide incomplete information about a measured quantity and all data must be combined to obtain a best value. Additionally, some students were found to use mixed reasoning (a combination of both pointlike and setlike reasoning). [5]. Many students in this study retained their pointlike views of experimental physics after instruction, and only about 20% of physics majors were found to exhibit a more setlike view of measurement uncertainty postinstruction [16].

A partial version of the PMQ was implemented at the University of Colorado, Boulder, to examine the impact of a transformed laboratory course. Researchers observed a shift from mixed reasoning to setlike reasoning after instruction for both the traditional and transformed courses. Very few students exhibited solely pointlike reasoning in any case (even before instruction), though mixed reasoning was not uncommon. They also found that the course transformation had a positive impact in that students shift toward more sophisticated reasoning in the transformed version of the course [27–31]. Overall, these student responses differed significantly from the student responses from South Africa, which could be ascribed to significant differences in the student populations surveyed. Thus, while the pointlike and setlike reasoning paradigm might be a useful classification scheme for student reasoning in some cases, it does not capture the full range of students' ideas and skills with measurement uncertainty.

Other work has built on the PMQ, using some of the same probes and adding additional fine-grained probes to determine student proficiencies, as well as examining the pointlike and setlike paradigm, specifically relating to data processing [24]. This work found that students face challenges with specific areas of measurement uncertainty, such as using the mean as the best approximation in a repeated measurement experiment. On this open-response assessment, students were often unable to articulate why the mean might be used beyond providing a definition of the mean.

Some PMQ probes were also recently used in exploring upper-division students' views of measurement and uncertainty at multiple institutions in the United States [32]. They found that while both introductory and advanced students frequently used setlike reasoning, advanced students often provided more sophisticated reasoning, especially on a question pertaining to two sets of data with different means and the same spread. Advanced students were more likely to correctly identify sources of uncertainty in this situation. Overall, very few students at any level discussed uncertainty as an inherent property of experimentation. The researchers conclude that, while advanced students do perform better than introductory students on some of the probes, there is still a significant amount of improvement that could be made in helping students master certain concepts in measurement uncertainty, especially related to the shapes of data distributions.

Another RBAI, the Laboratory Data Analysis Instrument (LDAI), was developed in Israel to assess first-year students' understanding of data analysis procedures. It consists of 30 multiple-choice and true or false questions that are contextualized in real laboratory reports. The four objectives of this assessment are that students should (i) understand the meaning of, and ways to calculate, measures of central tendency, (ii) understand the meaning of error and uncertainty, as well as how to compute this and distinguish between statistical and systematic uncertainties, (iii) be able to choose and decipher graphs, and (iv) understand regression lines and how to fit them [13]. The LDAI requires students to write an open-response explanation to accompany their choice of true or false questions in order to receive credit, which, as with the PMQ, presents logistical challenges in widespread administration due to the open-response nature of the assessment.

One implementation of the LDAI in Thailand found that introductory physics students, in particular, faced challenges in fully understanding uncertainty, while undergraduate students of all levels struggled with linear regressions, even after taking at least one laboratory course. However, first-year students performed significantly worse on this assessment than second- and third-year students, indicating that instruction does improve skills to some extent since first-year students have not had as much learning experience with data analysis [33].

Another important RBAI in the laboratory space is the Physics Laboratory Inventory of Critical Thinking (PLIC), a ten-question assessment designed to examine student learning in physics lab courses [17,18]. The PLIC is aimed at analyzing students' laboratory skills as a whole rather than focusing on measurement uncertainty specifically. It examines four skills: evaluating data, evaluating methods, evaluating conclusions, and proposing next steps. The PLIC shows that students have not fully mastered measurement uncertainty, including conflating systematic error, random uncertainty, and human mistakes [18]; this broad study includes matched pre-post responses from several thousand students at 29 institutions and includes both first-year and beyond-first-year courses. In a large-scale administration of the PLIC, there were no observed statistically significant shifts in performance from pre- to postinstruction. However, students enrolled in a lab course specifically designed to teach skills measured by the PLIC do show statistically significant improvements on this assessment [17], showing the importance of aligning laboratory instruction with the desired learning goals.

Finally, the Concise Data Processing Assessment (CDPA) was also developed to probe student ideas related to measurement uncertainty, focusing mainly on error propagation [34]. Research using this assessment has found curriculum-dependent student challenges dealing with measurement uncertainty. For example, many students excelled at questions involving measurement error in linear fits but struggled with questions involving power laws; an examination of the curriculum for the course in which students were surveyed noted an emphasis on the former and no instruction on the latter [35].

Further, the CDPA was used to investigate gender gaps in physics [19]. The CDPA did reveal a significant gender gap at both the pre- and post-test levels. While all students did improve on the CDPA postinstruction, the gender gap remains unchanged: men still outperform women. The authors posit that one reason women do worse on the CDPA is due to a lack of confidence due to previous work showing that women generally report lower confidence in themselves in terms of their physics knowledge than their male counterparts.

SPRUCE fills a void in the current space of RBAIs in that it is easily administrable and scorable on a large scale, aimed at introductory laboratory students, and focuses solely on measurement uncertainty. No other RBAI currently meets all of these goals. For example, the PMQ and LDAI require open-response text, meaning that scoring them for thousands of students is too laborious. The PLIC and CDPA do have some questions about measurement uncertainty, but this is not the sole focus of these assessments. Further, the CDPA is more appropriate for higher-level students than introductory.

Research about students' understanding of measurement uncertainty also exists outside of the space of RBAIs.

One study found that students enter university courses frequently believing they must take exactly three measurements. Postinstruction, this belief was often corrected in that students understood that three trials may not always be sufficient. However, many students did not improve from pre- to postinstruction in other areas, including an understanding of the importance of reporting uncertainties and using uncertainty to determine whether measurements agree with one another [23].

Another study determined that even postinstruction, students tend to establish a hierarchy of measurements and do not fully understand the need to take several measurements. Instead, students judge their first measurement as the most important and use subsequent measurements as a check of their first one. They are also unable to distinguish between random and systematic errors. Students tend to state that the more measurements that a person makes, the better the result is, without fully understanding how or why more data are better [25].

Most of this prior work examines student proficiencies with measurement uncertainty in the context of nonquantum courses, but the performance of students in quantum courses is also important to investigate and is the subject of several recent papers. For example, one collaboration between researchers at Cornell and California State University, Fullerton, looked at student responses to questions about measurement uncertainty in both classical and quantum contexts and found that an updated definition of the pointlike and setlike paradigm might help advance understanding of student views on this topic, especially due to the binary nature of this paradigm and the prominence of mixed-reasoning among students. This research also indicates that instructors need to clarify the meaning behind "more data are better" so that students can understand when, exactly, this is true. The researchers also noted in a similar vein that a clearer discussion of standard deviation and standard error might help students with differentiating these quantities. Further, they found that students often conflate quantum uncertainty (e.g., the Heisenberg uncertainty principle) with measurement uncertainty in quantum mechanical experiments and, therefore, care should be taken in advanced laboratory courses to help students distinguish between these concepts [36].

Other work related to this collaboration has shown that in classical physics, students often state the limitations of the experimental setup as the major cause of uncertainty, while in quantum mechanics, students often explained measurement uncertainty as related to the principles of the physics theory underlying the experiment, as well as statistics [37]. The researchers concluded that there is a split in student reasoning about classical and quantum experiments, and instructors should work to bridge this gap by providing additional instruction about the relationships between experiment, measurement uncertainty, and theory in courses at all levels, especially because statistical limitations and experimental setup limitations affect both quantum and classical experiments.

Other research has highlighted the benefits of using the term "uncertainty" instead of "error" when describing measurement variability. They posit that using "error" might be a cause of students' pointlike reasoning, as it has a connotation of making mistakes rather than uncertainty as an inherent aspect of measurements [5]. Further, using uncertainty to describe inherent limitations and random variability, systematic effects to describe assumptions or approximations, and measurement mistakes to describe actual human errors might further aid student understanding of these concepts [7].

Overall, many prior studies have illuminated student strengths and weaknesses surrounding measurement uncertainty. We aim to add to this growing body of research by presenting results from SPRUCE.

B. SPRUCE

1. General overview and development

SPRUCE is an RBAI centered on measurement uncertainty and was designed to be administered pre- and postinstruction. Previous work has commented on the development, format [14,15], and validation [38] of SPRUCE, though a brief summary is contained below.

SPRUCE is a fully online assessment that takes students about 19 min to complete (median[1] = 1120 s). It consists of 19 items in a variety of formats, including multiple choice, multiple response, numeric open response, coupled multiple choice, coupled multiple response, and coupled numeric open response.

SPRUCE was developed using an adaptation of evidence-centered design [39], beginning with the researchers conducting interviews with introductory laboratory instructors to determine which areas of measurement uncertainty they find important. Based on these interviews, we created assessment objectives for SPRUCE [14]. Assessment objectives, or AOs, are "concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment [40]:" they are statements that are easy to directly assess in such a survey.

These SPRUCE AOs were then refined during the process of writing and revising SPRUCE itself. Table I shows the final AOs for SPRUCE after iteration and refinement. They are divided into three categories—sources of uncertainty, handling of uncertainty, and distributions and repeated measurements—and cover a wide variety of measurement uncertainty concepts, while still maintaining a cohesive thematic structure to be able to target them in

[1]Median is used here to remove effects from students who leave the assessment open on their computers for multiple days, heavily skewing the mean and making it an inappropriate statistic to report.

TABLE I. SPRUCE assessment objectives, organized by assessment objective category.

	Sources of uncertainty
S1	Estimate size of random/statistical uncertainty by considering instrument precision
S2	Identify actions that might improve precision
S3	Identify actions that might improve accuracy
	Handling of uncertainty
H1	Propagate uncertainties using formulas
H2	Report results with uncertainties and correct significant digits
	Distributions and repeated measurements
D1	Articulate why it is important to take several measurements during experimentation
D2	Articulate that repeated measurements will give a distribution of results and not a single number
D3	Calculate and report the mean of a distribution for the best estimate of the measurement
D4	Appropriately use and differentiate between standard deviation and standard error
D5	Determine if two measurements (with uncertainty) agree with each other

one assessment. All items (i.e., questions) on SPRUCE probe at least one of these AOs. More details about the validity of these AOs in relation to SPRUCE items are covered in other work [38].

2. Scoring

SPRUCE is scored using couplet scoring, a scoring scheme discussed at length in a previous paper [41]. Briefly, this scheme first identifies which AOs an item aims to measure. It then scores the responses to the item based on that AO only. The score for that one AO on one item is called a item-AO couplet. Items may address one or more AOs and thus have one or more scored item-AO couplets. Items that addresses multiple AOs will be scored multiple times, and the method of assigning points based on students' responses might differ for each couplet.

An example item and scoring scheme are shown in Fig. 1 and Table II, respectively; this example has been highlighted in previous papers [38,41]. In this item, we address two different AOs on SPRUCE: *H1—Propagate uncertainties using formulas* and *H2—Report results with uncertainties and correct significant digits*. Students need to answer this multiple-choice item only once, but we are able to draw conclusions about proficiencies along two different axes (i.e., AOs) from their answers. The scoring scheme itself is provided in Table II. This example illustrates how, for one item, multiple answers might be scored as correct depending on the AO and what answer is considered correct depends on what AO is being scored.

For AO H1 [propagate uncertainty], the answers that are given credit are those where students have appropriately propagated error, in this case dividing by 20. Thus, options A, C, and E present choices where students have shown proficiency in error propagation and receive credit for couplet item 3.3—AO H1. On the other hand, for AO H2 [significant figures], the answers that are given credit are those where students have provided an answer with correct significant figures. In this case, the answer options with matching decimal places in the result and uncertainty are options C and F, so students selecting either of those would receive credit for couplet item 3.3—AO H2.

Students only answer this item once. If they pick the "correct" overall answer, which is option C, they would receive credit on both couplets. However, they can receive credit on one couplet, but not the other, by providing other

You and your lab mates decide to measure 20 oscillations at a time. Using a handheld digital stopwatch, you measure a time of 28.42 seconds for 20 oscillations. You estimate the uncertainty in your measurement of 20 oscillations to be 0.4 seconds, based on an online search for human reaction time. What value and uncertainty do you report for the period of **a single oscillation**?

◯ 1.421 ± 0.02 s ◯ 1.42 ± 0.02 s ◯ 1.4 ± 0.02 s
◯ 1.421 ± 0.4 s ◯ 1.42 ± 0.4 s ◯ 1.4 ± 0.4 s

FIG. 1. SPRUCE item 3.3 (with alternate numbers to protect test security), in which students are attempting to determine the period of oscillation for a mass hanging vertically from a spring. This single item addresses two AOs, H1 and H2, which handle error propagation and significant figures, respectively.

TABLE II. Example scoring for couplets of item 3.3, showing how one multiple-choice item results in information about two separate measurement uncertainty topics based on the different answers students might give.

