

## Survey of physics reasoning on uncertainty concepts in experiments: An assessment of measurement uncertainty for introductory physics labs

Michael Vignal<sup>1,2</sup>, Gayle Geschwind<sup>1,2</sup>, Benjamin Pollard<sup>3</sup>, Rachel Henderson,<sup>4</sup>  
Marcos D. Caballero,<sup>4,5,6</sup> and H. J. Lewandowski<sup>1,2</sup>

<sup>1</sup>JILA, National Institute of Standards and Technology and the University of Colorado,  
Boulder, Colorado 80309, USA


<sup>2</sup>Department of Physics, University of Colorado, 390 UCB, Boulder, Colorado 80309, USA

<sup>3</sup>Department of Physics, Worcester Polytechnic Institute,  
100 Institute Road, Worcester, Massachusetts 01609, USA

<sup>4</sup>Department of Physics & Astronomy and CREATE for STEM Institute,  
Michigan State University, East Lansing, Michigan 48824, USA

<sup>5</sup>Department of Computational Mathematics, Science, & Engineering,  
Michigan State University, East Lansing, Michigan 48824, USA

<sup>6</sup>Department of Physics and Center for Computing in Science Education,  
University of Oslo, 0315 Oslo, Norway

 (Received 14 February 2023; accepted 22 August 2023; published 4 October 2023)

[This paper is part of the Focused Collection on Instructional labs: Improving traditions and new directions.] Measurement uncertainty is a critical feature of experimental research in the physical sciences, and the concepts and practices surrounding measurement uncertainty are important components of physics lab courses. However, there has not been a broadly applicable, research-based assessment tool that allows physics instructors to easily measure students' knowledge of measurement uncertainty concepts and practices. To address this need, we employed evidence-centered design to create the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE). SPRUCE is a pre-post assessment instrument intended for use in introductory (first and second year) physics lab courses to help instructors and researchers identify student strengths and challenges with measurement uncertainty. In this paper, we discuss the development of SPRUCE's assessment items guided by evidence-centered design, focusing on how instructors' and researchers' assessment priorities were incorporated into the assessment items and how students' reasoning from pilot testing informed decisions around item answer options. We also present an example of some of the feedback an instructor would receive after implementing SPRUCE in a pre-post fashion, along with a brief discussion of how that feedback could be interpreted and acted upon.

DOI: [10.1103/PhysRevPhysEducRes.19.020139](https://doi.org/10.1103/PhysRevPhysEducRes.19.020139)

### I. INTRODUCTION

Measurement is a central component of experimental scientific research, as all experimental measurements have some uncertainty. Proper consideration and handling of measurement uncertainty (MU) is critical for appropriately interpreting measurements and making claims based on experimental data. While some techniques for determining and using MU can be quite sophisticated, it is still possible (and desirable [1]) to teach basic MU techniques in introductory science labs. In experimental physics, MU informs comparisons of multiple measurements [2]

or between measurements and values predicted by models [3], and so instruction around MU can help students better understand the nature of experimentation. This and other important features of MU have led to policies and recommendations to include MU in introductory science courses [4,5]. As developing proficiency with MU practices becomes an even more important goal in undergraduate physics labs, it is critical to be able to assess the level to which students are reaching this goal.

Educators often wish to evaluate student learning around important concepts and practices—often articulated as learning goals or learning objectives [6]—in order to inform and improve their instruction. To help instructors determine if learning goals are being achieved, the physics education research community has often developed and employed research-based assessments instruments (RBAs), which Madsen, McKagan, and Sayre define as “an assessment that is developed based on research into

---

Published by the American Physical Society under the terms of the *Creative Commons Attribution 4.0 International* license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

student thinking for use by the wider...education community to provide a standardized assessment of teaching and learning” [7]. It is important to note that RBAs provide researchers and instructors with opportunities to assess student learning across time, institutions, and curricular and pedagogical changes in order to inform and improve instruction; they are not intended to evaluate individual students for the purpose of assigning grades.

Developers of RBAs often employ a theoretical framework during assessment development, such as evidence centered design (ECD) [8], the three-dimensional learning assessment protocol [9], or the framework described by Adams and Wieman [10]. Such frameworks “facilitate communication, coherence, and efficiency in assessment design and task creation” [8], typically by outlining steps or stages of assessment development, including exploratory research, data collection, and item development through to assessment delivery, scoring, and validation. ECD, the framework used in this work, also provides a structure for establishing evidence-supported claims about student reasoning based on student responses to the assessment: these claims are grounded in evidentiary arguments (a major focus of this paper) and contribute to the validity of the assessment instrument.

