Evidence for validity and reliability of a research-based assessment instrument on measurement uncertainty

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 16 May 2024; accepted 12 September 2024; published 4 October 2024)

The Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) was designed to measure students' proficiency with measurement uncertainty concepts and practices across ten different assessment objectives to help facilitate the improvement of laboratory instruction focused on this important topic. To ensure the reliability and validity of this assessment, we conducted a comprehensive statistical analysis using classical test theory. This analysis includes an evaluation of the test as a whole, as well as an in-depth examination of individual items and assessment objectives. We make use of a previously reported on scoring scheme involving pairing items with assessment objectives, creating a new unit for statistical analysis referred to as a "couplet." The findings from our analysis provide evidence for the reliability and validity of SPRUCE as an assessment tool for undergraduate physics labs. This increases both instructors' and researchers' confidence in using SPRUCE for measuring students' proficiency with measurement uncertainty concepts and practices to ultimately improve laboratory instruction. Additionally, our results using couplets and assessment objectives demonstrate how these can be used with traditional classic test theory analysis.

DOI: 10.1103/PhysRevPhysEducRes.20.020125

I. INTRODUCTION

Improving physics instruction at the undergraduate level has been a longstanding goal within the community of physics educators and physics education researchers. However, assessing existing teaching practices to facilitate meaningful enhancements remains a difficult task. Research-based assessment instruments (RBAIs) are a vital tool to help assess the effectiveness of instruction in physics courses. Many RBAIs have been developed for use in a wide variety of physics lecture courses, such as Newtonian mechanics [1,2], thermodynamics [3], electricity and magnetism [4,5], and quantum mechanics [6,7], as well as for lab courses regarding critical thinking [8], handling of measurement uncertainty [9,10], handling of data [11], modeling [12], and views about experimental physics [13]. As lab courses have a large range of varied learning goals, there continues to be a need for more research-based assessment tools spanning this space.

To address the need for assessment tools designed for laboratory courses, we recently designed the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE). This assessment provides a measure of physics laboratory students' proficiency with measurement uncertainty concepts and practices [14,15].

Prior research on these topics has revealed challenges, such as students' lack of understanding regarding the importance of taking multiple measurements during experimentation [16] and the belief in a singular "true" value [17,18]. Inauthentic lab practices, such as artificially inflating uncertainties to match their experimental results to theoretical values, likely also contributes to student challenges with understanding measurement uncertainty [19]. Based on this research, we created SPRUCE in order to better quantify and characterize students' ideas around measurement uncertainty.

A crucial aspect of developing RBAIs, such as SPRUCE, involves establishing evidence for the validity and reliability of the instrument. This evidence is taken into account at every phase of the development process, starting from defining the scope of what the instrument will measure, progressing through item creation, and ultimately extending to the utilization of statistical testing on student responses.

Previously, we have shown that researchers can map student responses to different reasoning elements for each answer option on all SPRUCE items. Additionally, we showed evidence for content validity, via instructor input, and face validity, via item creation and alignment with specific objectives. More detail about each of these types of validity, as well as an in-depth analysis of the evidence we have for each of these from SPRUCE can be found in our previous work [15].

Evaluating the external validity of assessments is a well-established practice [20], and is important before a full deployment of the instrument in order to assure the accuracy of results obtained. Here, we use classical test theory (CTT) to provide that evidence for SPRUCE. We will discuss the validity of SPRUCE, based on various CTT metrics, such as discrimination, stability, and internal consistency. We discuss these metrics for the entire SPRUCE assessment in addition to a component-by-component analysis. Unlike traditionally scored assessments, where an item (question) would serve as the component, SPRUCE is scored using couplets. In couplet scoring, each item is scored separately based on each assessment objective (AO) it probes, and most items probe more than one AO. An AO can be thought of as a single concept the assessment tool aims to measure, or "*concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment* [21]." For example, one of SPRUCE's AOs is *Articulate why it is important to take several measurements during experimentation*. For SPRUCE, the couplet, which is the item score for a particular AO, is then the unit of analysis. In addition to providing evidence on the validity of SPRUCE, we will demonstrate how CTT can be used with couplet scoring.

Our research questions for this work include

1. RQ1: What is the evidence that SPRUCE is a reliable and valid assessment tool for the population included in the study?
2. RQ2: How can we adapt CTT for an assessment that uses couplet scoring?

The results of the analysis presented here will allow for future studies on student learning of measurement uncertainty using SPRUCE as a tool, as well as serve as an example for adapting CTT to an assessment which utilizes couplet scoring.

II. BACKGROUND

A. RBAIs in physics

Research-based assessment instruments are essential tools used by educators to help evaluate and improve instruction. These assessments are developed by identifying instructor priorities and student thinking in order to create a tool that can be used by the wider community [22]. Further, RBAIs allow researchers to compare instructional outcomes across many institutions and courses, and can also be used to evaluate the effectiveness of course transformations. However, they are specifically not intended to evaluate or to grade individual students. Instead, their intended use is in aggregate, to examine populations of students.

Widely used examples of RBAIs in physics include: the Force Concept Inventory (FCI) [1] and the Force and Motion Conceptual Evaluation (FMCE) [2], both designed to evaluate introductory physics students' understanding of simple Newtonian mechanics; the Physics Measurement Questionnaire (PMQ) [9], the Physics Lab Inventory of Critical Thinking (PLIC) [8,23–25], and the Concise Data Processing Assessment (CDPA) [11], intended to evaluate students' handling of measurement uncertainty and general experimental skills; and the Colorado Learning Attitudes about Science Survey (CLASS) [26] and Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) [13], which both evaluate student attitudes and beliefs about science in different contexts.

These assessments have dramatically altered the landscape of physics education at the undergraduate level. For example, the FCI showed a clear lack of conceptual understanding of basic introductory physics and helped introduce changes to the standard didactic lecture form of instruction [1,27].

RBAIs are often created through a rigorous development process. One such development process is evidence-centered design (ECD) [28]. This framework was used to guide the design and implementation of SPRUCE. Other examples of assessment frameworks include the three dimensional learning assessment protocol [29] and the framework described by Adams and Wieman [30]. All of these frameworks outline assessment development and design, including steps for exploratory research on the assessment topic, item development and refinement, and distribution and validation.

B. SPRUCE

SPRUCE aims to measure students' proficiency with measurement uncertainty concepts and practices. While some assessments, such as the PMQ [9], the PLIC [8,23–25], and the CDPA [11] also aim to measure introductory laboratory students' ideas around laboratory skills and measurement uncertainty, none of the three fills the specific space SPRUCE is designed for. We previously discussed the affordances and limitations of these other instruments [15]. SPRUCE aims to fill this gap in assessments by offering a test that is focused solely on measurement uncertainty skills, broad in its coverage of measurement uncertainty topics, widely administrable, easily scorable, and designed for lower-division (first two years of college) physics labs.

SPRUCE is administered in a fully online format that takes students about 15 minutes to complete. There are six distinct response formats for items on SPRUCE: multiple choice, multiple response, numeric open response, coupled multiple choice, coupled multiple response [31], and coupled numeric open response. More information about the development of the items and the types of items on SPRUCE is discussed in Vignal *et al.* [15].

SPRUCE consists of 19 items grounded in four experiments, which probe 10 AOs. The SPRUCE AOs are shown in Table I and are organized into three categories of measurement uncertainty: sources of uncertainty, handling uncertainty, and distributions and repeated measurements. These were designed based on findings from instructor interviews [14]. AOs provide many affordances, as detailed in [15,21].

For clarity, we note that these AOs are slightly different than those previously reported for SPRUCE [15]. During the process of scoring and validating the assessment, we determined that collapsing some of the AOs together created more reliable results. Previously, we had 14 AOs. We collapsed three objectives that all dealt with standard error and standard deviation into one AO (D4, see Table I). Additionally, we collapsed three AOs that all handled error propagation using equations. We believe that ten constructs, rather than 14, provides a more sound basis for validation.