		Score	
Answer option		H1	H2
A	1.421 ± 0.02 s	1	0
B	1.421 ± 0.4 s	0	0
C	1.42 ± 0.02 s	1	1
D	1.42 ± 0.4 s	0	0
E	1.4 ± 0.02 s	1	0
F	1.4 ± 0.4 s	0	1

TABLE III. AO couplets and score options. Each AO is targeted by different numbers of couplets and therefore has different total possible scores. Some AOs offer partial credit, which is then rounded to the nearest integer after summing all couplet scores for that AO, such that all final AO scores are integers.

	Number of couplets	Possible scores, before rounding	Possible scores, after rounding
S1	3	[0, 1, 2, 3]	[0, 1, 2, 3]
S2	5	[0, 0.25, 0.50, 0.75, …, 5]	[0, 1, 2, 3, 4, 5]
S3	4	[0, 0.25, 0.50, 0.75, …, 4]	[0, 1, 2, 3, 4]
H1	4	[0, 1, 2, 3, 4]	[0, 1, 2, 3, 4]
H2	3	[0, 1, 2, 3]	[0, 1, 2, 3]
D1	2	[0, 0.25, 0.50, 0.75, …, 2]	[0, 1, 2]
D2	2	[0, 0.25, 0.50, 0.75, …, 2]	[0, 1, 2]
D3	2	[0, 1, 2]	[0, 1, 2]
D4	4	[0, 1, 2, 3, 4]	[0, 1, 2, 3, 4]
D5	2	[0, 1, 2]	[0, 1, 2]

answers, or they receive no credit if they select answer options B or D. In this way, we can separate student proficiencies in two different areas of measurement uncertainty by scoring along these axes to obtain information about them separately. All items in SPRUCE are scored according to these conventions, by first aligning the items with AOs and then scoring items as couplets. This leads to 31 item-AO couplets scored on SPRUCE from its 19 items. These couplet scores are then treated similarly to conventional item scores on a traditional assessment, in that they form the base unit of scoring.

After all of the couplets are scored, we combine them to create ten different AO scores: one score for each AO on SPRUCE. These AO scores are obtained by simply adding up all of the couplet scores pertaining to each AO for each student. We then round these scores to the nearest integer using typical rounding conventions (i.e., 0.5 rounds up), as some couplets allow for partial credit, as discussed in our prior paper [38]. These integer scores are reported as the AO-level scores; in some analyses, we normalize these to 1 by dividing by the number of couplets in each AO for easier comparisons. Typically, in reporting raw scores, these are normalized to 100, whereas when we perform other statistical analyses (such as ordinal logistic regression), we keep these as non-normalized integers. Table III shows the number of couplets and possible scores for each AO on SPRUCE; this table also shows that several couplets on SPRUCE allow for partial credit in increments of 0.25, rather than simply 0 or 1 as scores.

In order to calculate one overall test score on SPRUCE, we add the normalized AO scores together and then normalize this overall score to 100. Although the AO scores provide more fine-grained information than one single overall score, we still provide an overall score as a measure of student proficiency in measurement uncertainty as a whole, which is helpful for instructors and interesting from a research perspective. We also used this overall score in validating SPRUCE via classical test theory [38]. This method of calculating the overall score (using the normalized AO scores) weights each AO equally, rather than weighting each couplet equally, in order to remove biases from some AOs that are sampled more than others. By weighting each AO equally, we produce a final score that accounts equally for all ten areas of measurement uncertainty and is therefore a good measure of overall student proficiency with measurement uncertainty and is also consistent with instructor expectations. We acknowledge that this scoring scheme is complex, but this method provides instructors and researchers with a rich set of data about student performance at different grain sizes.

III. METHODS

A. Data collection and cleaning

We collected data from 31 physics laboratory courses at 20 institutions in the United States during the Spring 2023 and Fall 2023 semesters (see Table V for details on these institutions). Of the courses, 23 were introductory (accounting for $1379/1576 = 87.5\%$ student responses) and 8 were beyond introductory (accounting for $197/1576 = 12.5\%$ student responses). Courses were solicited via the authors' contacts, as well as through posting advertisements on the Advanced Laboratory Physics Association (ALPhA) listserv and two American Physical Society (APS) discussion boards (Forum on Education and Topical Group on Physics Education Research). Student demographics, including gender, race, and major, are presented in Table IV. We note that these demographics, which represent the 1576 matched pre-post responses, are representative of the full sample of completed post-test responses. For both gender and race, students were able to select as many options as they wanted to from the multiple-response question (including a not listed text box option), and therefore, these numbers in the table do not add up to 100%. Additionally, the population of students who participated in SPRUCE is not reflective of the current racial makeup of

TABLE IV. Student demographics: Race, gender, year, and major [$N = 1576$]. Because all demographic questions except year in school allow multiple responses and because these questions were optional, the numbers will not add up to 100%.

	Number of students	Percent students
Gender		
Man	905	57.4
Woman	614	39.0
Nonbinary	52	3.3
Not listed	9	0.57
Race		
White	1206	76.5
Asian	248	15.7
Hispanic/Latino	138	8.8
Black	59	3.7
American Indian or Alaska Native	19	1.2
Native Hawaiian or other Pacific Islander	11	0.70
Not listed	37	2.3
Year in school		
First year	477	30.3
Second year	551	35.0
Third year	320	20.3
Fourth year	170	10.8
Fifth year	31	2.0
Sixth year or beyond	14	0.89
Major		
Engineering	606	38.5
Physics	204	12.9
Biology	184	11.7
Computer science	127	8.1
Math or applied math	99	6.3
Astrophysics	96	6.1
Biochemistry	92	5.8
Chemistry	71	4.5
Engineering physics	45	2.9
Astronomy	33	2.1
Geology or geophysics	23	1.5
Physiology	23	1.5
Other science	172	10.9
Nonscience major	41	2.6
Open option/undeclared	36	2.3

the United States, which is a limitation of the data we hope to address with future data collection. In particular, black students are underrepresented. Additionally, major and year in school are correlated and, therefore, cannot be independently included in our analysis.

We collected 3733 total pretest responses and 2710 total post-test responses for a total of 6443 total responses. We then removed responses based on the following conditions. First, students who did not consent to having their data used for research were excluded, resulting in a loss of 691 responses (10.7%). Second, students who either did not answer the filter question (i.e., closed the survey before reaching that question) or answered the filter question incorrectly were excluded; this step removed a total of 1153 of the 6443 responses (17.9%). The filter question is placed after three of the four experiments on SPRUCE, ensuring students have answered at least 11 of the 19 items and therefore are scored on at least 21 of the 31 couplets and asks students to enter a specific three-digit number into a text box to ensure they are reading the questions. Finally, in order to examine the impact of instruction, we matched students using their student names and ID numbers to have matched pretest and post-test responses for students. If students took only one of these (either only the pretest or only the post-test), their results were excluded. Thus, we present an analysis of 1576 matched pre-post responses from the two semesters of data collection or about 48.9% of total responses to SPRUCE in that time frame.

TABLE V. Institution information [$N = 20$] including highest degree offered and minority-serving status. HSI indicates a Hispanic serving institution and AANAPISI indicates an Asian American and Native American Pacific Islander serving institution.

	Number of institutions
Highest degree	
Ph.D.	6
Master's	5
Bachelor's	8
Associate's	1
Minority serving status	
HSI	4
AANAPISI	1

We choose to use matched data in order to more accurately report changes from pretest to post-test. However, as previously stated, the demographics of the unmatched data do not differ significantly from those in the matched dataset, and the average scores (at both the overall score level and the AO level) also do not change significantly when comparing these datasets. Thus, this provides evidence that none of the analysis methods used, which are described in further detail below, are biased by including only matched data.

We also conducted student interviews during the Fall 2022 semester while SPRUCE was in beta testing, and some of these interview data are used in the work presented here. These 27 interviews each lasted approximately 1 h. Students were solicited for interviews from courses in which SPRUCE was currently being piloted. During these think-aloud interviews, students took SPRUCE while sharing their screen with the interviewer and were asked to explain their reasoning for each item to which they responded. These interviews provided evidence of student reasoning for each answer option, both correct and incorrect [15].

Finally, we collected information about each course from instructors who participated in SPRUCE administration, including the goals of the course, the level of the course, and the importance they place on different aspects of measurement uncertainty. In particular, instructors were asked to evaluate the importance of each of the AOs on a five-point Likert scale (extremely important, very important, moderately important, slightly important, and not at all important) for their course. In our analysis, we collapse these responses to a three-point scale, where extremely and very important are combined, and slightly and not at all important are combined. One limitation of our data is that *D4: Appropriately use and differentiate between standard deviation and standard error* existed in a different form in prior iterations (three other AOs were collapsed to form this one), and therefore, we have no data about instructor emphasis on this AO. These three AOs were collapsed as we determined that we could not accurately differentiate the original three AOs that handled this topic. D4 is treated separately in certain sections of this work in order to account for this change.

B. Analysis methods

To answer our first research question regarding students' overall proficiency with measurement uncertainty, we analyze postinstruction data only. We use only matched postinstruction responses to maintain a single student population for the entire paper, though results with all post-test responses are similar. Here, we examine the student scores on each AO and their overall scores on the assessment. AO scores are calculated by summing the couplet scores for each AO, rounding to the nearest integer, and normalizing to 1 (here, we normalize the AO-level scores to 1 to allow easier comparisons between AOs). Finally, the overall score is the sum of these normalized AO scores, also normalized to 1. The scoring processes are detailed further in Sec. II B 2 and in previous work [41].

Statistically, with post-test data, we report normality statistics in the form of the Anderson-Darling test, as well as skewness (the third moment) and kurtosis (the fourth moment). The Anderson-Darling test can detect whether data are normally distributed and, rather than a binary outcome, provides a significance level that gives information about the degree to which the data presented are normal [42,43]. Skewness is a measure of the asymmetry of a distribution and kurtosis is a measure of the tails of the data compared to a normal distribution. Normally distributed scores allow easier analysis methods, as many methods assume normality.

To answer one component of our second research question, regarding the impact of instruction, we look at the significance of the shifts from pretest to post-test scores both at the level of the overall score and at the level of each individual AO, with scores calculated as described above. We perform a Wilcoxon signed-rank test [44] to compute this significance. This is a nonparametric test of the null hypothesis that for randomly selected scores from two populations (in this case, the pretest and the post-test scores are the two different populations), the probability of one being greater than the other is the same as the reverse. It can be considered a nonparametric version of the dependent t test. In this case, because we are comparing populations that are not expected to be equal (assuming instruction has an impact on student performance on SPRUCE), we anticipate that the null hypothesis will fail—that is, we would expect that the distributions of these two groups are not identical.

In order to determine how much of an impact instruction has, we also utilize Cohen's d as a measure of effect size [45]. Effect size is a measure of the magnitude of the shift, as opposed to the Wilcoxon signed-rank test, which simply indicates whether the shift is statistically significant (as a binary).

Generally, unless otherwise noted, uncertainties are given as standard errors throughout this paper (68% confidence interval).

1. Analysis of covariance

Another component of our second research question requires analysis of student performance on SPRUCE overall (using the post-test overall score) and how this correlates with students' pretest score, major, and gender. To do so, we use a two-way analysis of covariance (ANCOVA), due to the relatively continuous nature of overall scores (as opposed to the small integer-only nature of the non-normalized AO scores). ANCOVA decomposes the dependent variable's variance into a part explained by the covariate, a part explained by the independent variables, and a residual variance [46,47]. In our case, our dependent variable is post-test overall score, our categorical independent variables are student major and gender, and our covariate is the pretest overall score. Using ANCOVA, we can explore whether student major or gender is correlated with post-test performance on SPRUCE, while controlling for pretest performance. We conducted this analysis in both PYTHON and R to validate the results were the same with two different statistics packages, and we find the results to be in agreement with one another. We note that interaction terms might be significant, and these are addressed in more detail in Sec. IV.

The general model we implement for ANCOVA is

$$S_{\text{post}} = \beta_0 + \beta_1 S_{\text{pre}} + \beta_2(\text{Gender}) + \beta_3(\text{Major}), \quad (1)$$

in which we relate the post-test to a student's score on the pretest, their major, and their gender. The β coefficients give the relative importance of each of these factors. Additionally, we obtain information about the amount of variance explained by each of these predictors in the form of partial η^2.

ANCOVA has several assumptions that must be met in order for it to be an appropriate statistic to use. The details of these assumptions and our data's adherence to them are discussed in Appendix A.