Of the RBAs employed in physics labs, several focus on measurement and MU, albeit to varying extents. The Physics Measurement Questionnaire (PMQ) has been fundamental in articulating the pointlike and setlike reasoning paradigms [11] and in measuring the success of course transformations aimed at helping students shift toward more setlike reasoning [2,12–14]; the Physics Lab Inventory of Critical Thinking (PLIC) [15] has been used to assess the effectiveness of a scaffold and fade approach to teaching critical thinking in a physics course [16]; and the Concise Data Processing Assessment (CDPA) [17] has been used to identify changes in student performance around MU [18] and to look at student performance across genders [19].

While each of these assessments deals with MU in some way, there is not currently a widely administrable RBA that focuses explicitly on MU in introductory (first and second year) physics laboratory courses. To address this gap in assessments, we have developed the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) using the assessment development framework of ECD [8]. It is our hope that SPRUCE will help instructors and researchers identify and improve instruction around measurement uncertainty concepts and practices that are challenging for introductory physics lab students.

In this paper, we present SPRUCE’s assessment questions (hereafter referred to as “assessment items” or simply “items”) and discuss their development. The goals of this paper are to demonstrate (i) a need for a widely administrable assessment of measurement uncertainty and how SPRUCE will satisfy that need; (ii) the assessment item

development and refinement process, as guided by ECD; (iii) examples of evidentiary arguments, formed from student reasoning, that support our ability to make claims about student knowledge based on student responses to the assessment items; and (iv) an example of feedback for instructors and how that feedback might be interpreted.

We begin by discussing, in Sec. II, the need for a new MU RBA and how the framework of ECD can facilitate the development of such an assessment. In Sec. III, we describe the first three layers of ECD (*domain analysis*, *domain model*, and *conceptual assessment framework*) and how we gathered information and made decisions to support the development of assessment items and evidentiary arguments. The development and refinement of these items and arguments are discussed in Sec. IV. In Secs. V and VI, we briefly discuss components of validity and how instructors might interpret and use the results of the instrument. In the final section, Sec. VII, we summarize the work discussed in this paper and provide information for instructors and researchers who may be interested in using SPRUCE.

Future papers will discuss details of SPRUCE scoring, statistical validation, and claims about student learning, which all require a discussion of SPRUCE’s scoring scheme using a new scoring paradigm. As discussion of this paradigm is beyond this paper’s scope of introducing SPRUCE, in this paper, these topics are discussed only briefly to highlight how they informed item development.

## II. BACKGROUND

Over the last 30 years, research-based assessment instruments (RBAs) have been used in physics classrooms to probe areas of interest and import for physics education researchers and physics instructors. Particularly, notable examples of RBA use include Mazur’s use of assessment questions from the Force Concept Inventory [20] to probe student understanding of Newton’s third law in his introductory physics lecture at Harvard [21], which led him (and others) to rethink what instruction in a lecture setting should look like [22] and Eblen-Zayas’ use of the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS [23]) in her advanced lab courses, where she found that introducing metacognitive activities in an open-ended lab course had a positive impact on student enthusiasm and confidence [24].

More broadly, RBAs have been developed and deployed in the areas of mechanics [20,25], electrostatics [26,27], quantum mechanics [28], and thermodynamics [29]. In addition, assessments have been used to probe quantitative reasoning [30], beliefs about physics and physics courses [31], experimental research and lab courses [23], modeling [32,33], and concepts and practices used in laboratory courses [11,15,17]. These and other assessments can be found on the PhysPort website [34], and many of them are also accessible on LASSO [35].

In this section, we provide more detail about RBAs that probe student proficiency in working with measurement uncertainty (MU). We highlight the strengths of these existing assessments, while also (as stated in our first research goal) arguing that there is still a need for a new assessment specifically probing MU in introductory physics labs. We then present initial work on the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE), which is our response to the need for a new MU assessment instrument, and discuss how the framework of evidence-centered design (ECD) [36] informed this work.

### A. Research-based assessment instruments in physics labs

The following sections discuss three existing RBAs that include some assessment of MU topics. While each of these RBAs has contributed to our collective understanding of student reasoning around measurement uncertainty, they each have limitations that point to a need for a widely administrable assessment of measurement uncertainty for introductory physics labs.