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

	Sources of uncertainty
S1	Estimate size of random or 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

SPRUCE was designed by iterating through the ECD framework [28]. As described in previous work [14,15], a process of iterative steps were taken in order to understand the important aspects of measurement uncertainty in the introductory laboratory community, determine a set of areas to probe with the assessment (which eventually turned into AOs), write items for the assessment, and refine these items based on a series of student interviews and beta testing. The next stage is validation of the assessment, which this paper aims to provide.

As part of ECD, after item creation, SPRUCE was beta tested through online administration in several courses, as well as through student interviews to determine reasoning elements for all correct and incorrect answer options. Thus, for each item on SPRUCE, we are confident about student reasoning for each answer option they could select [15]. This important step of determining evidentiary reasoning is a critical part of the ECD process.

C. Classical test theory

Classical test theory is an important validation tool for RBAIs. It helps researchers determine whether the assessment they have created has evidence for validity: i.e., is the assessment evaluating what we think it is in a meaningful way. The underlying theory assumes that the total test score consists of two components: a true score and some random error [32]. These three factors (the total test score, the true score, and the random error) and the relationships between them can be used to determine various information about the quality of the assessment.

According to Englehardt, a high quality test must have reliability, validity, discrimination, comparative data, and be tailored to the population one hopes to measure [33]. In order to determine whether SPRUCE is a high quality test, we examine these requirements.

Four of these five qualities—reliability, validity, discrimination, and suitability for the intended introductory laboratory audience—are the main focuses of this paper. We will discuss each of these important aspects of conducting a thorough CTT analysis of SPRUCE. We are currently in the process of collecting a large database of comparative data to fulfill the last part of Englehardt's requirements for a high quality assessment. Below, we define these qualities as they are used for CTT.

Reliability describes how consistently an assessment measures what it is intended to measure (e.g., student proficiencies, in the case of SPRUCE). In other words, if a student takes the same assessment multiple times without recalling previous attempts, they should get the same score each time (assuming no new learning happens) in order for an assessment to be considered reliable. Further, reliability is dependent on the students taking the assessment—if they have a wide range of levels of proficiency, the reliability will be higher than if they have a narrow range [34].

To address this, we administered SPRUCE in a wide variety of courses at many different types of institutions.

Validity is related to the conclusions researchers can draw from the scores students get on the assessment. Statistical validations quantify how well the assessment measures the specific topics it is intended to.

Discrimination refers to the assessment's ability to distinguish between high and low student performance, both on the scale of the full assessment, as well as on the scale of each individual item.

Finally, suitability for the intended audience indicates a need for an assessment designed with the target population in mind. For example, giving introductory physics students an assessment with graduate-level questions will result in poor performance for all students and therefore the data will not be useful to instructors. Further, CTT does not handle floor and ceiling effects well—if many students are at a very high or very low range of scores, CTT is inappropriate [34], which is another reason the test should be targeted appropriately. It is also important to note that this quality is closely related to discrimination—if the test is too difficult for all of the students, then it will not discriminate well.

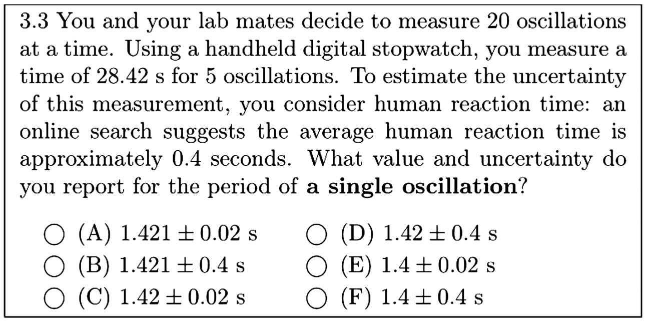
While many researchers are turning to item response theory (IRT) to validate assessments [32], CTT is an important first step before further validating the assessment using other methods. Additionally, CTT requires considerably fewer data than IRT.

D. Scoring by couplet

As discussed in recent work [35], SPRUCE uses a scoring paradigm that takes into account assessment objectives (AOs) for each item. There are 19 items on SPRUCE and ten AOs (see Table I). Each item addresses between two and five of the AOs covered by SPRUCE. Instead of simply scoring each item once and calculating an overall assessment score on SPRUCE for each student by adding together all item scores, the items are scored once per AO they address and average AO scores are presented to instructors in a final report. We refer to the individual item-AO pairs as couplets.

An example item (item number 3.3) from SPRUCE is shown in Fig. 1 to illustrate the scoring method. This item addresses two AOs: *H1: Propagate uncertainties using formulas* and *H2: Report results with uncertainties with correct significant digits*. Table II shows a breakdown of the scoring scheme for this item.

The first AO assesses whether students can identify the proper method of error propagation, in this case division by 10. If a student selects A, C, or E, they have correctly propagated the uncertainty and therefore receive credit for the couplet. The other AO assesses whether students report results with proper significant figures. If students select C or F, they have demonstrated understanding of significant figures, and thus receive credit for this couplet. We therefore score this item twice: first, couplet item 3.3—AO H1 (or simply 3.3 H1), and second, couplet item 3.3—AO H2 (or 3.3 H2). A student who selects C receives full credit (one point) for each couplet. A student who selects A, E, or F receives credit for only one couplet. A student who selects B or D receives no credit on either couplet. Thus, one item on the assessment becomes two independent couplets in terms of scoring. Students only have to answer this item once, but we are able to assess their understanding across multiple skills independently.



3.3 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 s for 5 oscillations. To estimate the uncertainty of this measurement, you consider human reaction time: an online search suggests the average human reaction time is approximately 0.4 seconds. What value and uncertainty do you report for the period of **a single oscillation**?

○ (A) 1.421 ± 0.02 s ○ (D) 1.42 ± 0.4 s
○ (B) 1.421 ± 0.4 s ○ (E) 1.4 ± 0.02 s
○ (C) 1.42 ± 0.02 s ○ (F) 1.4 ± 0.4 s

FIG. 1. Example SPRUCE item 3.3. This item addresses two assessment objectives—one regarding error propagation and the other regarding correct use of significant figures—with different correct answers for each. The exact numbers have been changed to protect the security of the assessment.

We complete a similar process for each item: an item was compared to the list of SPRUCE AOs, matched appropriately, and scored based on each AO that item addressed. This led to 31 item-AO couplets from 19 items. Instead of item scores, these couplet scores form the basis unit of scoring and are used in the statistical validation presented in this paper. Vignal *et al.* [35] present a more in-depth analysis and discussion of this scoring scheme.

This method of scoring serves many purposes that are discussed in more detail in our recent work [35]. This scoring scheme helps reduce the number of questions students answer—despite students only having to answer 19 items, we are able to score along 31 couplets, thereby increasing the amount of information researchers can extract about student understanding, while keeping the actual assessment a reasonable length. Additionally,

TABLE II. Example scoring for couplets of item 3.3.

	Score	
Answer option	H2	H3
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

we show in this work that the base unit of scoring—the couplet—is able to be treated similarly to item scores for validating assessments.

III. METHODS

A. Establishing validity

In order to ensure the content validity of SPRUCE, or, in other words, that the entire assessment measures the intended content domain, we worked closely with instructors through all phases of development. SPRUCE was developed using an evidence-centered design process, or ECD [28], a framework for creating RBAIs. The first phase of this was interviewing introductory laboratory instructors to develop the objectives of SPRUCE. These instructors indicated which areas of measurement uncertainty are most important for their students to learn, which led to the initial set of assessment objectives; these were then refined during further development of SPRUCE. Further, we had independent researchers map the SPRUCE items to the objectives to ensure full coverage. Face validity, or ensuring that items measure their intended constructs, was similarly determined by this matching process, as well as via soliciting instructor feedback during the entire item development process. Finally, external validity deals with generalizing results beyond the pilot population. This type of validity is the focus of the work presented here.