We note a limitation in the data regarding gender. SPRUCE includes a multiple-response item that asks students to report their gender. We received responses from 897 (56.9% ±2.4%) students who selected only man, 607 (38.5% ±2.4%) students who selected only woman, and 59 (3.7% ±0.9%) students who selected another option (either nonbinary, not listed with an opportunity to write their preferred gender in a text box, or some combination of the above responses). Exactly 13 students (0.8% ± 0.4%) did not respond to this question. We do not have enough responses in any category other than only man or only woman to analyze these responses. Thus, we treat gender as a binary and include only those students who answered along this binary in the gender analysis. We hope to collect more data in the future to be able to include other categories as well.

2. Ordinal logistic regression

As a final component of addressing our second research question, we examine the correlations between students' major and gender and their AO-level post-test scores, as well as determine the impact of an instructor's reported importance of that AO on these scores. To do this, we perform ordinal logistic regression [48]. We take SPRUCE AO scores as ordinal (the scores have a clear order attached to them, as a student who scores a one has a higher performance than a student who scores a zero), but the AO scores are not continuous (i.e., within an AO, only integer scores are possible before normalizing). The total possible scores for each AO are presented in Table III. For ease of analysis, we use these rounded, non-normalized integer AO-level scores.

The explanatory variables for the ordinal logistic regressions performed in this work are major (categorical), gender (categorical), and the importance of an AO to an instructor (ordinal, based on Likert-scale data). Additionally, the ordinal pretest scores were included as an independent variable. This analysis was conducted in both PYTHON and R to verify that the results are the same with two different statistics packages; the results were in agreement with each other in both programs.

The ordinal logistic regression model we fit to our data is as follows:

$$\log\left(\frac{\Pr(S_{\text{post}} \le j)}{\Pr(S_{\text{post}} > j)}\right) = \alpha_j + \beta_1 S_{\text{pre}} + \beta_2(\text{Gender}) + \beta_3(\text{Major}) + \beta_4(\text{Importance}), \quad (2)$$

where $\Pr(S_{\text{post}} \le j)$ is the cumulative probability that the single AO post-test score is either j or lower (where j is an integer score on that AO), $\Pr(S_{\text{post}} > j)$ is the cumulative probability of the score being higher than j, α_j is the y intercept for score integer j, β_1 is the coefficient for the pretest score on that particular AO, β_2 is the coefficient for students' gender, β_3 is the coefficient for students' major, and β_4 is the coefficient for the importance variable (i.e., how important instructors ranked that particular AO on a Likert scale).

In order to examine the impact of interaction terms, we also model the following for each AO:

$$\log\left(\frac{\Pr(S_{\text{post}} \le j)}{\Pr(S_{\text{post}} > j)}\right) = \alpha_j + \beta_1 S_{\text{pre}} + \beta_2(\text{Gender}) + \beta_3(\text{Major}) + \beta_4(\text{Importance}) + \beta_5(\text{Major} \times \text{Gender}) + \beta_6(\text{Major} \times S_{\text{pre}}). \quad (3)$$

The variables in this equation are identical to those in Eq. (2), but we have added extra terms to account for interactions between students' major and gender and students' major and pretest scores. We include only these interaction terms in our model because they are the ones that have reasonable theoretical explanations for the correlation. For example, the importance of a specific AO to a course is conceptually distinct from a student's gender. This is true for all other possible interaction terms not included in the model.

In terms of gender and major interaction (which is similarly seen in the ANCOVA analysis), this interaction can be explained conceptually by noting that physics majors are more likely to be men, and physics majors are more likely to do well on SPRUCE. In terms of the major and pretest interaction, this can conceptually be explained by the fact that physics majors are more likely to do well on SPRUCE even before taking a course in which SPRUCE is administered, likely due to high school preparation or prior coursework in physics.

In our ordinal logistic regression analysis, we report odds ratios, which are calculated as e^{β} for each β coefficient. In this analysis, the order of the categorical groups must be chosen. Odds ratios present the likelihood of improving a level (in this case, going from one score on the post-test to a score an integer higher on the post-test on that particular AO) as a multiplicative factor based on changing from one group to an adjacent group (e.g., from men to women or engineering major to physics major). Thus, the odds ratios with confidence intervals that cross one are not statistically significant. Those that are greater than 1 show that there is an increased chance of a greater AO score by moving from one group of majors to an adjacent group in a particular direction (e.g., from engineering majors to physics majors), based on the ordering of the groups. Those with confidence intervals strictly less than 1 show a higher chance of

decreasing the post-test score on that AO while comparing those adjacent groups in the same direction.

One positive aspect of logistic regression, as compared with linear regression, is that logistic regression is a nonparametric technique, meaning that there are no assumptions necessary about the underlying distribution of the data. Not only does this mean that the data do not need to be normal, but it also means that we do not require homoscedasticity, or constant variance of the residuals in the data [49]. This is because logistic regression uses maximum likelihood estimation (MLE), an iterative procedure to find the solution, instead of ordinary least squares (OLS) regression. MLE maximizes the likelihood that individual students have scores given by the dependent variable (in this case, post-test scores) based on their scores on the predictor variables (in this case, pretest scores, major, gender, and the level of importance their instructor places on that AO). Logistic regression does, however, have several assumptions that must be met in order for it to be applied to our data. We discuss these assumptions as well as the adherence of our data to them in detail in Appendix B.

IV. RESULTS AND DISCUSSION

A. Overall student proficiency with measurement uncertainty

Here, we examine the first research question by using only post-test data. We determine areas of measurement uncertainty where students excel, as well as areas where additional support could help improve their proficiency. Thus, we report the mean (both the overall post-test score as well as post-test AO scores) and comment on the results.

The mean overall post-test score on SPRUCE, as calculated based on methods described in Sec. II B 2 above, is 52.3 ± 0.5 with a standard deviation of 18.9. A histogram showing the distribution of overall post-test scores is shown in Fig. 2, where the scores have been normalized to 100 in this case only for ease of understanding. This distribution is indicative that our data appear visually normal. To quantify this, we perform an Anderson-Darling test for normality and find the post-test scores are normal to a significance level of 1.0% with a skewness of -0.49 ± 0.12 and a kurtosis of -0.55 ± 0.12, indicating normal data. For both skewness and kurtosis, values between -1 and 1 generally indicate normality [50].[2] Similar figures showing the distributions for all ten AO scores are shown in Appendix C. We also present the average score on each AO in Table VI; note that these are normalized to one, with uncertainty presented as the standard error.

[2]Note that this reference uses a measure of kurtosis that adds three to the method we use and therefore states that normality is present for kurtosis values of 2 to 4; this corresponds to our values when we subtract 3.

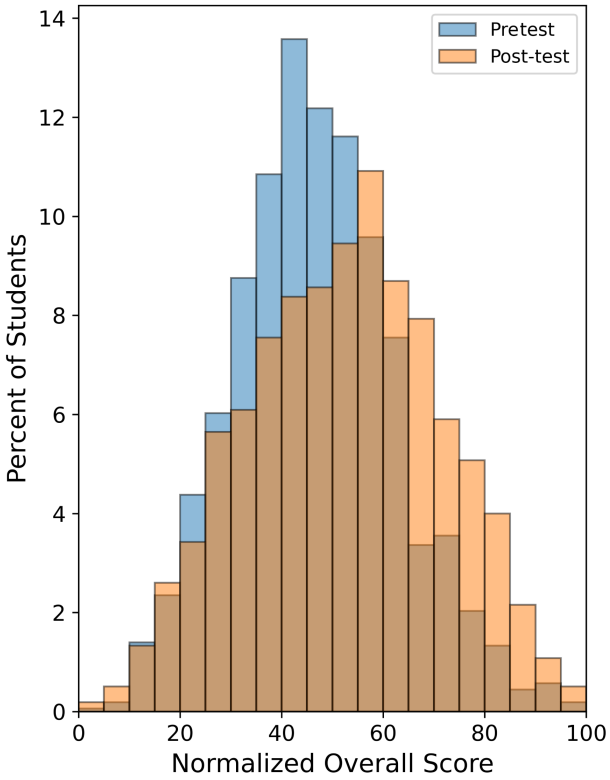


FIG. 2. Pretest (blue) and post-test (orange) overall scores on SPRUCE normalized to 100. In aggregate, students improve from pretest to post-test, as can be seen by the clear shift in the histogram. The distributions themselves are considered normal, with skewness and kurtosis levels for both pre- and post-test distributions well within the limits of normality and Anderson-Darling tests showing that both distributions are normal to a significance level of 1.0%. The ranges of scores show that SPRUCE does not suffer from ceiling or floor effects in the overall score.

These data help demonstrate student proficiency with measurement uncertainty. For example, students tend to do well at *D3: Calculate and report the mean of a distribution for the best estimate of the measurement*, which is the only AO with an average post-test score greater than 70%. On the other hand, students are less successful on AOs *D5: Determine if two measurements (with uncertainty) agree with each other*, *H1: Propagate uncertainties using formulas*, and *H2: Report results with uncertainties and correct significant digits*, which are perhaps areas instructors might focus on for improvement. All three of these AOs have postinstruction scores of less than 40%.

B. Impact of instruction

In this section, we explore the answer to the second research question, relating to the impact of instruction on student proficiency with measurement uncertainty. We first examine the significance of the shifts from pretest to post-test both at the overall score level and at the AO level.

TABLE VI. AO average scores, pretest and post-test [$N = 1576$], normalized to 100. Error presented is standard error, shown as uncertainty in the last digit (e.g., $51.4(7) = 51.4 \pm 0.7$). AO labels follow from Table I. D3, which relates to students reporting the mean, has the highest score for both the pretest and post-test, and likely exhibits ceiling effects. H1 (error propagation) and D5 (measurement comparison) have the lowest scores and thus represent student proficiencies that have substantial room for improvement.

	Average score, pretest	Average score, post-test
S1	51.4(7)	54.9(7)
S2	59.0(6)	62.5(5)
S3	37.1(6)	41.6(6)
H1	28.9(6)	38.9(6)
H2	32.6(7)	39.9(7)
D1	40.2(9)	45.8(8)
D2	60.3(8)	63.5(9)
D3	84.5(7)	81.9(7)
D4	43.8(6)	50.7(6)
D5	29.0(9)	35.7(8)
Overall	46.7(4)	52.3(5)

We then examine the correlations gender and major have with post-test score using ANCOVA (overall test) and ordinal logistic regression (AO level). Additionally, we examine the impact of the importance an instructor places on a specific AO on student performance on that AO using ordinal logistic regression.

As shown in the pretest and post-test distributions in Fig. 2, there is a clear shift of overall SPRUCE scores from pre- to postinstruction. We can quantify the significance of this shift using the Wilcoxon signed-rank test, as described in Sec. III B, and find that the pre-post shift is significant at $p \ll 0.0001$ with an effect size of $d = 0.33 \pm 0.04$.

Pre- and post-test scores along with the effect sizes (Cohen's d) of the shifts for each AO are shown in Fig. 3. Again, the Wilcoxon signed-rank test shows that all of the pre-post shifts at the AO level are significant with $p \ll 0.0001$ aside from AOs D2 (which is significant at $p = 0.0009$) and D3 (which is significant at $p = 0.004$).

We note that AO D3 [calculate mean] likely has ceiling effects, which results in a lower effect size due to students excelling at this AO both pre- and post-instruction. Further, from this plot's Cohen's d values, we can also see that while AOs D5 [measurement comparison] and H1 [propagate uncertainty] had similar post-test outcomes (with students struggling the most with these two AOs), instruction is having a significantly large positive impact on students surrounding AO H1 [propagate uncertainty] whereas their impact, though positive, is much smaller for AO D5 [measurement comparison].