#### 1. Physics Measurement Questionnaire

The Physics Measurement Questionnaire (PMQ) consists of multiple-choice and open-response items adapted from the Procedural and Conceptual Knowledge in Science (PACKS) Project [37] for use with students at the University of Cape Town, South Africa [11]. These items present decisions that students might face in a lab course and ask students which option they agree with (in a multiple-choice format) and then ask them to explain their reasoning (in an open-response format). Validation of the PMQ consisted of student interviews to “check students’ understanding of the questions and the interviewer’s interpretation of their responses” and to “confirm that the probes presented sufficient alternatives covering a wide enough range of possibilities” [11].

One of the most important findings to come out of the PMQ was the articulation of the point and set paradigms for student reasoning. These paradigms classify many types of student reasoning as being either pointlike, indicating students believe that quantities measured have a true value that can be obtained with a single, perfect measurement; setlike, a typically more expertlike view that measurements will always have uncertainty and that a true value (if it exists) can never be perfectly known; or something else, usually with elements of both pointlike and setlike perspectives.

Despite the successes of the PMQ in articulating this paradigm and helping to inform course transformations, the assessment has two large limitations: the PMQ covers only a narrow range of ideas related to MU (primarily around distributions of results from repeated measurements), and the assessment is open response and therefore laborious to score. This second limitation is compounded

by variance in student responses observed at different institutions, sometimes requiring instructors and researchers to first modify the scoring scheme provided by the developers of the PMQ [2].

#### 2. Physics Lab Inventory of Critical Thinking

The Physics Lab Inventory of Critical Thinking was developed by physics education researchers at Cornell University and Stanford University to “assess how students critically evaluate experimental methods, data, and models” [34]. The developers of the PLIC conducted multiple rounds of interviews and full-course piloting with several hundred students, as well as distributed the instrument to experts, to establish various forms of validity of the instrument including construct and concurrent validity [15]. The PLIC is contextualized in a small number of experiments, about which students are asked multiple questions, and the assessment is administered in an online format.

The PLIC has been used to evaluate a “scaffold and fade approach” to instruction around making comparisons between measurements, or between measurements and models, for students in an introductory physics lab course [16]. This approach involves a structured, explicit focus on a concept or practice initially (the “scaffold”), which then “fades” over the course of instruction as student proficiency develops. Students who received this scaffold and fade instruction around making comparisons were much more likely to think critically about their results and propose possible improvements to their experimental setup than were students who had taken the course the previous year and not received this instruction [16].

The PLIC was explicitly designed to assess critical thinking, which the authors define as “the ways in which one uses data and evidence to make decisions about what to trust and what to do.” The authors aim to assess critical thinking in a lab setting, and while this includes components of MU, MU is not the primary focus of the assessment [15].

#### 3. Concise Data Processing Assessment

The Concise Data Processing Assessment (CDPA) was developed by researchers at the University of British Columbia (UBC) to assess student proficiency around MU (primarily related to error propagation) and data handling [17]. It consists of multiple-choice questions and can be presented in a pre-post format so as to probe student learning in a course. The CDPA was developed to complement the learning goals of a “rigorous” introductory physics lab, and the researchers used full-class piloting and student interviews to refine the assessment items. The CDPA developers established validity with data from 12 faculty and 11 graduate students who completed the assessment.

The CDPA has been employed to explore if improvements in student proficiency with MU had an impact on



their scores on E-CLASS [18]. While there were not enough matched preinstruction and postinstruction data to make comparisons of improvement on these two assessments, no correlation was found between CDPA scores and E-CLASS preinstruction scores. However, the CDPA was found to be able to measure shifts in student proficiency, specifically positive shifts around content that was emphasized in the courses and negative shifts in content that was not emphasized in instruction. This study was conducted with participants in their second- or third-year laboratory course at the University of Helsinki.

As stated above, the CDPA was developed to complement an intensive introductory physics lab, but even still, it is a challenging assessment: as part of assessment development, graduate students at UBC were administered the assessment and scored, on average, just over 50%, with post-test scores for first-year students averaging less than 40%. In the second study discussed above, second- and third-year physics majors showed no improvement in CDPA scores from the preassessment to postassessment (with an overall score of around 40%). As such, the CDPA may not be appropriate for many introductory physics labs, as its difficulty may limit its ability to identify trends and provide usable feedback for instructors.

### B. Assessment development framework: Evidence centered design

To help guide the development of SPRUCE, we employed the assessment development framework of evidence centered design (ECD) [8] to help us incorporate instructor priorities around MU into the assessment instrument and to support the gathering of evidence of student reasoning that informs our interpretation and evaluation of student responses to the assessment items. Throughout this paper, we refer to these explanations that link student reasoning to student item responses as *evidentiary arguments*.