B. Data collection and cleaning

We recruited instructors across 22 institutions teaching 36 different courses to administer SPRUCE during the Fall 2022, Spring 2023, and Fall 2023 semesters using a pre-post format. We solicited instructors who had previously expressed interest in this project, as well as by posting advertisements on the Advanced Laboratory Physics Association (ALPhA) listserv and two American Physical Society discussion boards (forum on education and topical group on physics education research). For the Spring 2023 and Fall 2023 semesters, we also required instructors to fill out a brief survey about their course for future work on analyzing the results of SPRUCE in comparison with information about the course itself (e.g., examining the differences in student learning gains between courses intended for physics majors and courses intended for other science majors). Table III shows brief information about the institutions that administered SPRUCE.

TABLE III. SPRUCE Institution types ($N = 22$). For each of the 36 courses at 22 institutions during the Fall 2022, Spring 2023, and Fall 2023 semesters, we present information about the highest degree of the institutions, as well as the numbers of institutions that are minority serving. HSI indicates a Hispanic serving institution and AANAPISI indicates an Asian American and Native American Pacific Islander serving institution.

	Number of institutions	Number of students
Highest degree		
Ph.D.	7	2152
Master’s	4	325
Bachelor’s	9	86
Associate’s	2	33
Minority serving status		
HSI	4	48
AANAPISI	1	29

In the work presented here, we used only the post-test data from all three semesters. Validation results typically do not take into account pretest data because students have not yet learned the material they are being tested on. From these three semesters of post-test data, we received 3644 responses, of which 2596 were analyzed. Students were excluded from analysis for not consenting to have their data used in research, not answering the filter question (i.e., they closed the assessment before making it that far), or answering the filter question incorrectly. We also excluded duplicates (i.e., students who took SPRUCE more than once—only their final attempt is included in this analysis). This led to excluding 28.8% of the responses received (of which 460 or 43.9% of exclusions were due to nonconsent to research).

Additionally, we also collected expert responses in order to establish validity of our scoring scheme. We asked faculty members at a wide variety of institutions, as well as those on the ALPhA email list, to anonymously take the assessment. We also provided them a text box for additional feedback they might have. We specifically targeted instructors of introductory laboratory courses, as well as physicists who run experimental research groups. We received 36 complete responses from these experts.

C. SPRUCE scoring scheme and CTT

The scoring scheme we used for SPRUCE as discussed in Sec. II D takes into account the fact that SPRUCE is a multiconstruct assessment. Similar to the CDPA [11], the PLIC [25], and the FCI [36], instead of probing one single topic, SPRUCE assesses a variety of topics, in this case all under the umbrella of the topic of measurement uncertainty. The method of constructing scores from a student answer to an item all the way through to an overall assessment score is depicted in Fig. 2 and further explained below.

First, students answer assessment items in the usual way (since couplet scoring does not change how the instrument appears to students), as shown in the lowest layer of Fig. 2. We then use these answers to score item-AO couplets, as described in the methods section and shown in the second-lowest layer of Fig. 2. It is important to note that, for this work, the individual unit of scoring is the couplet, rather than the item, as is the case for traditional scoring schemes. Items may be scored multiple times in the couplet-scoring

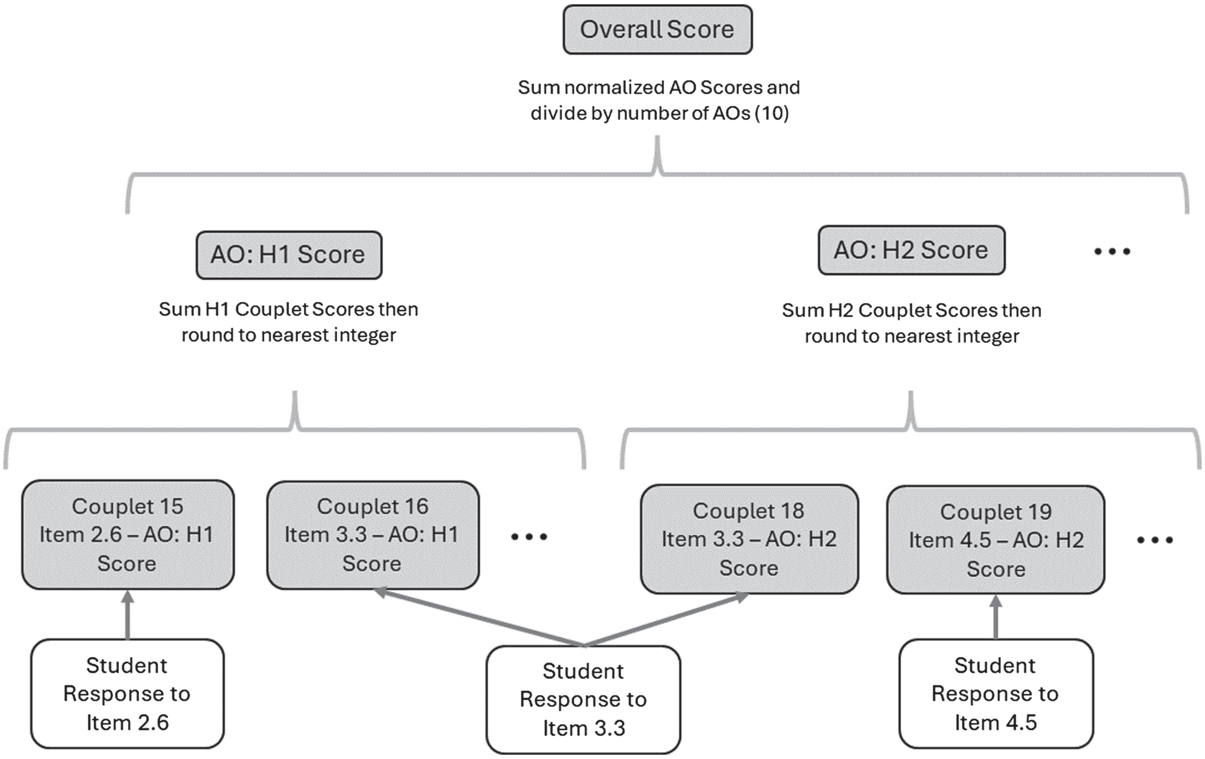

Overall Score
Sum normalized AO Scores and divide by number of AOs (10)
AO: H1 Score
Sum H1 Couplet Scores then round to nearest integer
AO: H2 Score
Sum H2 Couplet Scores then round to nearest integer
• • •
Couplet 15 Item 2.6 – AO: H1 Score
Couplet 16 Item 3.3 – AO: H1 Score
• • •
Couplet 18 Item 3.3 – AO: H2 Score
Couplet 19 Item 4.5 – AO: H2 Score
• • •
Student Response to Item 2.6
Student Response to Item 3.3
Student Response to Item 4.5

FIG. 2. Flowchart indicating how to proceed from student responses to an item to couplet scores, AO scores, and an overall score. Students respond to items, which are then paired with AOs. These item-AO pairs are scored as couplets. The couplet scores for each AO are then summed, rounded, and normalized to 1. Finally, the AO scores themselves are summed and normalized to form an overall score. This simplified flowchart illustrates only some of the couplets scored for two SPRUCE AOs, H1 and H2; in scoring SPRUCE, more couplets than indicated are included in these two AOs, and all ten AOs are included in the overall score.

paradigm. A couplet is one such score on an item along a specific AO. These couplets form the base scoring units to which classical test theory is applied.

While most of the couplets are given either full credit (1 point) or no credit (0 points), 10 of the 31 couplets—resulting from three items—allow for partial credit in 0.25 point increments. These three items are all in the coupled multiple-response format. A list of number of couplets associated with each AO, as well as possible un-normalized scores on that AO are shown in Table IV.