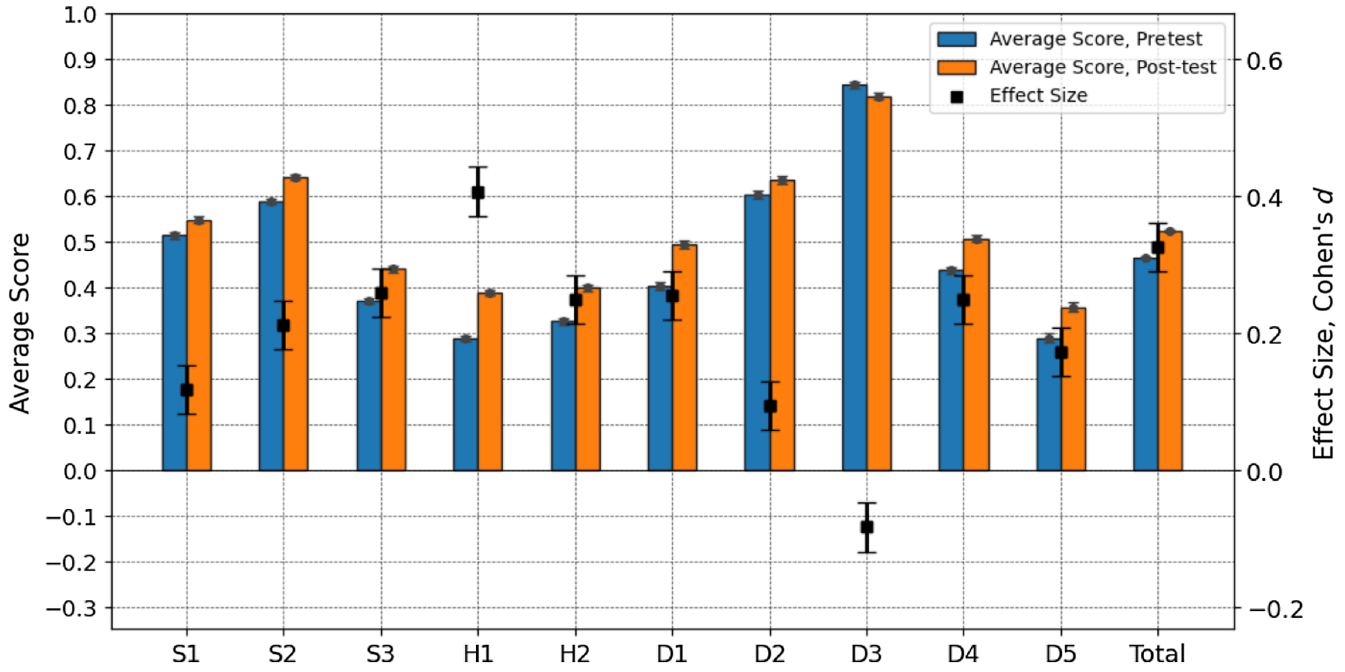


FIG. 3. AO scores and overall score for both pre- and post-test [$N = 1576$]. Error bars on the scores represent the standard error. AO labels follow from Table I. Effect size (calculated via Cohen's d) is shown with black squares using the scale on the right, with error bars representing the standard error. The pre-post shifts were significant for all AO scores and for overall scores (as determined via the Wilcoxon signed-rank test), with varying effect sizes for these shifts. All shifts were positive aside from AO D3, which exhibits ceiling effects.

1. Impact of instruction on overall post-test score: Correlation with major and gender

We perform an ANCOVA to determine the impact of gender and major on overall SPRUCE post-test score.

Only students who responded to both demographic questions about major and gender and selected only man or only woman for gender were included in this analysis, which resulted in $N = 1503$ responses from the original $N = 1576$. A breakdown of number of students by gender and major is shown in Table VII.

Based on previous work on analysis of laboratory assessments, We split student majors into the following categories [22]:

1. Physics, engineering physics, astrophysics.
2. Other engineering.
3. Other science and math (including astronomy and computer science).
4. Nonscience majors.

In order to assign students to one of the groups above, we first examined whether they selected a major in the first group (physics, engineering physics, and astrophysics). If they did, they were placed into the first group, even if they have other majors as well. If not, we determined whether they should be in the second group, then the third, then the fourth.

The ordering of the groups must be chosen for this analysis (and in the ordinal logistic regression analysis discussed later). In both analyses, we order the majors as follows: nonscience, other science or math, other engineering, and physics/engineering physics/astrophysics as is done in other assessment analyses (e.g., the PLIC [22]).

We also split students into man and woman categories. We have ordered these categorical variables such that an odds ratio greater than 1 indicates that men outperform women.

Initially, we used the model $\text{Post} \sim \text{Pre} + \text{Gender} + \text{Major} + \text{Gender} \times \text{Major}$ (which indicated a post-test score dependent variable, with independent variables of pretest score, gender, and major and interaction term gender × major), with an interaction term between major and gender included to test for its significance. We find this interaction term to be borderline in its significance (F test, $p = 0.046$). Because 0.05 is an arbitrary cutoff and this p value is on the edge, we have chosen to treat it as not significant. If it were significant, we would have to split the data and do six separate ANCOVA analyses for each category (for example, we would need to examine men only in the model $\text{Post} \sim \text{Pre} + \text{Major}$, and similarly for the other major and gender categories). This would obfuscate the conclusions one can draw from the data. Therefore, we choose to treat the borderline interaction term as not significant and use only the model presented in Eq. (1).

Results of the ANCOVA analysis are shown in Table VIII. Partial η^2 is an indicator of the effect size of each of these predictors (pretest, major, and gender) on the post-test score by indicating the amount of variance each explains in the post-test score. A partial η^2 of at least 0.01 indicates a small effect, and anything above 0.06 is at least a medium-strength effect [47]; reported partial η^2 values should be considered a lower bound due to shared variance between the covariate (pretest score) and independent variables (major and gender), as discussed further in Appendix A. All p values in this table are calculated via the F test, with gender having 1 degree of freedom and major having 3 degrees of freedom. Instruction is not accounted for in this model and is likely the cause of much of the residual variance.

While pretest score is a significant predictor of post-test score (both by p value and by partial η^2), when we control for this, we find that both major and gender are predictors of post-test score. Major is more significant and accounts for more variance than gender does. However, much of the

TABLE VII. Number of students by gender and major [$N = 1503$]. The genders indicated are for students who selected only one single gender, whereas the majors indicated might be one of several majors selected by students, but students were placed into only one group based on their major such that no students are double counted in this table. The numbers add up to 1503 due to 73 of the 1576 matched responses either did not include their demographic information or did not select majors compatible with this analysis. Physics includes students majoring in physics, engineering physics, and astrophysics. Other engineering includes all other types of engineering students. Other science includes students majoring in a science not listed above, including chemistry, astronomy, and computer science. Finally, nonscience includes all possible majors outside of science.

	Men	Women	Total
Physics	227	113	340
Other engineering	401	172	573
Other science/math	241	292	533
Nonscience	28	29	57
Total	897	606	1,503

TABLE VIII. ANCOVA results, including p values and partial η^2, a measure of the amount of variance explained by each of the predictors. We find that pretest is a significant predictor of post-test score and explains much of the variance in the post-test scores. Similarly, major and gender also are significant predictors of post-test score but account for less of the variance. All variance not explained by these three predictors must be explained by some other variables not included in the model, such as the impact of instruction.

Predictor	p	Partial η^2
Pretest	<0.001	0.360
Gender	0.022	0.003
Major	<0.001	0.012

Num. Couplets	AO	Pretest	Gender	Major	Importance
2	D1†	2.58 [2.23, 3.00]*	1.41 [1.15, 1.72]*	1.27 [1.13, 1.44]*	1.21 [0.87, 1.68]
	D2	3.34 [2.83, 3.93]*	1.22 [0.99, 1.50]	1.29 [1.14, 1.46]*	0.79 [0.62, 1.00]
	D3	2.95 [2.45, 3.54]*	1.11 [0.87, 1.41]	1.14 [0.99, 1.31]	1.17 [0.87, 1.57]
	D5	3.66 [3.17, 4.22]*	0.94 [0.76, 1.17]	1.08 [0.95, 1.22]	1.48 [0.97, 2.25]
3	S1	2.88 [2.54, 3.26]*	1.14 [0.93, 1.39]	1.13 [1.00, 1.27]*	0.85 [0.77, 0.96]*
	H2†	2.13 [1.89, 2.40]*	0.92 [0.76, 1.12]	1.34 [1.19, 1.51]*	1.05 [0.85, 1.29]
4	S3†	1.91 [1.73, 2.11]*	1.37 [1.13, 1.66]*	1.19 [1.06, 1.33]*	0.94 [0.77, 1.14]
	H1	1.59 [1.44, 1.76]*	1.37 [1.13, 1.66]*	1.48 [1.31, 1.66]*	1.07 [0.96, 1.19]
	D4	1.81 [1.65, 1.99]*	1.02 [0.85, 1.23]	1.34 [1.20, 1.50]*	– – – – –
5	S2	1.83 [1.68, 1.99]*	1.20 [0.99, 1.39]	1.24 [1.11, 1.39]*	0.93 [0.77, 1.12]

FIG. 4. Odds ratios for major, gender, and importance. Values shown are 95% confidence intervals. We denote significance with an asterisk (∗) (i.e., the confidence interval does not cross 1) and † indicates a significant interaction term present in the model. This table is arranged in order of number of couplets (logistic levels) in each AO (two, three, four, five). Yellow cells indicate a positively correlated predictor for that AO and blue cells indicate a negatively correlated predictor for that AO; odds ratios were calculated using Eq. (2). Pretest is a significant predictor in all models, whereas gender, major, and importance only play a role in certain AOs. In particular, importance is only significant for AO (S1) and it is an inverse predictor.

variance in post-test score is not accounted for by any of the three predictors used in the model, leading us to presume that instruction (which is not included in the model) plays a significant role in post-test scores. Instruction helps students improve their SPRUCE overall scores, indicating some success in increasing students' proficiency in measurement uncertainty.

2. Impact of instruction on AO-level post-test scores: Correlation with major, gender, and AO importance to instructor

We performed an ordinal logistic regression analysis with pretest AO score, major, gender, and importance of the AO to the instructor as explanatory variables for the post-test AO score. Again, only students who responded to both the demographic questions about major and gender were included in this analysis, with the additional requirement that the instructor for the course responded to the course instructor survey, meaning that $N = 1486$ for this analysis (73 of the 1576 matched responses did not include the gender and major demographic information or did not respond with only man or only woman for their gender and were thus excluded from this analysis, and a further 17 students were enrolled in classes in which the instructor did not respond to the question in the instructor survey regarding importance for each AO).

The model we use is described by Eq. (2).[3]

We report the odds ratios for each of the AOs in Fig. 4. It is important to note that odds ratios cannot be compared between AOs with different numbers of ordinal levels associated with them. We can, however, compare the odds ratios for AOs with the same number of ordinal levels (e.g., S3 and H1 can have their odds ratios compared, but H1 and H2 cannot). An area of nascent research in educational statistics is determining how to compare statistical analyses from groups with different number of levels.

As described in the Methods, we also examine interaction terms using Eq. (3). Three AOs have significant interaction terms ($p < 0.01$). AOs S3 and D1 show significant interaction between pretest and major, and AO H2 shows a significant interaction between gender and major. The odds ratios reported are for the model in Eq. (2), that is, the model without the interaction term, because odds ratios for the stand-alone terms in models with interaction terms do not have meaning due to the collinearity between the stand-alone and interaction terms. However, significant interaction terms indicate that when interpreting odds ratios, one should take caution to remember that the full effect is not explained by these two variables alone, but rather that one impacts the other.

We find that pretest score is a significant predictor of post-test scores for all ten AOs, which is expected—students who start with better scores also end with better scores. Further, we find that gender is a significant predictor for several AOs: S3 [accuracy], H1 [propagate uncertainty], and D1 [several measurements]; in all cases, men perform better than women. Major is a predictor for nearly every AO, aside from D3 and D5. Again, in these cases, physics majors have a higher likelihood of better performance than engineers, engineers have a higher likelihood of better performance than other science and math majors, and so on.

Finally, and most notably, the instructor-rated importance of an AO is only a significant predictor of post-test score for AO S1 [estimate size of uncertainty]. However, it is actually an *inverse* predictor in this case; that is, instructors who rated this AO as important were more likely to have students perform *worse* on this AO on the

[3]The model for AO D4 looks much the same as the model in the referenced equation, but with β_1 set to zero and no data about importance included in the model (due to our lack of data collected about this).

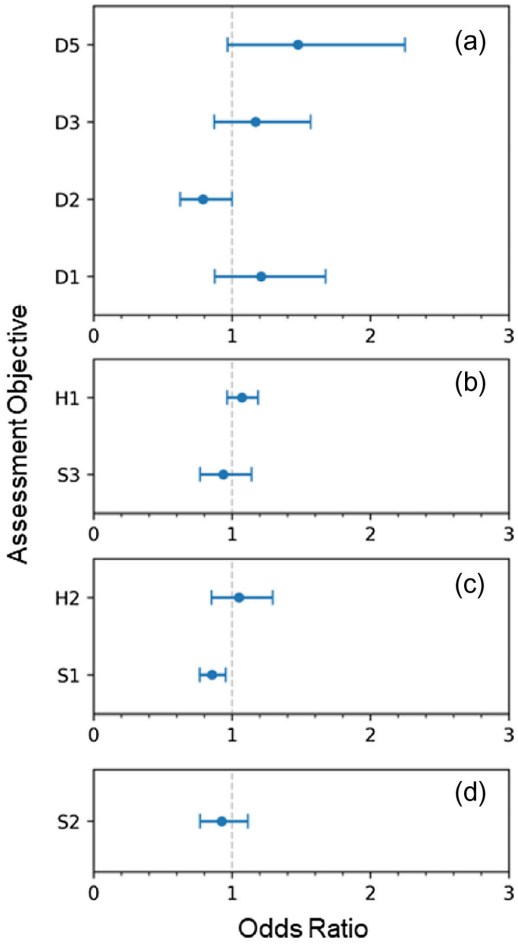


FIG. 5. Odds ratios for importance, separated by the number of couplets per AO—two (a), three (b), four (c), and five (d) couplets. These odds ratios show the impact of instructor-reported importance of an AO on students' final post-test score. Ideally, these odds ratios would be greater than 1 with 95% confidence. However, none of them are statistically significant, indicating that the level of importance placed on an AO by instructors is not correlated with student performance on that AO, with a single exception: S1 scores are slightly negatively correlated with importance, due to its odds ratio lying below 1 with 95% confidence.

post-test compared to students in other courses. This, combined with the lack of significance on the other AOs, indicates that students are not necessarily achieving the instructor's stated goals for the course with respect to measurement uncertainty. One potential cause of this is that instructors who do not report a particular AO as important still teach those concepts as well as those who do report it as important. Overall, instructors are not having a significant, positive impact on the areas of measurement uncertainty that they deem important as compared to instructors who do not report those areas as important. To further highlight this, the odds ratios for importance are presented in Fig. 5.