ECD consists of five layers to facilitate “communication, coherence, and efficiency in assessment design and task creation” [8]. We list and briefly summarize these layers below:

- *Domain analysis*: gather information on the topic to be assessed, including from current instructors.
- *Domain model*: organize *domain analysis* data by writing narrative assessment arguments that describe proficiencies to be measured (which we do via assessment objectives [29,38]), acceptable evidence of such proficiencies, and the methods for gathering this evidence.
- *Conceptual assessment framework*: operationalize assessment arguments to determine appropriate assessment features and item formats.
- *Assessment implementation*: write then iteratively pilot and revise assessment items while establishing evidentiary arguments that link observable data (student responses) to targeted claims about student

reasoning, which will eventually be quantified via a scoring scheme.

- *Assessment delivery*: finalize implementation of assessment, scoring scheme, and instructor reports.

The first layer of ECD, domain analysis, is the topic of a previous paper [1] and briefly summarized below. Domain model, conceptual assessment framework, and especially assessment implementation constitute the bulk of the work presented here, with a strong emphasis on piloting and evidentiary arguments. Our development of the quantitative scoring scheme used with SPRUCE (part of assessment implementation) and the fifth layer (assessment delivery) are briefly discussed in this paper and are instead the focus of upcoming papers.

## III. SPRUCE DEVELOPMENT

### A. Domain analysis

The first steps toward developing an RBAI on MU were presented in a previous paper [1]. In that work, we conducted and analyzed interviews with 22 physics lab instructors at institutions that spanned a range of sizes, the highest degrees offered, selectivity, and student body demographics. In these interviews, we sought to identify instructor priorities when it came to the teaching and learning of MU. These interviews were semistructured in nature and typically lasted around one hour.

Preliminary coding of these interviews was done to identify which concepts and practices instructors described as priorities or aspirational priorities for their courses. Instructors also talked about challenges for students and for instruction, including dealing with ideas taught in high school that students need to unlearn or refine, which informed our decisions of what content to include (or not include) in SPRUCE.

### B. Domain modeling

After the domain analysis, domain modeling involves “articulat[ing] the argument[s] that [connect] observations of students’ actions in various situations to inferences about what they know or can do” [8]. These assessment arguments are narrative in structure and describe the concepts and practices (i.e., the constructs) to be assessed, how evidence of student proficiency with respect to those concepts and tasks might be gathered, and how the items will allow students to demonstrate such proficiencies. It is in this stage that specific instrument items begin to take shape, as ideas gathered in the *domain analysis* are reexpressed in terms of specific tasks.

To more explicitly embody the assessment priorities of instructors, we expressed our assessment arguments in terms of assessment objectives [29]. Assessment objectives (AOs) are “concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment” [38]: essentially, the

TABLE I. Final 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	Identify when to use fractional versus absolute uncertainty.
H2	Propagate uncertainties using formulas.
H3	Report results with uncertainties and correct significant digits.
H4	Use concepts of uncertainty propagation to identify the largest contributor to uncertainty in a calculated value.
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	Articulate that the standard deviation is related to the width of the distribution of measurements.
D5	Report the standard error (standard deviation of the mean) for the uncertainty in the mean of a distribution.
D6	Calculate the standard error from the standard deviation.
D7	Determine if two measurements (with uncertainty) agree with each other.

AOs are the instrument's constructs. AOs are similar in concept and grain size to learning objectives [6], but they are designed to "span the space of feasible, testable outcomes" of an assessment [29]. As discussed in Ref. [38], AOs also provide a number of additional benefits for assessment development beyond the organization of ideas collected in the domain analysis.<sup>1</sup>

For SPRUCE, our AOs emerged from the qualitative codes developed during the domain analysis and from the list of concepts and practices noted as being important to experts that were developed in Ref. [1] and from a survey of instructor priorities around MU. These AOs were iteratively refined during item development, piloting, and development of our evaluation scheme.

Ultimately, we identified four main areas of concepts and practices into which all of our AOs can be organized, and these categories and their AOs resemble the dimensions and concepts developed to model MU content in secondary science education [39]:

- Sources of uncertainty: estimating the size of uncertainty and identifying ways to reduce it.
- Handling of uncertainty: uncertainty propagation and significant digits.
- Distributions and repeated measurements: mean, standard deviation, standard error, and the importance of taking multiple measurements.
- Modeling: comparisons between explicit externalized models and the data.