After all couplets are scored, we can then form AO scores by summing the couplet scores for each AO individually, which are shown as the second layer in Fig. 2. After we calculate each student's AO score, we round it to the nearest integer. We do this after calculating an AO score, rather than at the couplet level (i.e., we do not round the couplet scores) because it allows for a more fine-grained examination of scores. It also allows students to get, for example, 0.25 points on four different couplets, which could then add a point to that AO score, rather than rounding all of those down to zero before calculating the AO score. Rounding the AO scores allows us to make more comparisons between the different AOs without losing information; all of the CTT statistics presented below were calculated with and without rounding, and with multiple different methods of rounding. Aside from difficulty, which changes as one might expect (rounding up brings the scores up), none of the other statistics—including measures of discrimination—change in a statistically significant way due to rounding, either to the half integer or to the integer.

TABLE IV. AO score possibilities both before and after rounding. Each AO is targeted by a different number of couplets, and therefore has a different total possible score. 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.

AO	Num. 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]

TABLE V. Statistics at each level that we present in this work. Which CTT statistics are calculated differ based on which level of the assessment they are applied to—the individual couplets, the AOs, or the overall assessment score.

	Assessment	AO	Couplet
Difficulty	✓	✓	✓
Ferguson's delta	✓	···	···
Discrimination index	···	✓	✓
Pearson coefficient	···	✓	✓
Cronbach's alpha	✓	···	···
Test–retest stability	✓	···	···
Split-halves reliability	✓	···	···

Thus, we choose to round to the nearest integer, with 0.5 rounding up.

We gather ten AO scores, one for each AO. From this, we then compute an overall score by simply normalizing and then summing these ten AO scores, as shown in the top layer in Fig. 2. We calculate the overall score in this way because it weights each AO equally, which is more desirable than weighting the AOs differently depending on how many times each is probed. This would result in artifacts from test construction heavily biasing the score towards certain AOs. We provide statistical evidence that this method of reporting the overall score—normalizing and then summing the AO subscores rather than simply adding up couplet scores—provides a valid and reliable score.

In the following section, we report descriptions of statistical validation, as well as the results we obtain from these tests. A summary of these statistics and to which level of score they are applied—the entire assessment, the AO level, or the couplet level—is presented in Table V.

IV. RESULTS AND DISCUSSION

Below, we discuss the results of applying CTT to SPRUCE; a summary of these results can be found in Table VII.

A. Analysis of instructor responses

In order to establish expert alignment with our scoring scheme, we collected 36 responses from experts, with data collection for this section described above.

Of these responses, we analyzed only 27. We excluded nine responses from our analysis after implementing a system to exclude responses based on incorrect answers to the most straightforward questions. For example, one such item presented four stereotypical "bullseye" targets showing different levels of accuracy and precision and asks the user which bullseye represents high precision and low accuracy. All items used to exclude expert responses were either multiple choice or multiple response; we explicitly chose not to use open response items for this. We excluded responses with two or more incorrect answers to this subset of questions from our analysis.

After removing these responses, we calculated an average score for each couplet from the 27 responses. Only three couplets had less than 80% correct: couplets 3, 12, and 30 (see Appendix for couplet numbering).

Couplet 12 asked about the impact on both accuracy and precision when going from 200 to 1000 measurements. Issues with this couplet were addressed by updating the wording in the question statement to clarify it, based on both a lower average than anticipated, as well as feedback received in the feedback box at the end of the assessment. We believe with this added clarification experts would be better able to answer this item appropriately.

Couplet 30 required students (and experts) to compare numerical measurements with uncertainty. It is a different representation as couplet 31, but with identical comparisons presented in both. However, couplet 31 is presented pictorially rather than numerically. A full analysis of student responses to these couplets is described in prior work [37]. The expert average on the pictorial version of this couplet was 81%. Therefore, we believe that the low average of 59% on the numerical version of the couplet is due to experts moving too quickly through the assessment rather than taking the time to do the calculations required for this item. Because experts agree with our answer when given the same item pictorially, we elected not to change the numerical version of the item. Further, many of the incorrect answers received for the numerical version of the item selected an answer where the two measurements being compared were vastly different, with error bars very far apart, another indication that experts did not fully engage with this item.

Finally, couplet 3 is a coupled multiple-choice item that presents an unusual experimental setup. In this item, students are given a piece of string and have to determine how much mass hanging off the string will break it; the masses they have are given in 20 g increments. The string does not break under a 520 g load, but does break under a 540 g load. Experts and students have to correctly answer two multiple choice items in order to receive credit for this couplet—one for the breaking mass and one for the uncertainty in that mass. The correct answer is 530 g ± 10 g (where they must put 530 g for one question on the assessment asking about mass, and 10 g on the next which asks about uncertainty). The most common incorrect answer (six out of the 11 incorrect responses) was 520 g ± 20 g. We believe that this answer is clearly incorrect, because the string did not break at 520 g, so it is clearly able to support weights between 500 and 520 g; these values should not be included in a final determination of the breaking mass of the string. Because the situation is not typical as one does not directly measure the mean value, we chose to keep this couplet as is in the assessment.

B. Overall score

From three semesters of data collection, we analyzed a total of 2596 post-test responses to SPRUCE. In Fig. 3,

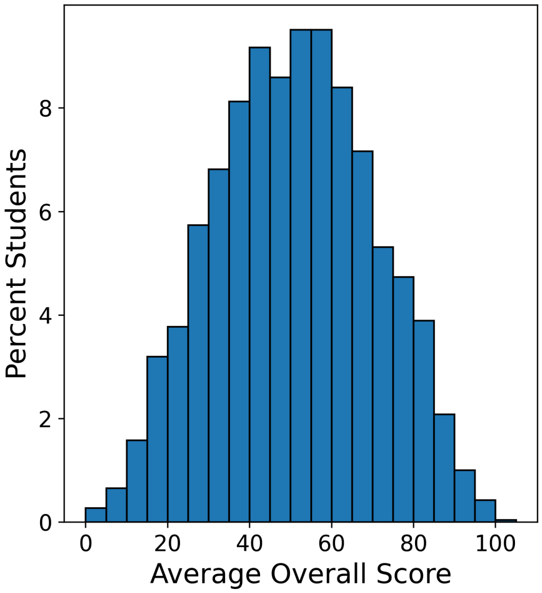


FIG. 3. Histogram showing the distribution of overall post-test scores on SPRUCE ($N = 2596$). This distribution is normal, as determined by the Anderson-Darling test, skewness, and kurtosis.

we show a distribution of overall assessment scores. The overall score is calculated as discussed above. The average overall score for all students, normalized to 100, is 50.9 ± 0.4, with a standard deviation of 19.1.

Based on the overall score statistics, the assessment appears to be properly tailored to the population in that the scores cover a wide range and the average score is about 50%. This means that the assessment is neither too easy nor too difficult for the intended introductory physics student population. The general guidance is for this average score to be between 30% and 80%. Further, the range of scores (from 0.0 to 100, once normalized to 100) covers the entire spread of possible scores, which lends evidence that the discrimination of the assessment is good.

Finally, we tested the distribution of overall scores for normality via the Anderson-Darling test [38,39], as well as determined the skewness and kurtosis (the third and fourth moments of the distribution). We do this because normal data are simpler to analyze in most cases. The Anderson-Darling test shows that the data are normal to a significance level of 1.0%. The skewness is 0.010 with 95% confidence interval [−0.084, 0.105], and the kurtosis is −0.577 with 95% confidence interval [−0.673, −0.481]. We conclude that the distribution of the overall score data is very close to normal as the skewness and kurtosis are between −1 and 1 [40].[1]

C. Internal consistency: Matching assessment objectives and items

A key component of the validation process is to ensure the assessment is measuring what we believe it measures with respect to our AOs. In our case, part of that validation is to determine whether there is expert agreement on which AOs are probed by each item. In order to validate our matching of the SPRUCE items with their corresponding AOs, we performed interrater reliability (IRR) testing on assignment of the couplets. We provided two independent (i.e., had never worked on SPRUCE) physics education researchers with Ph.D.s in experimental science with a list of SPRUCE items and AOs, and we asked them to list all AOs they believed are probed by each item. We obtained 91% agreement with our matching of AOs initially, which then rose to 99% after brief conversations to clarify specifics about some of the AOs. For example, one such clarification was in regards to AO D3: *Calculate and report the mean of a distribution for the best estimate of the measurement*. The raters coded some items with this AO that related to calculating a mean, but did not require students to report it. After discussion, there was full agreement on couplets containing this AO. Through this IRR process, we were able to demonstrate that the assignment of AOs to items was robust and that the items do in fact probe the AOs they have been assigned. Note that this method of interrater reliability was performed due to the fact that there are too few items per AO to perform statistical analysis, such as a Cronbach's alpha calculation within each AO to determine the internal consistency of all of the items within a particular AO.