We again note that the odds ratios for AOs with different numbers of couplets should not be directly compared.

C. AOs of interest

The results for several of the AOs are particularly interesting and require a deeper investigation. Some of these are examined in greater detail in the following sections. The end of each AO analysis subsection provides recommendations for instructors pertaining to that AO.

1. AO S1: Estimate size of random or statistical uncertainty by considering instrument precision

This AO is particularly interesting when considering the results of the ordinal logistic regression since it is the only one with a significant odds ratio for importance. However, the odds ratios for this AO indicate that the more importance an instructor places on this AO, the lower their students' scores on this AO. This is opposed to an instructor's aims when teaching a lab course.

This AO had a mean post-test score of 0.549 ± 0.007 with only a slight improvement from pretest (effect size $d = 0.13 \pm 0.04$, pretest mean $= 0.514 \pm 0.007$). While this is not the AO students struggle with the most, it is frequently reported to be important by instructors, and students still have many difficulties when determining instrument precision and incorporating this into their uncertainty.

This AO is probed 3 times by SPRUCE. Two of these items are coupled numerical open response and one is coupled multiple choice. In all cases, students are asked to provide a measurement and an uncertainty associated with that measurement based on a specific instrument shown.

One of these items presents students with two graduated cylinders filled with water showing a "before" and an "after" measurement in order to determine the volume of an irregular object. Students report the volume shown in both of these circumstances, as well as the uncertainty in the measurement for both. Importantly, it is the same graduated cylinder in both measurements, with measurement markings every 100 ml.

Students frequently believe that a certain type of instrument always has a specific uncertainty, regardless of the markings on that instrument. For example, one student during an interview entered 0.05 ml as the uncertainty for both measurements and explained that this was because

> It's 0.05 for, what was that, a volumetric flask or a graduated cylinder.

This indicates that they believe all graduated cylinders have the same uncertainty, regardless of measurement markings on the instrument. This is directly opposed to AO S1, which requires students to understand that the precision of their

measurements is directly related to the precision of the specific instrument they are using. In this case, 0.05 ml is much too precise for a graduated cylinder with markings every 100 ml.

Further, some students believe that the instrument precision changes depending on the value being shown. For this same question, another student entered an uncertainty of 5 ml for the "before" measurement and 6 for the "after" measurement. They explained this choice as

> The way that I picked this first uncertainty for the before was I definitely know that it's more than 1500 milliliters, but it looks like it's less than halfway. But because we can't really accurately judge where halfway is, I just kind of said that it was 5 milliliters to get that span. And then this second one is a little closer to halfway so I just extended it one more milliliter.

While both 5 and 6 ml are still too precise for the instrument, this student's reasoning is also interesting because, despite using the exact same graduated cylinder, they believed the precision of the device had changed simply because the amount of liquid in it had changed. This is also not aligned with mastery of AO S1.

Another similar item that probes this AO shows students a digital scale with a reading of 74.2 g and asks students to enter the value and uncertainty from this scale.

Some students employed incorrect reasoning on this item. Similarly to the first item discussed, some students believe that all digital scales have the same uncertainty. For example, one student entered 0.01 g said,

> I just remember that 0.01 is a general value, like a generalized uncertainty… A generalized uncertainty is 0.01,

showing that they believe that digital scales in general have this uncertainty associated with them regardless of the precision of the output. This value is too precise for this particular scale, and the student's reasoning does not show proficiency in AO S1.

Finally, some students believe digital instruments have no uncertainty whatsoever. For example, one student entered 0 g for the uncertainty and said,

> Uncertainty comes from either a scale that's giving you a bunch of different readings and you have to take measurements over time or use a bunch of different scales and see what you'd get. For uncertainty for this, I don't understand that there would be any,

which shows that this student understands uncertainty from multiple devices or from one device with a flickering display, but does not consider a single instrument's precision when determining the uncertainty of a measurement (especially when it is a digital device). Interestingly, this student did provide uncertainties for the graduated cylinder question, showing that their belief in the lack of uncertainty for a single static instrument is related specifically to digital instruments.

Thus, while estimating the size of statistical uncertainty based on instrument precision is important to instructors, current instruction seems to be somewhat ineffective in raising student scores extensively (especially for instructors who rate this AO as important). Major is a significant predictor for student performance on this AO, with physics majors outperforming engineering majors, etc. However, this effect is relatively small—the odds ratio is 1.13 [1.00, 1.27], indicating that with uncertainty, it is possible that all majors perform identically (odds ratio of 1.00). We hope that illustrating some common student misconceptions surrounding this AO will help improve instruction in this area. Explicit instruction surrounding how the precision of the measurement instrument, including digital instruments, can aid in determining the uncertainty of the measurement and can potentially help students overcome these challenges.

2. AO D1: Articulate why it is important to take several measurements during experimentation

This AO is especially interesting for comparing SPRUCE interview data with the previously discussed PMQ paradigms of pointlike and setlike reasoning because the PMQ deals extensively with this topic. Pointlike reasoning is employed when students believe there is one true value for an experimental measurement, and setting up an ideal experiment will yield that true value. Setlike reasoning is aligned with expert views, in which students believe that any experimental setup will yield a distribution of results. We find that this AO had a post-test mean of 0.458 ± 0.008 with a medium-size shift between pre- and post-test scores (effect size $d = 0.26 \pm 0.04$, pretest mean $= 0.402 \pm 0.009$). Logistic regression shows that gender and major are both significant predictors for post-test scores on this AO, with a slightly larger effect from gender. However, this AO showed significant interaction terms between gender and major, so caution must be taken when analyzing these results—each of these variables impacts the other, leading to the final post-test score.

Two SPRUCE items probe this AO. Both are coupled multiple response in which students are asked a multiple-choice question and then are asked a follow-up multiple-response question to ascertain their reasoning behind their answer to the first question.

During the interview phase, we found a common response to questions of collecting more data was simply a blanket statement about having more data being the best

practice without providing reasoning as to why it is better to have more data, similar to results from prior studies [25,36]. For example, one student said,

> I've always understood that it's best practice to take measurements multiple times in experiments.

Another common line of reasoning for students is the desire to take several measurements because of human error. For example, one student stated,

> Usually when I do experiments I like to do three trials because sometimes… there might be a human factor involved in it. It's just always good to do three trials so you can look at your data and compare.

Again, this student has a similar understanding that taking more data is better and provides minimal evidence as to why, while seeming to employ pointlike reasoning of taking multiple measurements to ensure the reproduction of the "true value." This student specifically quotes three measurements, similar to findings in prior studies [23]. Further, this quote exemplifies the issues regarding using "error" as opposed to "uncertainty" when describing random and systematic effects, as previously explored [5,7]. Both of these students left off key parts of a fully correct response to items probing this AO. For example, in this second case, the student did not select answer options about calculating the uncertainty and reducing the impact of outliers as reasons for collecting more data.

Another student discussed the true value of a measurement:

> Measuring something one time is not super—it's something that I learned not just in physics classes but in many different science classes, if you can measure something more than once you should, multiple trials are great because you can find a mean value which will be as close to the true value as possible… if the mean is close to the mode, like the most common value… I can use that to my advantage statistically.

This student has a pointlike view of measurement, in which there is one true value for the measurement. They hope that their mean might be close to this true value and intend to test this hypothesis by determining whether the mode is close to the mean. This reasoning shows a lack of understanding about why one should take many measurements; comparing the mean and mode is statistically irrelevant. This student also leans toward labeling the mode as the true value as they go on to state that they might use the mode in further calculations if the mean varies significantly from the mode.

To reduce students' reasoning based on a single measurement and better emphasize that uncertainty is not the result of mistakes, we suggest that instructors focus on explaining *why* collecting more data is better (rather than simply stating that more data is better or instituting minimum requirements for data collection without proper explanation) in order to help students become more proficient in this area.

3. AO D5: Determine if two measurements (with uncertainty) agree with each other

This AO was previously examined in prior work [51], which is expanded upon here with a more complete dataset and using only matched pre-post responses. It is probed by two isomorphic items that are shown in Fig. 6. The first asks students whether their measurement agrees with other groups' measurements using a numeric representation (NRI) while the second asks students the same question using a pictorial representation (PRI).

While students generally perform poorly on both of these items, they tend to perform better on the pictorial version despite the items being identical in content. The post-test mean for AO D5 is 0.357 ± 0.008, the lowest score of all AOs. Of the 1576 post-test responses to these items, 354 students ($23 \pm 2\%$) answered both items correctly, while 103 students ($7 \pm 1\%$) correctly answered only the NRI, 315 students ($20 \pm 2\%$) correctly answered only the PRI, and 804 students ($51 \pm 3\%$) answered both items incorrectly. Further, this AO shows some improvement from pretest to post-test, but this improvement was small (effect size $d = 0.17 \pm 0.04$, pretest mean $= 0.290 \pm 0.009$).

Figure 7 shows a heat map of the 905 most common student responses on the post-test for both the NRI and the PRI. One might expect responses to occur only along the diagonal, indicating students who selected the same answer combination for both the NRI and the PRI, but this is not the case. Instead, students frequently select different answers to these items, indicating a need for further instruction in measurement comparison. Of the 1576 total responses, only 433, or about 27% of students, selected the same answer combination for these items (whether a correct or incorrect combination).

During interviews, we frequently found that students who correctly answered the numeric version of the item discussed a mental pictorial version despite not yet encountering the PRI on SPRUCE. For example, one student said,

> I just looked at the values and saw it—like I kind of picture if they have that little bar with their error bars to see if they overlap.

NRI Using your values for the mass and period (and uncertainties), you use the formula:

$$k = \frac{4\pi^2 m}{T^2}$$

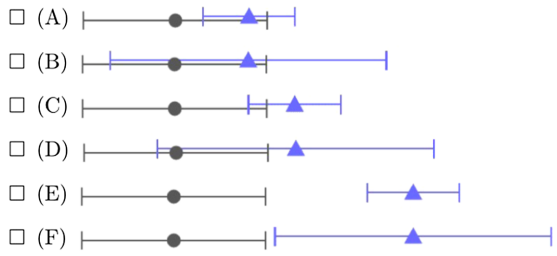
to calculate your spring constant and uncertainty, and you get the following value:

$$k = 3.62\frac{\mathrm{N}}{\mathrm{m}} \pm 0.11\frac{\mathrm{N}}{\mathrm{m}}$$

Several other lab groups took different approaches to calculating the spring constant. Their values (with estimated uncertainty) are shown below. Select **all** of these values you believe **agree** with your measured value.

- ☐ (A) $3.71\frac{\mathrm{N}}{\mathrm{m}} \pm 0.06\frac{\mathrm{N}}{\mathrm{m}}$
- ☐ (B) $3.71\frac{\mathrm{N}}{\mathrm{m}} \pm 0.17\frac{\mathrm{N}}{\mathrm{m}}$
- ☐ (C) $3.76\frac{\mathrm{N}}{\mathrm{m}} \pm 0.06\frac{\mathrm{N}}{\mathrm{m}}$
- ☐ (D) $3.76\frac{\mathrm{N}}{\mathrm{m}} \pm 0.17\frac{\mathrm{N}}{\mathrm{m}}$
- ☐ (E) $3.91\frac{\mathrm{N}}{\mathrm{m}} \pm 0.06\frac{\mathrm{N}}{\mathrm{m}}$
- ☐ (F) $3.91\frac{\mathrm{N}}{\mathrm{m}} \pm 0.17\frac{\mathrm{N}}{\mathrm{m}}$
- ○ (G) None of these agree with my data

PRI You decide to compare your group's estimate of m_{breaking} with six other groups by sketching your results (gray circles) next to their results (blue triangles) on six different graphs, shown below. The error bars in the graphs represent the uncertainty in the measurements. Select **all** graphs that depict **agreement** between your data and data from other groups in your class.