Because the modeling category pertained primarily to explicit comparisons between externalized models and data, we determined that AOs in this category fell outside of the scope of this assessment instrument. However, there are still elements of modeling, as defined by the Experimental Modeling Framework [40], that remain as integral parts of the other categories. Additionally, as described in Ref. [38] and discussed further in Sec. IV, some individual AOs in the other categories were also removed because of difficulties in establishing clear evidence of student reasoning. The finalized AOs are presented in Table I.<sup>2</sup>

In practice, rather than a strictly narrative structure, our assessment arguments included a narrative description of the task that would be presented to students, the AOs the item would assess and which responses would constitute evidence of proficiency, and a paragraph describing the rationale for why the item is appropriate. In a sense, these assessment arguments represent a hypothesis regarding a claim that the assessment will be able to make: if we present task X to students and they provide response Y, then we can conclude Z about their knowledge and reasoning around a particular AO. The connection between student responses and student reasoning comes from evidentiary arguments, which are developed during assessment implementation and described in Sec. IV A.

While the literature on ECD portrays a fairly linear progression from one layer to the next, we took a more iterative approach in which we revisited and revised our work in previous layers (including domain modeling) as we worked on subsequent layers.

### C. Conceptual assessment framework

The third layer of the ECD framework involves operationalizing the assessment arguments developed in the second layer to inform the development of assessment items. This process includes deciding on the format of the assessment and the individual items and selecting a scoring paradigm.

<sup>1</sup>In Ref. [38], we described the articulation of assessment arguments as being part of the conceptual assessment framework: we now believe that it more appropriately belongs in the domain model.

<sup>2</sup>Using the AOs outlined in Table I, we can describe the PMQ as covering S1 and most of distributions and repeated measurements (with the exception of D5 and D6), the PLIC as covering sources of uncertainty and distributions and repeated measurements (again with the exceptions of D5 and D6), and the CDPA as focusing primarily on handling of uncertainty (as well as graphical representations of data).

TABLE II. SPRUCE experiment descriptions.

Experiment	Description
Cart acceleration (experiment 1)	A cart is released from rest to roll down a ramp as part of an experiment to determine the acceleration of the cart. Students are asked about taking multiple measurements and to identify the source of greatest uncertainty in their calculation of the acceleration.
Mug density (experiment 2)	The density of a mug is to be computed by measuring its mass and volume. Students are asked to identify uncertainties in each measurement and then propagate those uncertainties.
Spring constant (experiment 3)	A mass hangs from a spring and the period of (vertical) oscillation is used to determine a spring constant. Students are asked to estimate and propagate uncertainties and make comparisons between results.
Breaking mass (experiment 4)	Successive masses are added to a mass hanger until the string holding the mass hanger breaks. Students are asked to estimate the uncertainty of a single measurement, make comparisons between results, and answer questions about taking many more measurements.

In order to ensure a compact survey and reduce the cognitive load on students, we contextualized all of SPRUCE’s assessment items within four experiments (as opposed to each item being a unique experimental context). Initial experimental contexts aligned with contexts discussed by instructors in the *domain analysis* and were refined as needed to support the establishment of evidentiary arguments. The four experiments are summarized in Table II and described in more detail in Sec. IV C.

To develop an assessment that is easy to administer to a large number of students—twice, as SPRUCE is intended to be used preinstruction and postinstruction—we opted for an online format for the assessment [41,42] using the survey platform Qualtrics. We embedded digital calculators in all items in which students select or enter a numeric response, and we selected six potential item formats that facilitate automated evaluation.

The first three item formats are multiple choice (MC), multiple response (MR), and numeric open response (NOR). These formats contain a single prompt (or “stem” [43]) to which students respond by selecting a single answer (for MC items) or multiple answers (for MR items) from a list of answer options or by entering a number into a text box (for NOR items). MC and MR items are the most common types of items on assessments, as they are straightforward to develop and evaluate. While NOR items were more complicated to evaluate, Qualtrics is able to exclude non-numeric

responses from text boxes, meaning that student responses were sufficiently constrained that these items could be evaluated using a simple algorithm.

The next three item formats involve the coupling of two questions: coupled multiple choice (CMC) items that have two coupled MC parts, coupled multiple response (CMR) that have an MC part followed by an MR part, and coupled numeric open response (CNOR) that have two NOR parts. These item types are examples of two-tier questions [44] in which, rather than considering the answer options selected in either question independently, it is the combination of selections from the coupled questions that are evaluated.