D. Difficulty

Difficulty is simply a measure of the average score, which can be calculated for a couplet, an AO, or the entire assessment. The entire assessment difficulty (i.e., the average overall score) was discussed previously.

Couplet difficulty is a measure of how many students got the answer to each item-AO couplet correct. In other words, the difficulty on each couplet is its average score, which falls between zero and one. Note that this means a higher difficulty indicates an *easier* couplet, which can be slightly counterintuitive. Couplets with difficulty values of about 0.50 are generally the best for discrimination, although this is not always the case. This idealized difficulty of 0.50 also assumes couplets are not correlated with each other, which is not true for SPRUCE due to the nature of scoring some items more than once, and also because many of the AOs are often conceptually related to one another. We aim to have the couplet difficulty be between ~0.25 and 0.9 [41]. If the difficulty is greater than 0.9, the couplet may be too easy, and if the difficulty is less than 0.25, the couplet may be too difficult. Caution must be taken here because many of the multiple choice items on SPRUCE have more than four potential answer options, and therefore this general

[1] 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 two to four; this corresponds to our values when we subtract three.

statement about item difficulty is not always applicable, especially because the main rationale for removing very easy and very difficult items is because they are generally poor for discrimination [32,41].

Doran describes an additional schema for determining "good" values of difficulty, where a distribution of difficulties amongst the items is ideal [41]. This distribution should be tailored for the intent of the assessment and the level of instruction. Instead of hard cutoffs, Doran *et al.* recommend having questions of varying difficulty at all levels. To this end, we show the distribution of couplet difficulties in Fig. 4, which shows a good spread of difficulties in line with Doran's advice. Additionally, a low difficulty value might indicate a couplet that is useful, but addresses an area that students struggle with or improvements in instruction are needed. Regardless, we used these numbers as an indication to investigate couplets that fall outside this range in order to determine why this may have happened and whether the couplets should be kept as is, changed in some way, or removed.

Individual couplet difficulties are provided in Appendix in Table VIII. They fall between the values of 0.22 and 0.86, which is reasonable by the above cutoffs and schema, especially because many of the low and high difficulty couplets have reasonable discrimination values. The average couplet difficulty was 0.49 ± 0.03, again showing an acceptable level of whole-test difficulty.

Next, we present AO-level difficulty in Table VI. AO-level difficulty refers to the average score on each AO after it has been normalized to one. Similar to couplet difficulties, we aim to have a reasonable spread of AO difficulties, with a desired average of around 0.50, which would indicate that the assessment is designed appropriately for the desired student population.

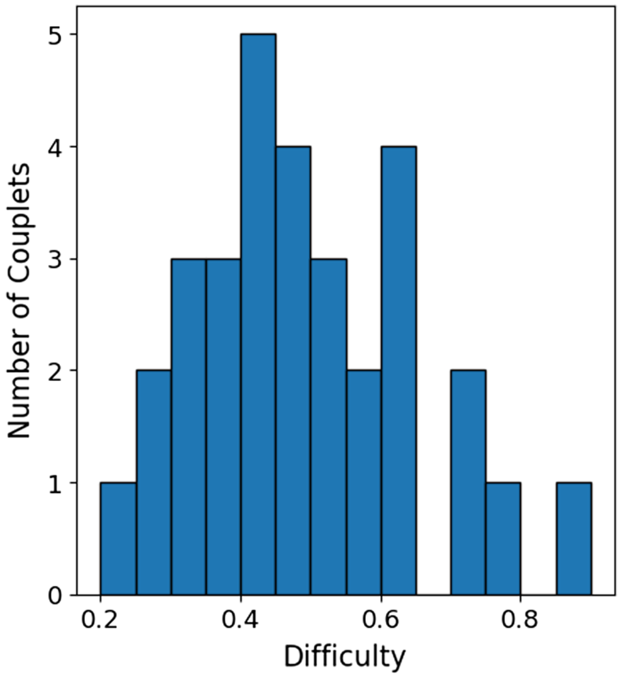


FIG. 4. Individual couplet difficulties. The histogram shows the distribution of couplet difficulties. A large spread of difficulties, as seen here, is a sign of a robust assessment.

TABLE VI. Statistics at the AO level. This table presents the difficulty, discrimination index, and Pearson coefficient for each of the ten AO scores. Error presented is standard error, shown as uncertainty in the last digit [e.g., $0.54(1) = 0.54 \pm 0.01$].

AO	Difficulty	Discrimination index	Pearson coefficient
S1	0.54(1)	0.42	0.56(1)
S2	0.62(1)	0.43	0.68(1)
S3	0.43(1)	0.50	0.71(1)
H1	0.38(1)	0.30	0.49(2)
H2	0.40(1)	0.35	0.48(2)
D1	0.49(1)	0.70	0.74(1)
D2	0.62(1)	0.62	0.71(1)
D3	0.79(1)	0.46	0.57(1)
D4	0.49(1)	0.47	0.62(2)
D5	0.33(1)	0.51	0.52(1)
Average	0.509(4)	0.48(4)	0.71(1)

E. Discrimination

Discrimination refers to how well an assessment can distinguish between high and low student performance in a particular area. This can be calculated for the assessment as a whole, at the AO level, and for each individual couplet.

1. Overall test discrimination

First, we determine Ferguson's delta, a measure of the discriminatory power of the entire test. It determines how broadly overall scores are distributed over the entire possible range. A broader distribution indicates a test that is likely better at discriminating between students at different levels [32]. To determine this measure, we use the equation outlined in Ding and Beichner [32]:

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - N^2/(K+1)}, \qquad (1)$$

where N represents the total number of students in the dataset (in our case, 2596), K is the number of AOs (10), and f_i is the number of students whose overall score is i [32]. The Ferguson's delta for SPRUCE is $\delta = 0.947$. Because this is above 0.90 [42], we conclude that SPRUCE, as an entire assessment, provides good discrimination among students. In calculating Ferguson's delta for this assessment, we binned overall scores out of 10 into single integer bins (e.g., [0,1), [1,2), etc.). This is necessary for the statistic to be calculated.

2. Couplet-level discrimination

Couplet discrimination measures the power of a couplet in distinguishing between high and low student performance on the assessment as a whole. It is a correlation between performance on a particular couplet and performance on the entire SPRUCE assessment. We calculate couplet discrimination with two separate methods. First, we calculated the discrimination index, D. This is done by using data from only the top and bottom 27% of performers [33] on the assessment as a whole. The discrimination index for each couplet is the difference in the average couplet score for students in the top 27% minus the average couplet score for students in bottom 27%. Note that anything above about 0.3 [41] is considered to be good discrimination. When calculated using this method, the discrimination may be between −1 and 1. A negative discrimination on an item indicates that students who did worse on the assessment overall did better on that particular couplet. Results of the discrimination index for each couplet are shown in Appendix.