- ☐ (A)
- ☐ (B)
- ☐ (C)
- ☐ (D)
- ☐ (E)
- ☐ (F)
- ○ (G) None of these agree with my data

FIG. 6. Two isomorphic items on SPRUCE. These items probe student understanding of measurement comparisons with uncertainty by presenting the same data in two different representations—a numerically represented item (NRI) and a pictorially represented item (PRI). The students first encounter the NRI and then, after answering several unrelated items, they encounter the PRI in a different experimental context. Responses of "ABCD" and "ABCDF" receive full credit, while no other combinations receive any credit. Note that the answer options on the PRI are in a different order when presented to students (DAEBFCG) than shown here; we present them in the same order as the answer options for the NRI in this paper for ease of understanding.

This ability in being able to switch between different representations aided this student in correctly answering the numerically presented item; they also were able to correctly answer the pictorially presented item.

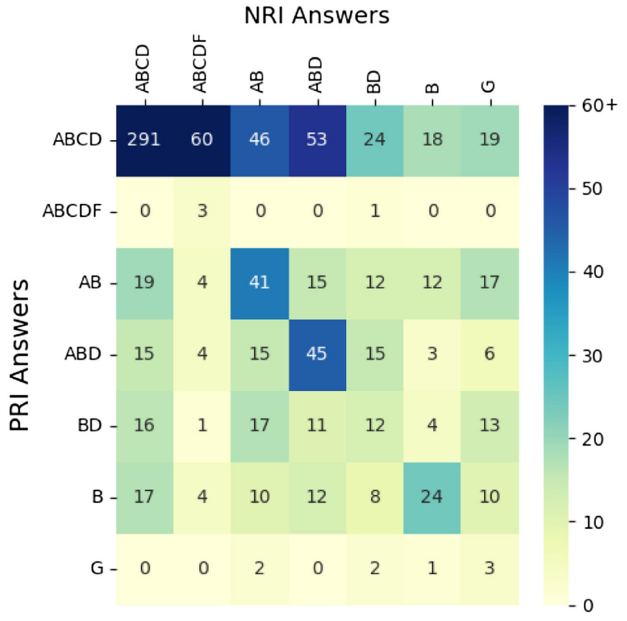


FIG. 7. Heat map showing the most common 905 of the total 1576 post-test responses to the NRI and PRI. Both ABCD and ABCDF are accepted as correct responses. Responses along the diagonal indicate students who selected the same answer combination for both the NRI and the PRI, while off-diagonal elements indicate students who selected different answer combinations for these two items.

Students often provided incorrect reasoning for one item and not the other. For example, one student selected all answer options (aside from "None of these agree with my data") on the NRI, and said,

> Honestly I would just say all of them… that's still at the end of the day what they got… We don't have enough data to say like 'no yours are all wrong because they don't exactly match ours' because there are a lot of factors that could have altered their numbers and their uncertainty. I know that's a very idealized way of thinking about science.

This student is unwilling to say that any of the measurements disagree because all students are performing the same experiment. However, this student provided expert-like reasoning regarding the overlap of the full range of each measurement when correctly answering the PRI, showing a clear difference in thinking about measurement comparison between the two representations.

One common line of incorrect reasoning was students implicitly valuing their own measurements more than others'. For example, one student who selected only "AB" on the numeric item said,

> For the other four groups… their values did not put them in the same range as my value with its

> uncertainty so I don't believe they agree with my value.

When comparing numeric measurements with uncertainty, they placed more weight on their own measurement; in order for measurements to agree, the other group's measurement had to be encompassed by their own error bars. When solving this problem, they only added and subtracted their uncertainty from their own value and then selected the two answers whose means fell within that range; they ignored the uncertainties in the measurements in the answer options. However, when answering the PRI, this same student selected a correct response of "ABCD" and provided expertlike reasoning. Thus, their reasoning changed with representation.

Prior research into students' handling of different representations of the same problem [52–56] shows that representation is very important in student proficiency at problem solving in all areas of physics, not just in measurement uncertainty. Because many scientific papers present results numerically with uncertainties, being able to compare numeric results between papers is a vital skill for students to learn in experimental physics courses, and we encourage instructors to help students learn to switch between different representations to bolster this skill. Further, instruction emphasizing that students should not prioritize their own measurements over others could again help students become more proficient in this area.

V. SUMMARY AND FUTURE RESEARCH

We have presented an overview of student proficiency in measurement uncertainty, including the impact of instruction and the correlation of students' gender and major with their performance on SPRUCE. We find that instruction does tend to lead to better scores on SPRUCE, both overall and at the individual AO level, though these effects vary between AOs. At the overall score level, we find students' scores improved from 46.7(4) on the pretest to 52.3(5) on the post-test, showing a statistically significant improvement and evidence of learning during one term of a lab course. Of the ten AOs, nine show improvement as well, and we find these increases to be statistically significant. Students improve most on AO H1 [propagate uncertainty], showing an increase from pretest to post-test of about 10%. Students improve the least on AO D2 [distribution of results], with an increase of about 3.2%. Students do worse on AO D3 [calculate mean] by about 2.6% on the post-test than on the pretest due to ceiling effects.

Further, we find that instructors rating specific areas as important do not correlate with student post-test scores (aside from one case in which this correlation is inverse). Overall, students excel at reporting the mean as a final answer and struggle with comparing measurements with uncertainties, propagating uncertainties using formulas, and correct use of significant figures. We also find that gender is a statistically significant but weak predictor of student performance on SPRUCE. Additionally, it is only correlated with performance on three of the ten AOs on SPRUCE. Altogether, these results about gender show a promising step toward improving issues associated with gender bias in physics courses and assessments.

Further, from student interview data and the analysis of outcomes on SPRUCE, we present several suggestions for instructors. First, because students struggle with understanding why collecting more than one data point is important, we suggest that instructors emphasize this rather than providing minimum requirements without justification. Next, instructors should note that teaching students both numeric and pictorial representation methods of comparing measurements with uncertainty, as well as teaching students how to switch between these representations provides students the best tools to properly analyze data. This has been shown in prior research and is apparent from our analysis of identical questions with different representations in SPRUCE. Additionally, instruction on comparing measurements with uncertainty could help bolster students' skills in this area, because even after a semester of instruction, students struggle with this concept. Finally, because students sometimes struggle with identifying the precision of a measurement in relation to the instrument used to make that measurement, instructors should be deliberate in their treatment of this topic. This includes using various types of the same instrument with different measurement markings (e.g., rulers with different scales) to show that the specific instrument precision is important rather than treating each type of instrument as being the same. Instruction about digital instrument uncertainty is important, as often students have more confidence in digital scales than the measurement uncertainty would suggest.

In the future, as more SPRUCE data are collected, we plan to perform further analyses about student proficiency in measurement uncertainty. With more data, we can perform more advanced statistical analyses such as a cluster analysis to identify groups of similarly thinking students within the data [57]. Further, we hope to be able to perform ordinal logistic regression and ANCOVA to examine the correlation between race or ethnicity and student performance (an analysis not presented in this paper due to not having enough data from non-white students), as well as including gender minorities in future iterations of this work. We also aim to update the analysis of gender and major correlations with SPRUCE scores presented within this paper.

Additionally, future papers will investigate specific AOs further. For example, we are investigating student ideas surrounding accuracy and precision, as related to AOs S2

and S3 specifically. In addition to items related to accuracy and precision, students are presented with an initial question with four stereotypical bulls-eye targets and are asked to select which image depicts high precision and low accuracy. From this, we can correlate student performance on other items about accuracy and precision to see whether students understand the difference between these concepts at least in the bulls-eye representation.

In addition to helping to collect more data for these studies, instructors and researchers who are interested in using SPRUCE in their teaching and/or research can visit the SPRUCE website at [58] for more information about how to use it in their own classes and studies.

In conclusion, we have presented initial data from SPRUCE along with plans for future data collection and research studies, as well as concrete suggestions for instructors based on statistical analysis of SPRUCE results and student reasoning elements gathered from interviews.

ACKNOWLEDGMENTS

This work is supported by NSF DUE 1914840, DUE 1913698, and PHY 2317149. We would also like to thank Robert Hobbs, Ben Pollard, and Rachel Henderson for their work contributing to the development of SPRUCE, as well as the instructors and students who contributed to the body of data upon which this assessment was built and refined.

APPENDIX A: ANCOVA ASSUMPTIONS AND ADHERENCE

Hahs-Vaughn and Lomax discuss several assumptions [47] for ANCOVA. We provide evidence of our data meeting the following assumptions:

1. Independence of observations.
2. Homogeneity of variance.
3. Normality of the residuals.
4. Linear relationship between dependent variable and covariate.
5. Independent variable(s) fixed by researcher.
6. Independence of covariate and the independent variable(s).
7. Measure of covariate without error.
8. Homogeneity of regression slopes.

First, ANCOVA requires independence of observations —that is, that observations are independent of one another both within and across samples. However, no dataset collected from students will ever be truly adherent to this requirement in the strictest sense [49]. We are, of course, introducing a sampling bias by testing only students in physics courses. Further, since our data tend to be dominated by large R1 institutions, students are more homogeneous than the general population would predict. However, this will introduce only slight effects in our results and is a general issue with any assessment analysis in physics education research. The effects of this are small. We can check this assumption more rigorously both by examining plots of the residuals by group, as well as by using the Durbin-Watson statistic to test for autocorrelation [59–61].

Plots of the residuals are shown in Fig. 8. Because these plots appear random with no correlations and data fairly evenly distributed above and below the zero line, we determine that we have independence of observation. Additionally, we check the Durbin-Watson statistic for autocorrelations, applied to the residuals of the ANCOVA fit. This statistic falls between 0 and 4, with 2 representing

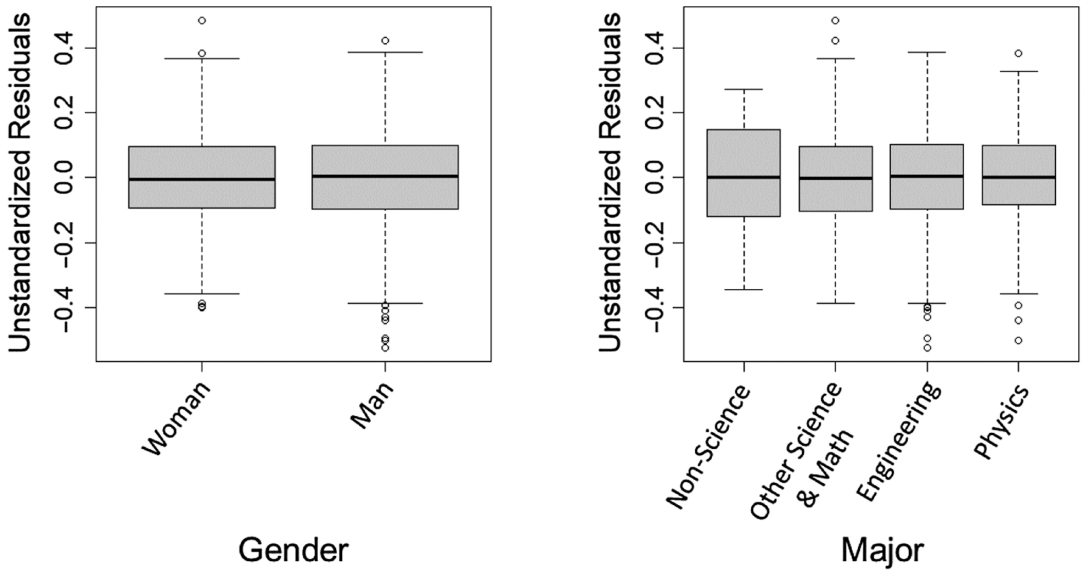


FIG. 8. Plots of the unstandardized residuals for the ANCOVA model. These show that our data conform to the independence of observation assumption as they fall randomly above and below the horizontal line at zero.

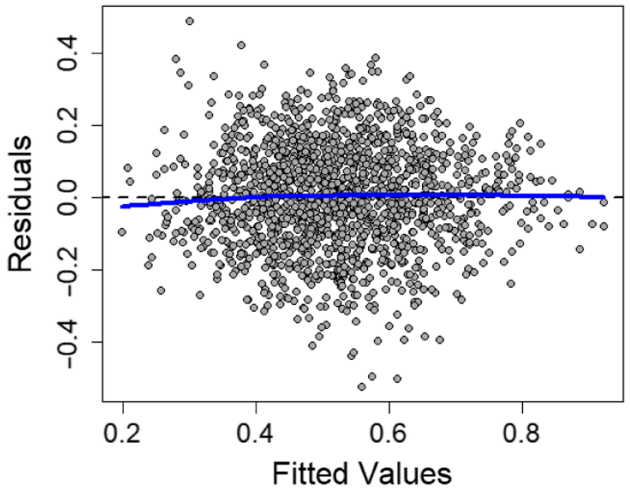


FIG. 9. Plot of the unstandardized residuals versus the fitted model. This shows that our data conform to the homogeneity of variance assumption. Shown as a solid blue line is the averaged data, which indicates no significant discernible patterns and is nearly perfectly aligned with the residuals = 0 line (black dashed line), indicating normality and homogeneity of variance assumptions have been met.