For SPRUCE’s CMR items, the multiple response answer options are *reasoning elements* that allow students to compose a justification for their response to the multiple choice question, a design used in other physics assessments [27,45,46] that allows for evaluating complex student reasoning in a format that can be evaluated by a computer (as opposed to, for example, a free response justification that must be evaluated by a person or well-trained machine learning algorithm [47]). We used CNOR items to compare student values and uncertainties to see if students reported these quantities using appropriate significant digits, though as these items ask students to respond in a text box, student browsers may store student responses from the preinstruction assessment and suggest or autofill them during the postinstruction assessment, and so quantities that factor into student responses on NOR and CNOR items are slightly different between preinstruction and postinstruction versions of the assessment.

#### D. A brief note on scoring

The selection of a scoring paradigm also impacts what types of items one might use in an assessment instrument. From the early stages of SPRUCE’s development, we decided to have items relate to potentially more than one AO and to score each item once for each of the item’s AOs. This approach resulted in the development of *couplet scoring* in which a couplet is essentially an item viewed and scored through the lens of a single AO. As discussed in a paper under review as of this publication [48], the couplet becomes the unit of assessment for scoring, validation, and reporting student proficiencies, and it offers a number of affordances in these and other aspects of assessment development. For example, we found that couplet scoring scaffolded item development and refinement and helped us craft the questions that we wanted within the constraints of MC, MR, and NOR items.

A simple example of couplet scoring for item 3.3 (Fig. 1) is presented in Table III. In this item, students are asked what value they would report for the period of oscillation (with uncertainty) for a mass attached to a spring that is oscillating up and down. Students must select an answer based on the information given in the prompt about a measurement of the time it takes the mass to complete



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 20 oscillations. You estimate the uncertainty in your measurement of 20 oscillations to be 0.4 s, based on an online search for human reaction time. What value and uncertainty do you report for the period of a **single oscillation**?

$1.421 \pm 0.02$  s      $1.42 \pm 0.02$  s      $1.4 \pm 0.02$  s  
  $1.421 \pm 0.4$  s      $1.42 \pm 0.4$  s      $1.4 \pm 0.4$  s

FIG. 1. Item 3.3 (with modified numbers) asks students a single MC question about reporting the value of a single period of oscillation of a mass on a spring based on a measurement of 20 oscillations. The scoring of responses to this item is depicted in Table III.

20 oscillations. Student responses are evaluated twice, once for “H2—Propagate uncertainties using formulas,” and once for “H3—Report results with uncertainties and correct significant digits,” as depicted in III. The independent scores from these couplets are not consolidated into a single item score (couplet scoring does not have “item scores”), rather they each contribute (with all other couplets targeting the same AO) to independent AO scores, as discussed in Sec. VI.

IV. ASSESSMENT IMPLEMENTATION

In the fourth layer of ECD, assessment implementation, assessment items are written and, iteratively, pilot tested, and refined as the developers construct the evidentiary arguments that facilitate meaningful interpretations of student responses. Items were constructed by expressing the assessment arguments developed in the *domain analysis* in terms of the item formats identified in the *conceptual assessment framework*.

A. Evidentiary arguments

As stated above, the key focus of this paper, and a key component of ECD, is the establishment of evidentiary

TABLE III. Example scoring scheme for couplets of item 3.3. Student responses are scored once for each of the item’s AOs “H2—Propagate uncertainties using formulas,” and “H3—Report results with uncertainties and correct significant digits”.


Answer option		Score	
		H2	H3
A	$1.421 \pm 0.02$ s	1	0
B	$1.421 \pm 0.4$ s	0	0
C	$1.42 \pm 0.02$ s	1	1
D	$1.42 \pm 0.4$ s	0	0
E	$1.4 \pm 0.02$ s	1	0
F	$1.4 \pm 0.4$ s	0	1

arguments, which allow researchers to map student reasoning to student responses. The primary source of evidence for evidentiary arguments in this work is student responses to the assessment items during pilot testing, though previous work with the PMQ, [2] and researcher expertise and experience also informed these arguments.

Data from pilot testing (discussed in the next section) were used to establish our evidentiary arguments, linking student reasoning to student responses for each answer option, for each item, for each of the item’s AOs. Interviews, especially, were used to probe student reasoning around not only students’ final responses but also their “second best” responses and other responses they considered.