The second method of calculating discrimination is using the Pearson coefficient, which is as follows:

$$r = \frac{\sum(x_i - \bar{x}) \sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \tag{2}$$

where the x_i refers to the score on a particular couplet by the ith student, $\bar{x}$ refers to the mean couplet score for a particular couplet being examined, y_i refers to the overall score on the assessment for the ith student, and $\bar{y}$ refers to the mean overall score. The sums are taken over all students who completed the assessment. Similar to the discrimination index, the Pearson coefficient can be between −1 and 1. Desirable results are $r \geq 0.2$ [43]. Note that the Pearson coefficient indicates the correlation between student scores on a particular couplet with their score on the entire assessment.

Individual couplet discrimination indices and Pearson coefficients are provided in Appendix.

Further, we calculated the minimum critical Pearson coefficient value [44] for each couplet: this is defined as being two standard deviations above zero, where the standard deviation is given by:

$$\sigma_r = \frac{1}{\sqrt{N-1}}, \tag{3}$$

where N is the sample size. This minimum places a lower bound on the Pearson coefficient, below which the couplet should almost certainly be removed or reworked. Our sample size is $N = 2,\ 596$, and therefore our minimum critical Pearson coefficient is $r_{\mathrm{min}} = 0.039$. All values are above this cutoff, and the average Pearson coefficient is significantly above this, showing an assessment with reasonable discriminatory power.

In addition to calculating both the discrimination index and Pearson coefficient for each couplet, we also show the average of these metrics over all of the couplets for the entire assessment: $\bar{D} = 0.45 \pm 0.03$ and $\bar{r} = 0.40 \pm 0.03$, which show that, on average, the items have good discriminatory power.

Figure 5 shows the discrimination index vs difficulty for each couplet on SPRUCE. Couplets that fall below the horizontal gray line (a discrimination of 0.3), to the left of the left dashed line (difficulty less than 0.25) or to the right of the right dashed line (difficulty greater than 0.90) should to be examined further, as these fall outside the normally accepted bounds for difficulty and/or discrimination. Below we discuss the couplets that are outside these ranges.

Couplet 13 (see numbering in Appendix) has a low difficulty (0.22 ± 0.01), indicating a difficult couplet, as well as a low discrimination ($r = 0.19$). Typically, low difficulty leads to low discrimination since most students do poorly on the couplet. This couplet probes AO H1, *Propagate uncertainties using formulas*. We chose to keep this couplet because it articulates a concept that many instructors described as important during interviews used to develop the AOs—more than half of instructors mentioned this concept [14], and because we aim to achieve a high spread of difficulties. Further, many couplets probe this AO, so one with poor discrimination does not hinder the results.

Couplet 16 has poor discrimination ($r = 0.10$) and a low difficulty (0.52 ± 0.01) for the number of answer options. This couplet also addresses AO H1, *Propagate uncertainties using formulas*. The error propagation occurs in an unusual context for students, leading to more student difficulties. It is important to note that the difficulty of this couplet should be considered as consistent with random guessing: out of the six answer options on this multiple choice item, three were given full credit for this particular couplet, which means 50% is random guessing. Thus, students found this couplet to be fairly difficult (even with a difficulty measure of 0.52), which partially explains its poor discrimination. This item also addresses another AO, and therefore we choose to retain the item on SPRUCE, as well as this couplet in our analysis, due to the importance of the topic it covers, despite student difficulties with this couplet. The difficulty students have with this couplet can also be used to inform instruction, and this information would be lost if the couplet were removed. Additionally, because the couplet is difficult for students, we anticipate low discrimination.

Couplet 8 also shows poor discrimination ($r = 0.12$) though with reasonable difficulty (0.61 ± 0.01). This couplet addresses AO S2, *Identify actions that might improve precision*. This is a multiple-choice question with five answer options, two of which are given full credit, so this difficulty shows that students are not randomly guessing (in which case we would expect a difficulty of about 0.40).

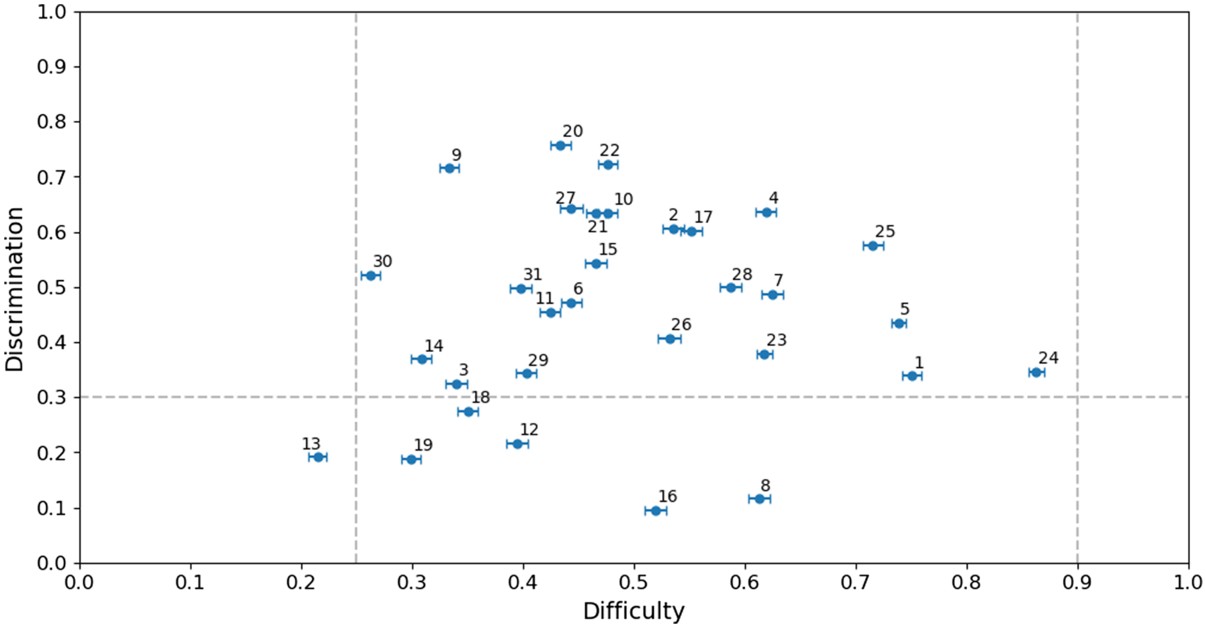


FIG. 5. Discrimination vs difficulty for each item-AO couplet on SPRUCE. Couplet number labels match the corresponding couplet number in Table VIII. Gray dashed lines indicate reference cutoff values.

The AO addressed in this couplet deals with precision. This item also is scored for accuracy (couplet 12). These two concepts can often be difficult for students to distinguish, which we believe may be the cause of the low discrimination for this couplet. We also choose to retain this couplet due to the importance of these results for instructors (i.e., students who perform better overall on the assessment may still struggle with precision and accuracy), as well as for future work, which will focus on student struggles with these concepts.

Couplet 12 shows poor discrimination ($r = 0.22$), though with an acceptable difficulty (0.40 ± 0.01). The item itself is a multiple-choice item, and the couplet investigates AO S3, *Identify actions that might improve accuracy*. This item has five possible answer choices in which two are given full credit, so the couplet difficulty is consistent with random guessing. We choose to keep this couplet because, despite the fact that students are struggling with this concept, instructors care about it: thus, this couplet still provides a measure of student performance that can inform instruction.

Finally, couplet 19 shows less than ideal discrimination ($r = 0.19$) and somewhat low difficulty (0.30 ± 0.01). This couplet investigates AO H2, *Report results with uncertainties and correct significant digits*. This item is a numeric open response item, which, in this context, is scored on student mastery of significant figures. This is a topic students often struggle with and may be uncorrelated to other measurement uncertainty topics, which leads to a lower discrimination. Further, because the couplet is difficult, we anticipate lower discrimination.