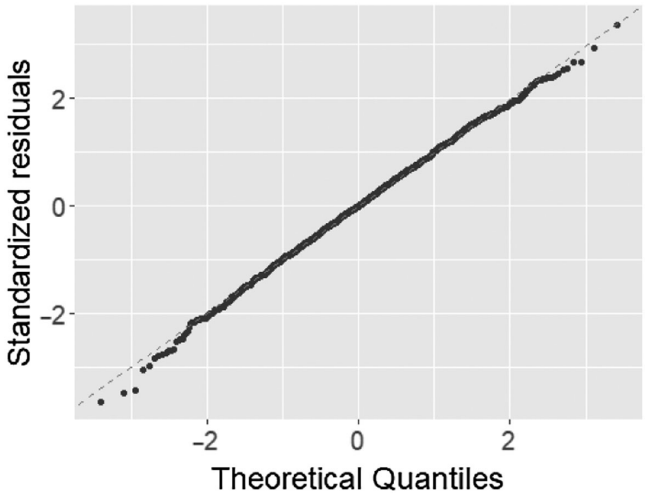


FIG. 10. Q-Q plot. These data indicate normally distributed residuals because the points adhere closely to the diagonal line; too many points deviating from this diagonal would indicate that the residuals are not normally distributed and may point to excess outliers.

uncorrelated data, 0 representing strongly positively correlated data, and 4 representing strongly negatively correlated data [59–61]. Our value of 2.04 is close to 2, thus, we can state that our data follow the independence of observation assumption.

The second assumption of ANCOVA is homogeneity of variance. The variances of each population must be the same (in our case, the variance of post-test scores among the different majors and genders). This requirement is also known as homoscedasticity. To test for this, we examine the plot of unstandardized residuals versus the fitted model. A random display of points without patterns suggests that both this assumption and that of normality (discussed below) are met. We find that our data do conform to this, as shown in Fig. 9. If our data did not meet the assumption, we would expect to see a "fanning out" of data points (i.e., clustered along $y = 0$ on the left and broadening out in the y dimension toward the right) in this plot.

Third, ANCOVA requires the residuals be normally distributed. ANCOVA is relatively robust to violations of this assumption, so only severe deviations from normality are cause for concern. Normality is tested via the Anderson-Darling test [42,43], as well as determining the skewness and kurtosis (the third and fourth moments of the distribution) of the residuals. The Anderson-Darling test indicates normality to a significance of 1.0%. Additionally, both the skewness (-0.12 ± 0.12) and kurtosis (0.08 ± 0.13) were both near zero, an indication that neither of these effects is dominant in the residuals [50]. We further examine the plot of residuals versus fitted values (Fig. 9) visually as a test of normality. As there are no patterns in this plot (such as a parabolic pattern), normality has been met. Finally, a visual inspection of the Q-Q plot (shown in Fig. 10) also indicates that the residuals are normally distributed.

The fourth assumption of ANCOVA is overall linearity of data; since ANCOVA is a linear regression, we require that the regression of post-test score on pretest score is linear. We test this by plotting post-test score versus pretest score and fitting a line. We do not require a perfect line for this assumption but rather that our data are linear enough, meaning that it tends toward linear rather curvilinear or

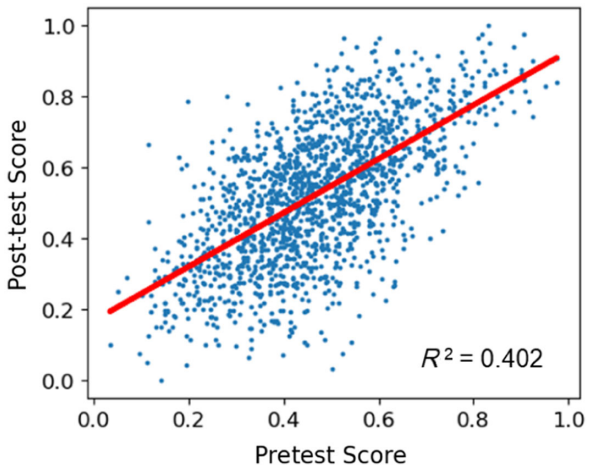


FIG. 11. Plot of post-test vs pretest overall scores (blue points) with linear fits (red line). This plot shows good evidence of linearity, therefore providing evidence that our data meet this fourth assumption for ANCOVA to be applied.

uncorrelated. A plot of post-test score vs pretest score is shown in Fig. 11. We fit a line to these data and determine $R^2 = 0.402$. Visual analysis of our data shows that it conforms to the assumption of linearity.

Next, we require that our independent variables are fixed by the researcher; this simply means that we determine the levels of the independent variable (i.e., the genders or majors) rather than randomly assigning groups. No test needs to be done to ensure our data adheres to this, due to the design of the analysis.

The sixth assumption is that the covariate and independent variables ideally would be independent from one another. In practice, in assessments, this is not the case; it is extremely common for the covariate and independent variable(s) to share variance. For example, a student's major and their pretest score are inherently linked. We accept that we must violate this assumption to some extent, with the outcome being that all partial η^2 values that describe the amount of variance explained by each of the predictors must be considered a lower bound instead of considered a particular value [62]. Thus, we might be underestimating some of the variance explained by major or gender when we report values from our ANCOVA analysis.

The next assumption is that the covariate (i.e., pretest score) must also be measured without error. In terms of assessment, this means that the assessment itself must be proven to have reliability and validity. In previous work [38], we have shown that SPRUCE has high reliability and validity, and therefore this assumption is met.

Finally, ANCOVA requires homogeneity of regression slopes. This means that we require no interaction between the covariate and the independent variable; unequal slopes would point to interactions that are being ignored. Another way of stating this is that the interaction term between the covariate (pretest) and the independent variable (major or gender) must be statistically insignificant. We find that these interaction terms are not statistically significant by building these interaction terms into our model (separately) and examining their significance via the F test (pretest: gender, $p = 0.726$; pretest:major, $p = 0.596$). Thus, our data conform to this assumption.

Based on testing of all of these assumptions, we have shown that ANCOVA is an appropriate statistical model for our data.

APPENDIX B: ORDINAL LOGISTIC REGRESSION ASSUMPTIONS AND ADHERENCE

The assumptions required to perform for ordinal logistic regression are [49,63]:

1. Independence of observations.
2. Noncollinearity of independent variables.
3. Independent variables are linear on the logit.
4. "Perfect" measurement.
5. Nonsparseness of data.

First, logistic regression requires independence of observations. This means that observations are independent of one another both within and across samples and also details sampling biases. We note that the same issues exist within this requirement as did for the same requirement in ANCOVA: no dataset collected from students will ever strictly adhere to this, but the effects are small, and this is a general issue with any assessment analysis in PER.

Next, we require noncollinearity of independent variables in cases of multiple predictors. This means that we require the pretest scores to be generally uncorrelated with gender, major, and importance to some degree; they can be slightly collinear without causing issues. In order to test this, we generate an OLS model and examine the variance inflation factor.

TABLE IX. Variance inflation factors for pretest with gender, major, or importance. These results show that they are not collinear, as all variance inflation factors are less than 10. We do not have the data for importance of D4, so it is excluded here.

	Variance inflation factor
S1 Gender	2.12
S1 Major	3.64
S1 Importance	3.14
S2 Gender	2.21
S2 Major	5.57
S2 Importance	5.63
S3 Gender	1.81
S3 Major	3.07
S3 Importance	2.94
H1 Gender	1.62
H1 Major	2.42
H1 Importance	2.33
H2 Gender	1.65
H2 Major	2.37
H2 Importance	2.30
D1 Gender	1.59
D1 Major	2.26
D1 Importance	2.31
D2 Gender	1.95
D2 Major	3.78
D2 Importance	4.00
D3 Gender	2.18
D3 Major	6.14
D3 Importance	8.64
D4 Gender	1.96
D4 Major	3.57
D4 Importance	—
D5 Gender	1.33
D5 Major	1.59
D5 Importance	1.58

In our case, this means that the pretest score cannot be highly collinear with gender, major, or importance of the AO. We determine our data meet this requirements by performing an OLS regression and examining the variance inflation factor (VIF) for each AO. VIF values greater than 10 indicate a violation [64]. VIF is defined as

$$\mathrm{VIF} = \frac{1}{1 - R^2}, \tag{B1}$$

where R^2 is the usual coefficient of determination in an OLS regression. Note that the inverse of VIF is tolerance, which is also sometimes used to determine collinearity (with thresholds, of course, being anything less than 0.1 indicating collinearity).

Our results show that our variance inflation factors are all less than 10, meaning that pretest is not collinear with major, gender, or importance, and therefore, we meet this requirement. The variance inflation factors themselves are presented in Table IX.

Further, in order to appropriately utilize logistic regression, our independent variables must be linear on the logit, meaning that they must vary linearly with the logit of the dependent variable. However, this assumption is only a requirement for continuous predictors [64], and our model for logistic regression does not have these. Therefore, this assumption is not relevant for our data.

Additionally, logistic regression requires "perfect" measurement. Typically, this means that we measure both our independent and dependent variables without error. In terms of assessment, this means that the assessment itself must be proven to have reliability and validity. In previous work [38], we have shown that SPRUCE has high reliability and validity, and therefore this assumption is met. Further, since students are self-reporting the demographics used in this analysis, we can assume that we measure this without error as well.

Finally, we aim for the data to not be sparse—that is, we aim to not have any full category cells that are empty (i.e., all categorical data intersections have at least one data point). For example, we hope that our data include cases of male physics majors at all possible scores on a particular AO for pre- and post-test. In our case, we do have some cells that are not populated. However, in these cases, the effect is that the error on the coefficients is larger and therefore underestimate the significance of some results; because we are underestimating significance (rather than overestimating), and because we cannot fix the issue of sparseness without collecting more data, we allow this condition to not be met with the caveat that in our logistic regression models, we might, in some cases, underreport significance.

Based on testing of all of these assumptions, we have shown that logistic regression is an appropriate statistical model for our data.

APPENDIX C: PREINSTRUCTION AND POSTINSTRUCTION SCORE DISTRIBUTIONS BY ASSESSMENT OBJECTIVE

In Fig. 12, we present the pretest and post-test distributions for all ten AOs on SPRUCE. Table I has information regarding each of the AOs. Due to the scoring scheme used, these AO scores can only be integers (see Table III) and therefore, these distributions are not continuous.

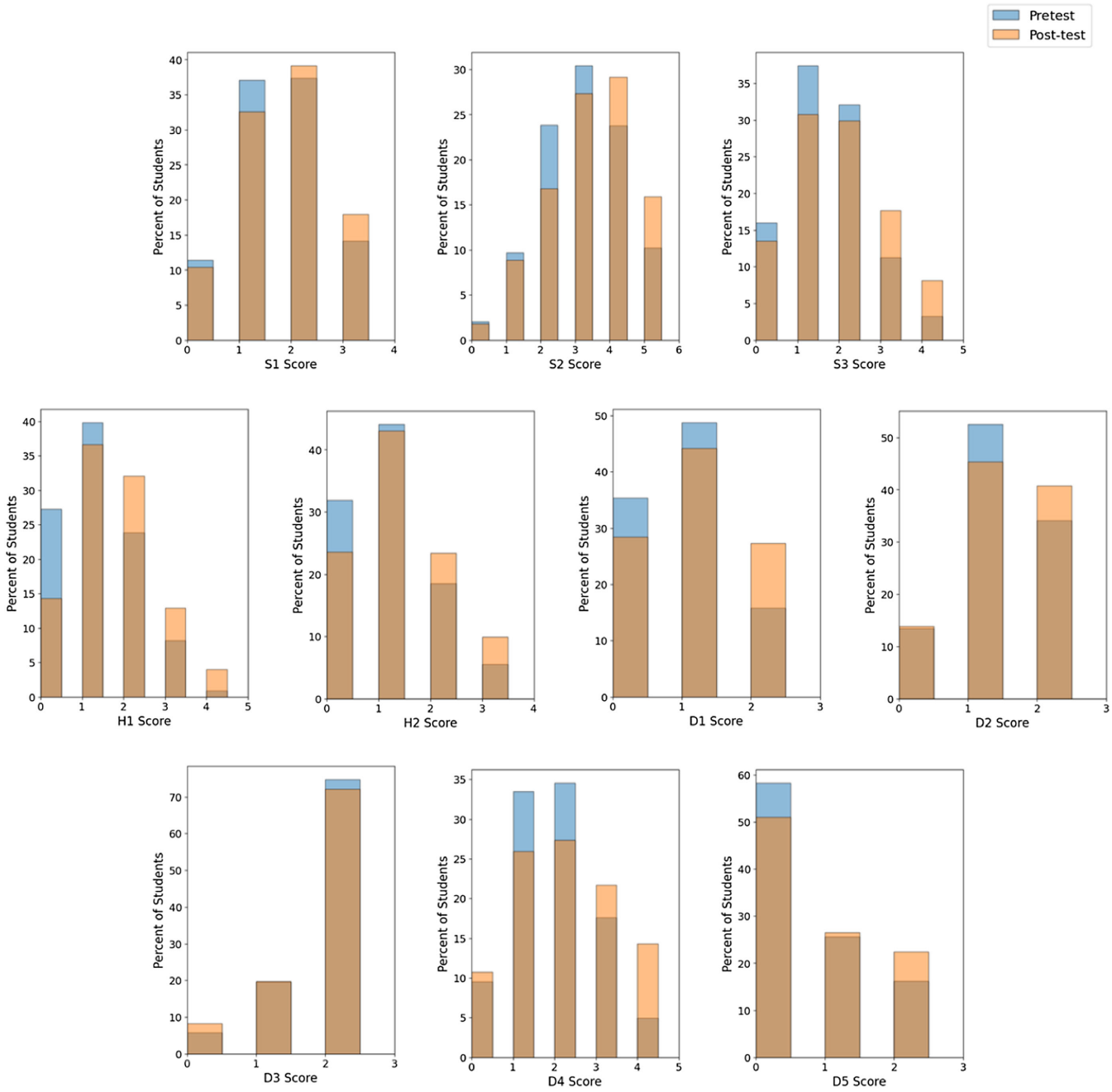


FIG. 12. Distribution of pretest (blue) and post-test (orange) scores for each assessment objective. These distributions are presented after rounding but before normalizing scores to 1.