In an ideal situation, researchers would be able to make a one-to-one mapping between specific student responses and specific lines of student reasoning to ensure that evaluation is based on a perfectly accurate interpretation of student responses. In reality, no amount of piloting will capture all possible responses and reasoning employed by students, and so the goal is to develop evidentiary arguments to map trends in observed responses to trends in expressed reasoning. As a result, most of our item revisions were to improve our mappings by addressing instances in

Your physics lab instructor found an old ball of string and wants to know how strong the string is. They cut it up and give each lab group 10 pieces of string, a 100 g mass hanger, and a large number of 20 g masses, as shown below:



Your lab instructor asks the class to find the “breaking mass,”  $m_{\text{breaking}}$ , for the string. They describe  $m_{\text{breaking}}$  by saying: “The string can support  $m_{\text{breaking}}$ , but the string will break if you add even a grain of sand more than  $m_{\text{breaking}}$ .”

4.1 Your first string is able to support 520 g, but breaks when you try to hang 540 g from it. What value do you report for your best estimate of  $m_{\text{breaking}}$ ?

520 g     521 g     530 g     539 g     540 g

What uncertainty do you report for your best estimate of  $m_{\text{breaking}}$ ?

0 g     1 g     5 g     10 g     19 g     20 g

FIG. 2. Item 4.1 went through iterations informed by multiple rounds of pilot testing with students.

TABLE IV. Example Evidentiary Arguments for item 4.1. The top and bottom halves of the table include the evidentiary arguments for the MC questions asking students, respectively, what mass and uncertainty they would report.

Answer option	Evidence-supported reasoning	Example of evidence
520 g	Maximum confirmed supported mass	“520 is the last value reported that this string is able to support before it breaks.... So that’s the closest value [to the breaking value] that we get before it breaks”
521 g	“Just over” maximum confirmed supported mass	“I guess it would be 521, since that wouldn’t be too far [off from 520].”
530 g	Midpoint of 520 g and 540 g (often justified in conjunction with an uncertainty of 10 g)	“We do know it’s within the range of 520 to 540, and so what this does, if we have it at 530 with an uncertainty of 10, means our minimum value is just over 520, and maximum value is just under 540.”
539 g	“Just under” minimum confirmed unsupported mass	“Maybe it’s 539, because it breaks when you hit 540- maybe that was just slightly too big.”
540 g	Minimum confirmed unsupported mass	“That’s the value that the string broke on”
0 g	There is no uncertainty	Common beta response (not seen in interviews)
1 g	Small but nonzero uncertainty	“It’s better to include some uncertainty than to just make assumptions. So it wouldn’t be zero, but it shouldn’t be too far off.”
5 g	Half of measurement increment’s “place”	“Like I said earlier, if it gives me one decimal place, my uncertainty would be the next one, like 05, so it can go up or down.”
10 g	Half of measurement increment	“I picked 10 because the smallest increments that we can go in this measurement tool is 20, so I took the 20, divided by 2, and got 10”
19 g	Noninclusively spans range (e.g., 521 g to 540 g)	“I would say 19 g...since it wouldn’t include 520 but it could be anywhere else in that range [of 520] to 540.”
20 g	Measurement increment	“So I said 20 because we don’t know what the- say like 521, 535, or 539, if that would also break. So there’s uncertainty there, which I found because 540–520 is equal to 20.”

which different students either provided the same response with different justifications or provided different responses with the same justification.

To clearly illustrate what we mean by evidentiary arguments and how they were constructed and employed, we provide an example of our evidentiary arguments for item 4.1 (shown in Fig 2) in Table IV. This item is in a CMC (coupled multiple choice) format and only has one AO: “S1—Estimate size of random/statistical uncertainty by considering instrument precision.”

As our planned evaluation scheme evaluates each item along potentially multiple AOs, we established these mappings for each AO relevant to each item. In a few instances, when a mapping could be made for one AO but not another, the item was retained and simply not evaluated along the AO for which we could not establish sufficient evidentiary arguments.

The following sections discuss the different stages of piloting and many of the specific changes made to items as we worked to establish evidentiary arguments.

## B. Piloting

We implemented six pilot versions of the assessment between January and November 2022. These pilots consisted of multiple rounds of interviews and classroom implementation (which we refer to as “beta piloting” or simply “betas”). The primary goals of piloting were to

ensure that our items were appropriately interpreted by students and to collect sufficient evidence of student reasoning such that we could form comprehensive evidentiary arguments.