Overall, while some of the couplets may fall outside of the range of ideal discrimination and difficulty, this is due to a variety of factors. First, we hope for a spread of difficulties of couplets; because some couplets have a high difficulty, we anticipate these couplets will have a low discrimination. However, the whole-test discrimination and difficulty fall into the ideal range, thus showing that having a few couplets outside of this range is not a significant issue. Further, these couplets can help inform instruction; it is important for instructors to be able to evaluate where students are struggling and, in particular, where students who may excel at most topics are struggling (i.e., the couplets with low discrimination). This information can help instructors make effective changes to their instruction.

3. AO-level discrimination

We present AO-level discrimination indices and Pearson coefficients in Table VI. Discrimination—both in the form of the discrimination index and the Pearson coefficient—validates the way we calculate an overall score, by giving equal weight to each AO score. We find that each AO score has excellent discriminating power.

Further, we previously discussed rounding at the AO level to determine AO scores. We calculated both discrimination indices and Pearson coefficients at the AO level using unrounded scores, scores rounded to the half integer, and scores rounded to the integer. All of these methods resulted in statistics that were not significantly different from one another, which provides additional evidence that rounding at this level is appropriate. We round for ease of future analysis with the data, such as utilizing ordinal logistic regression.

F. Reliability: Stability and internal consistency

Reliability is a method of generalizing the assessment to future administrations. Essentially, it is a way of determining if the current data from students who have taken the assessment are representative of data from potential future students. Engelhardt describes three types of reliability: stability, equivalency, and internal consistency [33]. Stability refers to consistency of scores over time; equivalency refers to relation of scores on two different versions of the assessment to each other; and internal consistency refers to homogeneity of items. For this work, we calculate reliability in the forms of internal consistency and stability. Only one form of SPRUCE exists, so determining equivalency is not possible.

1. Stability

Determining stability traditionally requires students to undergo an additional round of testing to obtain test-retest scores within a short time frame. However, this method has two major issues. First, it creates an extra burden on students and instructors, due to the necessity of another round of testing. Second, this method would not work well as students would recall the assessment and thus skew the results. We instead assume that the populations of students who participated in the administration of SPRUCE in Fall 2022 and Spring 2023 were equivalent and make the same assumption for the Fall 2022 and Fall 2023 populations. This is a reasonable assumption because the types of courses in terms of intended student population of the course (i.e., we had a mix of physics for life science majors and physics for physics and engineering majors across all semesters of administration) and the level of course (i.e., we had a similar set of students' year in school for all three semesters) surveyed in all three semesters are very similar. In addition, in many cases, the same courses participated in SPRUCE in all three semesters. We then use these pairs of data sets as test-retest data. To determine the stability, we examine only the post-test administration in these two terms.

Because the individual students taking SPRUCE during all three semesters of administration were not identical, traditional methods of calculating stability such as detailed in Englehardt [33] do not apply. Instead, we used an alternate approach as described in Day and Bonn [11]. We find a Pearson coefficient between two administrations of the same test in different semesters. We calculate the Pearson coefficient to determine the correlation of average AO scores between the Fall 2022 and Spring 2023 administrations of SPRUCE and Fall 2022 and Fall 2023. For example, we pair the average score of AO S1 in Fall 2022 and AO S1 in Spring 2023, AO S2 in Fall 2022 and AO S2 in Spring 2023, and so on. We find the stability of the test to be 0.985 ± 0.009 with $p \ll 0.01$. The stability is therefore an acceptable value greater than 0.70, the generally accepted cutoff [45]. For the Fall 2022 and Fall 2023 semesters, we find the stability of SPRUCE to be 0.98 ± 0.01 with $p \ll 0.01$. We chose these two sets of data for comparison to have one set of spring versus fall and one set of fall versus fall in order to control for different types of student populations that might be enrolled in the different semesters (for example, some courses have typical times they are offered when most students take the course, while the other term might be the "off" term with a different student population; this alternating semester effect was shown in recent work by Christman *et al.* [46]). In either case, the stability is both significant and high, showing that SPRUCE conforms to test–retest stability when calculated in this manner.

2. Internal consistency

To determine the internal consistency of the entire assessment, we used two methods. First, the assessment was fully scored and then randomly split into two halves, each with five of the ten AOs. The Pearson coefficient was then calculated between the average scores on the two sets of five AOs. Next, the Spearman-Brown prophecy equation was applied as a correction due to each half of the assessment having fewer items than the whole [33]:

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}}. \tag{4}$$

This equation for the internal consistency of the entire exam (r_{tt}) is given in terms of the correlation coefficient between the two halves (r_{hh}). We repeated this process 126 times, once for each possible split of ten numbers, and averaged all of the values obtained to find that $r_{tt} = 0.809 \pm 0.005$, which is above the accepted value of 0.70 for internal consistency [33,41]. Thus, we have strong evidence that SPRUCE, as a whole assessment, contains items that are internally consistent with one another. The items are homogeneous to an extent, measuring the same overarching concept of measurement uncertainty.

A second method of evaluating the internal consistency of an assessment is Cronbach's alpha. The goal of calculating this statistic is to determine whether the AOs are internally consistent with each other, which would indicate that each of the AOs measures one component of the overarching topic of measurement uncertainty. This both aids in validating SPRUCE as an assessment that probes measurement uncertainty, as well as validates our method of calculating an overall score by averaging AO scores. Cronbach's alpha was calculated using the following equation [33]:

$$\alpha = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum_{i=1}^{K} \sigma_i^2}{\sigma_t^2}\right). \tag{5}$$

This statistic takes into account the number of AOs on the assessment ($K = 10$), the total test variance (σ_t^2), and the variance for each AO (σ_i^2). Cronbach's alpha for SPRUCE is $\alpha = 0.83 \pm 0.01$, showing that the AOs are internally consistent with one another, since this is above the generally accepted cutoff of about 0.70 [33,41].

We further calculated the Cronbach's alpha with each AO removed to ensure that no single AO is artificially lowering the Cronbach's alpha for the assessment; all of these values are within error of the Cronbach's alpha for the assessment. Had any of these values been significantly higher than the Cronbach's alpha for the assessment, it would have indicated that the particular AO was lowering the internal consistency of the entire exam and therefore, likely not consistent with the other AOs.

V. SUMMARY AND FUTURE RESEARCH

In the work presented here, we have provided answers to our initial research questions. First, we have shown evidence for validity and reliability for SPRUCE as an assessment tool for the student population included in this study by calculating various CTT metrics, such as difficulty, discrimination, and internal consistency. These statistics were performed at various levels of scores—couplets, AO scores, and overall score—to provide evidence that SPRUCE is both valid and reliable. Second, we have shown methods of adapting CTT for an assessment using couplet scoring, including the use of various statistics with AO scores. Using our new base unit of the item-AO couplet from our scoring scheme, we evaluated couplet difficulty, couplet discrimination index, and couplet Pearson coefficient to perform a couplet-by-couplet analysis of SPRUCE and determined that all of the couplets show evidence of validity. We also calculated these same statistics at the AO-level and again found that the AOs show evidence of validity. Further, we calculated whole-test statistics, such as average couplet difficulty, average couplet discrimination, average couplet Pearson coefficient, Ferguson's delta, split-halves reliability, and test–retest stability in order to show evidence of validity and reliability for SPRUCE as a whole assessment.

Future research includes a full item response theory analysis of SPRUCE, once enough data are collected to make this feasible. In addition, future work is forthcoming regarding student learning gains compared between pre- and postinstruction administration of SPRUCE in laboratory courses, including a breakdown of several assessment objectives and a close examination of areas in which students most often struggle.

ACKNOWLEDGMENTS

This work is supported by NSF (Grants No. DUE 1914840, No. DUE 1913698, and No. PHY 2317149). We would also like to thank Rachel Henderson for her insights into how couplet scoring might be used in the context of statistical validation.

APPENDIX: INDIVIDUAL COUPLET STATISTICS

Here, we present statistics for all individual couplets on SPRUCE in Table VIII.