[1] J. C. for Guides in Metrology, JCGM 100: Evaluation of measurement data—Guide to the expression of uncertainty in measurement, JCGM, Technical Report, 2008.

[2] J. Kozminski, H. Lewandowski, N. Beverly, S. Lindaas, D. Deardorff, A. Reagan, R. Dietz, R. Tagg, M. EblenZayas, and J. Williams, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (American Association of Physics Teachers, College Park, MD, 2014), Vol. 29.

[3] Analyzing and Interpreting Data | Next Generation Science Standards (2022), https://www.nextgenscience.org/practice/analyzing-and-interpreting-data.

[4] Guide to instructional laboratories and experimental skills, available at https://ep3guide.org/guide-overview/instructional-laboratories-and-experimental-skills.

[5] A. Buffler, S. Allie, and F. Lubben, The development of first year physics students' ideas about measurement in

terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).

[6] B. Campbell, F. Lubben, A. Buffler, and S. Allie, Teaching scientific measurement at university: Understanding students' ideas and laboratory curriculum reform, Monograph, Afr. J. Res. Math. Sci. Math. Educ. (2005), https://pure.york.ac.uk/portal/en/publications/teaching-scientific-measurement-at-university-understanding-stude.

[7] N. G. Holmes and C. E. Wieman, Assessing modeling in the lab: Uncertainty and measurement, in *2015 Conference on Laboratory Instruction Beyond the First Year* (American Association of Physics Teachers, College Park, MD, 2015), pp. 44–47.

[8] N. G. Holmes, C. E. Wieman, and D. A. Bonn, Teaching critical thinking, Proc. Natl. Acad. Sci. U.S.A. **112**, 11199 (2015).

[9] N. G. Holmes and D. A. Bonn, Quantitative comparisons to promote inquiry in the introductory physics lab, Phys. Teach. **53**, 352 (2015).

[10] M. M. Stein, E. M. Smith, and N. G. Holmes, Confirming what we know: Understanding questionable research practices in intro physics labs, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Stein.

[11] K. N. Quinn, C. E. Wieman, and N. G. Holmes, Interview Validation of the Physics Lab Inventory of Critical thinking (PLIC) (2018), pp. 324–327.

[12] K. W. Kok, Certain about uncertainty, Ph.D. thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2022.

[13] H. Eshach and I. Kukliansky, Developing of an instrument for assessing students' data analysis skills in the undergraduate physics laboratory, Can. J. Phys. **94**, 1205 (2016).

[14] B. Pollard, R. Hobbs, R. Henderson, M. D. Caballero, and H. Lewandowski, Introductory physics lab instructors' perspectives on measurement uncertainty, Phys. Rev. Phys. Educ. Res. **17**, 010133 (2021).

[15] M. Vignal, G. Geschwind, B. Pollard, R. Henderson, M. D. Caballero, and H. Lewandowski, Survey of physics reasoning on uncertainty concepts in experiments: An assessment of measurement uncertainty for introductory physics labs, Phys. Rev. Phys. Educ. Res. **19**, 020139 (2023).

[16] T. S. Volkwyn, S. Allie, A. Buffler, and F. Lubben, Impact of a conventional introductory laboratory course on the understanding of measurement, Phys. Rev. ST Phys. Educ. Res. **4**, 010108 (2008).

[17] C. Walsh, K. N. Quinn, and N. G. Holmes, Assessment of critical thinking in physics labs: Concurrent validity, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Walsh.

[18] C. Walsh, K. N. Quinn, C. Wieman, and N. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[19] J. Day, J. B. Stang, N. G. Holmes, D. Kumar, and D. A. Bonn, Gender gaps and gendered action in a first-year physics laboratory, Phys. Rev. Phys. Educ. Res. **12**, 020104 (2016).

[20] V. Adlakha and E. Kuo, Critical issues in statistical causal inference for observational physics education research, Phys. Rev. Phys. Educ. Res. **19**, 020160 (2023).

[21] B. R. Wilcox and H. J. Lewandowski, Research-based assessment of students' beliefs about experimental physics: When is gender a factor?, Phys. Rev. Phys. Educ. Res. **12**, 020130 (2016).

[22] C. Walsh, H. J. Lewandowski, and N. G. Holmes, Skills-focused lab instruction improves critical thinking skills and experimentation views for all students, Phys. Rev. Phys. Educ. Res. **18**, 010128 (2022).

[23] R. L. Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, Nordic Stud. Sci. Educ. **2**, 40 (2006).

[24] N. Majiet and S. Allie, Student understanding of measurement and uncertainty: Probing the mean, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.-Majiet.

[25] M. Séré, R. Journeaux, and C. Larcher, Learning the statistical analysis of measurement errors, Int. J. Sci. Educ. **15**, 427 (1993).

[26] D. Deardorff, Introductory physics students' treatment of measurement uncertainty, Ph.D. thesis, North Carolina State University, 2001.

[27] B. Pollard, R. Hobbs, J. T. Stanley, D. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab course on students' understanding of measurement uncertainty, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.073.

[28] B. Pollard, R. Hobbs, D. Dounas-Frazer, and H. J. Lewandowski, Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses, presented ar PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Pollard.

[29] B. Pollard, A. Werth, R. Hobbs, and H. J. Lewandowski, Impact of a course transformation on students' reasoning about measurement uncertainty, Phys. Rev. Phys. Educ. Res. **16**, 020160 (2020).

[30] H. J. Lewandowski, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and B. Pollard, Student reasoning about measurement uncertainty in an introductory lab course, presented at PER Conf. 2017, Cincinnati, OH, http://dx.doi.org/10.1119/perc.2017.pr.056.

[31] A. Werth, B. Pollard, R. Hobbs, and H. Lewandowski, Investigating changes in student views of measurement uncertainty in an introductory physics lab course using clustering algorithms, Phys. Rev. Phys. Educ. Res. **19**, 020146 (2023).

[32] E. M. Stump, M. Hughes, G. Passante, and N. Holmes, Comparing introductory and beyond-introductory students' reasoning about uncertainty, Phys. Rev. Phys. Educ. Res. **19**, 020147 (2023).

[33] S. Jirungnimitsakul and P. Wattanakasiwich, Assessing student understanding of measurement and uncertainty, J. Phys. Conf. Ser. **901**, 012121 (2017).

[34] J. Day and D. Bonn, Development of the concise data processing assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).

[35] I. Kontro, Development of data processing skills of physics students in intermediate laboratory courses, in *Concepts, Strategies and Models to Enhance Physics Teaching and Learning*, edited by E. McLoughlin and P.

van Kampen (Springer International Publishing, Cham, 2019), pp. 101–108.

[36] A. Schang, M. Dew, E. M. Stump, N. Holmes, and G. Passante, New perspectives on student reasoning about measurement uncertainty: More or better data, Phys. Rev. Phys. Educ. Res. **19**, 020105 (2023).

[37] E. M. Stump, M. Dew, G. Passante, and N. Holmes, Context affects student thinking about sources of uncertainty in classical and quantum mechanics, Phys. Rev. Phys. Educ. Res. **19**, 020157 (2023).

[38] G. Geschwind, M. Vignal, M. D. Caballero, and H. Lewandowski, Evidence for validity and reliability of a research-based assessment instrument on measurement uncertainty, Phys. Rev. ST Phys. Educ. Res. (to be published).

[39] R. J. Mislevy and M. M. Riconscente, Evidence-centered assessment design: Layers, structures, and terminology, SRI International Center for Technology in Learning, Technical Report 9, 2005, https://padi.sri.com/downloads/TR9_ECD.pdf.

[40] M. Vignal, K. D. Rainey, B. R. Wilcox, M. D. Caballero, and H. J. Lewandowski, Affordances of articulating assessment objectives in research-based assessment development, presented at PER Conf. 2022, Grand Rapids, MI, 10.1119/perc.2022.pr.Vignal.

[41] M. Vignal, G. Geschwind, R. Henderson, M. D. Caballero, and H. J. Lewandowski, Couplet scoring for research based assessment instruments, arXiv:2307.03099 [J. STEM Educ. Res. (to be published)].

[42] T. W. Anderson and D. A. Darling, Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes, Ann. Math. Stat. **23**, 193 (1952).

[43] L. S. Nelson, The Anderson-Darling test for normality, Journal of Quality Technology **30**, 298 (1998).

[44] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. **1**, 80 (1945).

[45] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Mahwah, NJ, 1988), pp. 66–67.

[46] G. Keppel and T. D. Wickens, *Design and Analysis: A Researcher's Handbook*, 4th ed (Pearson Prentice Hall, 2004), pp. 311–344.

[47] D. L. Hahs-Vaughn and R. G. Lomax, *An Introduction to Statistical Concepts*, 4th ed. (Routledge, London, 2019), pp. 615–635.

[48] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, Cambridge, England, 2006).

[49] J. W. Osborne, *Best Practices in Logistic Regression* (SAGE Publications, Inc., Thousand Oaks, CA, 2015).

[50] D. L. Hahs-Vaughn and R. G. Lomax, *An Introduction to Statistical Concepts*, 4th ed. (Routledge, London, 2019), pp. 126–130.

[51] G. Geschwind, M. Vignal, and H. J. Lewandowski, Representational differences in how students compare measurements, presented at PER Conf. 2023, Sacramento, CA, 10.1119/perc.2023.pr.Geschwind.

[52] P. B. Kohl and N. D. Finkelstein, Student representational competence and self-assessment when solving physics problems, Phys. Rev. ST Phys. Educ. Res. **1**, 010104 (2005).

[53] P. B. Kohl and N. D. Finkelstein, Effects of representation on students solving physics problems: A fine-grained characterization, Phys. Rev. ST Phys. Educ. Res. **2**, 010106 (2006).

[54] P. B. Kohl, D. Rosengrant, and N. D. Finkelstein, Strongly and weakly directed approaches to teaching multiple representation use in physics, Phys. Rev. ST Phys. Educ. Res. **3**, 010108 (2007).

[55] P. B. Kohl and N. D. Finkelstein, Patterns of multiple representation use by experts and novices during physics problem solving, Phys. Rev. ST Phys. Educ. Res. **4**, 010111 (2008).

[56] N. Weliweriya, T. Huynh, and E. Sayre, Standing fast: Translation among durable representations using evanescent representations in upper-division problem solving, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.103.

[57] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[58] SPRUCE for Instructors | JILA—exploring the frontiers of physics, https://jila.colorado.edu/lewandowski/research/spruce-instructors-0.

[59] J. Durbin and G. S. Watson, Testing for serial correlation in least squares regression: I, Biometrika **37**, 409 (1950).

[60] J. Durbin and G. S. Watson, Testing for serial correlation in least squares regression. II, Biometrika **38**, 159 (1951).

[61] J. Durbin and G. S. Watson, Testing for serial correlation in least squares regression. III, Biometrika **58**, 1 (1971).

[62] G. A. Miller and J. P. Chapman, Misunderstanding analysis of covariance, J. Abnorm. Psychol. **110**, 40 (2001).

[63] D. L. Hahs-Vaughn and R. G. Lomax, *An Introduction to Statistical Concepts*, 4th ed. (Routledge, London, 2019) pp. 997–1063.

[64] D. L. Hahs-Vaughn and R. G. Lomax, *An Introduction to Statistical Concepts*, 4th ed. (Routledge, London, 2019), pp. 1017–1018.