TABLE VII. Summary of statistical test results for whole assessment, SPRUCE ($N = 2596$). We present the results of all statistical tests run at the whole-test level, including difficulty of the assessment, the average couplet difficulty, the average couplet Pearson coefficient, Ferguson's delta, Cronbach's alpha, split-halves reliability, and test–retest stability. Because test–retest stability is calculated twice, we show both values in this table—once for a comparison of data from Fall 2022 and Fall 2023 semesters, and once for a comparison of data from Fall 2022 and Spring 2023 semesters.

Statistic	Range	Desired values	SPRUCE value
Difficulty, overall score	[0, 1]	[0.25, 0.90]	0.509 ± 0.004
Average couplet difficulty	[0, 1]	[0.25, 0.90]	0.49 ± 0.03
Average couplet discrimination index, $\bar{D}$	$[-1, 1]$	≥ 0.30	0.45 ± 0.03
Average couplet Pearson coefficient, $\bar{r}$	$[-1, 1]$	≥ 0.20	0.40 ± 0.03
Ferguson's delta, δ	[0, 1]	≥ 0.90	0.947
Cronbach's alpha, α	[0, 1]	≥ 0.70	0.83 ± 0.01
Split-halves reliability	[0, 1]	≥ 0.70	0.809 ± 0.005
Test–retest stability (Fall22/Fall23)	$[-1, 1]$	≥ 0.70	0.98 ± 0.01
Test–retest stability (Fall22/Spring23)	$[-1, 1]$	≥ 0.70	0.985 ± 0.009

TABLE VIII. Summary of statistical test results for each SPRUCE couplet [$N = 2596$]. See Table I for full text of each assessment objective. This table includes the difficulty, discrimination, and Pearson coefficient for each SPRUCE couplet.

Couplet	Assessment objective (AO)	Difficulty, ±0.01	Discrimination index, D	Pearson coefficient, r, ±0.02
1	S1	0.75	0.34	0.31
2	S1	0.54	0.61	0.48
3	S1	0.34	0.32	0.27
4	S2	0.62	0.64	0.54
5	S2	0.74	0.43	0.52
6	S2	0.44	0.47	0.41
7	S2	0.62	0.49	0.39
8	S2	0.61	0.12	0.10
9	S3	0.33	0.72	0.66
10	S3	0.48	0.63	0.58
11	S3	0.42	0.46	0.39
12	S3	0.40	0.22	0.19
13	H1	0.21	0.19	0.20
14	H1	0.31	0.37	0.32
15	H1	0.47	0.54	0.43
16	H1	0.52	0.10	0.10
17	H2	0.55	0.60	0.48
18	H2	0.35	0.27	0.24
19	H2	0.30	0.19	0.17
22	D1	0.43	0.76	0.66
23	D1	0.47	0.63	0.59
24	D2	0.48	0.72	0.63
25	D2	0.62	0.38	0.43
26	D3	0.86	0.35	0.44
27	D3	0.72	0.58	0.50
28	D4	0.53	0.41	0.33
29	D4	0.44	0.64	0.52
30	D4	0.59	0.50	0.40
31	D4	0.40	0.34	0.27
32	D5	0.26	0.52	0.47
33	D5	0.40	0.50	0.42

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[2] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, Am. J. Phys. **66**, 338 (1998).

[3] K. D. Rainey, M. Vignal, and B. Wilcox, Validation of a coupled, multiple response assessment for upper-division thermal physics, Phys. Rev. Phys. Educ. Res. **18**, 020116 (2022).

[4] R. Chabay and B. Sherwood, Restructuring the introductory electricity and magnetism course, Am. J. Phys. **74**, 329 (2006).

[5] B. R. Wilcox and S. J. Pollock, Validation and analysis of the coupled multiple response Colorado upper-division electrostatics diagnostic, Phys. Rev. ST Phys. Educ. Res. **11**, 020130 (2015).

[6] E. Marshman and C. Singh, Validation and administration of a conceptual survey on the formalism and postulates of quantum mechanics, Phys. Rev. Phys. Educ. Res. **15**, 020128 (2019).

[7] G. Zhu and C. Singh, Surveying students' understanding of quantum mechanics in one spatial dimension, Am. J. Phys. **80**, 252 (2012).

[8] 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.

[9] B. Campbell, F. Lubben, A. Buffler, and S. Allie, Teaching scientific measurement at university: Understanding students' ideas and laboratory curriculum reform, Mono.

African J. Res. Math. Sci. Math. Educ. (2005), https://pure.york.ac.uk/portal/en/publications/teaching-scientific-measurement-at-university-understanding-stude.

[10] 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).

[11] J. Day and D. Bonn, Development of the Concise Data Processing Assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).

[12] B. Pollard, M. F. J. Fox, L. Ríos, and H. J. Lewandowski, Creating a coupled multiple response assessment for modeling in lab courses, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.Pollard.

[13] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).

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

[17] J. Leach, R. Millar, J. Ryder, M.-G. Séré, D. Hammelev, H. Niedderer, V. Tselfes, M. Bandiera, F. Dupré, C. Tarsitani, and E. Torracca, Survey 2: Students' images of science as they relate to labwork learning (1998).

[18] S. M. Coelho and M. Séré, Pupils' reasoning and practice during hands on activities in the measurement phase, Res. Sci. Technol. Educ. **16**, 79 (1998).

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

[20] G. J. Aubrecht and J. D. Aubrecht, Constructing objective tests, Am. J. Phys. **51**, 613 (1983).

[21] 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.

[22] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[23] 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.

[24] K. N. Quinn, C. E. Wieman, and N. G. Holmes, Interview validation of the physics lab inventory of critical thinking (PLIC), presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.076.

[25] 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).

[26] W. K. Adams, The design and validation of the Colorado Learning Attitudes about Science Survey, AIP Conf. Proc. **790**, 45 (2005).

[27] S. J. Pollock, Transferring transformations: Learning gains, student attitudes, and the impacts of multiple instructors in large lecture courses, AIP Conf. Proc. **818**, 141 (2006).

[28] R. J. Mislevy and M. M. Riconscente, Evidence-Centered Assessment Design: Layers, Structures, and Terminology, Tech. Rep. (SRI International Center for Technology in Learning, 2005).

[29] J. T. Laverty and M. D. Caballero, Analysis of the most common concept inventories in physics: What are we assessing?, Phys. Rev. Phys. Educ. Res. **14**, 010123 (2018).

[30] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).

[31] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020124 (2014).

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

[33] P. V. Engelhardt, An introduction to classical test theory as applied to conceptual multiple-choice tests (2009).

[34] D. Rindskopf, Reliability: Measurement, in *International Encyclopedia of the Social & Behavioral Sciences*, edited by N. J. Smelser and P. B. Baltes (Pergamon, Oxford, 2001), pp. 13023–13028.

[35] M. Vignal, G. Geschwind, M. D. Caballero, and H. J. Lewandowski, Couplet scoring for research based assessment instruments, arXiv:2307.03099.

[36] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional Item Response Theory and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010137 (2018).

[37] 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.

[38] 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).

[39] L. S. Nelson, The Anderson-Darling test for normality, J. Qual. Technol. **30**, 298 (1998).

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

[41] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction* (National Science Teachers Association, Washington, DC, 1980), pp. 97–104, eRIC Number: ED196733.

[42] P. Kline, *Handbook of Psychological Testing* (Routledge, New York, NY, 2000), p. 31, https://www.routledge.com/Handbook-of-Psychological-Testing/Kline/p/book/9780415211581.

[43] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Methuen, New York, NY, 1986), p. 143.

[44] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, Orlando, FL, 1986), p. 34, eRIC Number: ED312281.

[45] P. Kline, *The New Psychometrics: Science, Psychology and Measurement*, 1st ed. (Routledge, New York, NY, 1998), p. 29.

[46] E. Christman, P. Miller, and J. Stewart, Beyond normalized gain: Improved comparison of physics educational outcomes, Phys. Rev. Phys. Educ. Res. **20**, 010123 (2024).