While each assessment item was intended to be presented to students in a particular format (e.g., MC, CMR, etc.), during piloting, we often temporarily changed the response format to gather additional information about student reasoning and student responses. These formatting decisions, as well as other priorities of the various pilots, are described in Table V.

Even with fairly robust evidentiary arguments (as exemplified in Table IV) resulting from 39 interviews and beta testing with around 2000 students, it is likely there are examples of student reasoning that we did not observe. However, we worked to minimize such occurrences by recruiting as many students as possible from different types of institutions and introductory physics courses (using a database of instructors previously constructed [1] and since expanded upon). Additionally, as this assessment is intended to inform instruction at the classroom level, not assign grades or otherwise evaluate students at an individual level, the impact of this limitation is further reduced by reporting averages and aggregated data to instructors and researchers.

The Appendix contains information about these courses and the number of student participants in Table XI and student demographics in Table XII.



TABLE V. The item formats, primary goals, and number of student participants ( $N$ ) for each of the six pilots (presented in chronological order).

Pilot	Purpose(s)	$N$
Interviews 1	(Primarily open-response items) Check item clarity Establish evidentiary arguments Identify potential refinements	9
Beta 1	(Primarily closed-response items) Preliminary validation Identify potential refinements Pilot scoring scheme	911
Interviews 2	(primarily closed-response items) Check item clarity Expand evidentiary arguments	3
Beta 2	(Primarily open-response items) Confirm MC answer options	74
Beta 3	(Primarily final item formats) Pilot preinstruction implementation Expand evidentiary arguments Refine scoring scheme	1048
Interviews 3	(Primarily final item formats) Finalize evidentiary arguments	27

### 1. Pilot interviews

Pilot interviews took place at three distinct stages of SPRUCE’s development. The primary goal of these interviews was to gather evidence of student reasoning in order to establish evidentiary arguments linking student reasoning to student item responses.

#### Interviews: Round 1

The first round of interviews was conducted to ensure item clarity, identify potential item refinements, and begin developing evidentiary arguments.

Through course instructors, we solicited interview participants who had completed an introductory physics lab with a MU component in the previous 12 months. Nine students from four institutions were interviewed between January and February of 2022. Interviews were conducted with students completing the assessment on their computer while screensharing with the interviewer via Zoom. Interviews lasted between 30 min and 1 h and were video and audio recorded. Students were compensated for their time with an electronic gift card.

In the interviews, students worked through the assessment items while the interviewer observed their responses and prompted students to provide reasoning supporting their final responses as well as other answers they considered. The majority of items were presented to students in an open-response format.

#### Interviews: Round 2

A second set of interviews was conducted between June and August of 2022 to verify that our item distractors were sufficiently tempting and to again ensure that items and answer options were clear and understandable to students. We also further expanded our body of evidence of student reasoning by explicitly prompting students to explain their reasoning for not only their response but also, on many items, for a “second-best” response as well.

Interviews were solicited, conducted, and compensated in the same way as the first round of interviews. Despite the low number of participants, these interviews provided valuable data about student reasoning, especially for items that we had changed or were considering changing.

#### Interviews: Round 3

A final set of interviews to finalize our evidentiary arguments was conducted in October and November of 2022. We solicited interviewees (through instructors) from courses that participated in beta 3 (discussed below), so the majority of these students had already taken a prior version of the assessment. Twenty-seven interviews took place with students from eight courses across four institutions. These data provided substantial evidence of student reasoning and also identified a few items where our assessment was not capturing student reasoning as intended, prompting us to make a few minor modifications, and, as the interviews progressed, we began to see very few new ideas being expressed, indicating that we had likely conducted a sufficient number of interviews. These interviews were conducted and compensated in the same way as the previous interviews.

### 2. Full-class beta piloting

During the Spring, Summer, and Fall 2022 terms, we conducted three full-class beta pilots of the assessment, where instructors asked students to take the assessment (generally outside of class). We encouraged instructors to offer participation credit or extra credit to students who completed the assessment, and in most of the courses, the instructors did so. The assessment took most students around 20 min to complete, and they typically had at least a week in which to complete it. Instructors were given a list of students who had completed the assessment but were not given any information on individual student scores or responses.

For all three betas, students could complete the assessment for course credit (if awarded by the instructor) independent of if they consented to allow us to use their responses in our analysis, meaning students could complete the course assignment without granting us permission to use their responses in our analyses. We believe this contributes to some of the courses having a rather low response rate as reported in Table XI, where we report the number and percentage only of students who consented to allow us to use their responses in our research. Additionally, for betas 1 and 2, we removed students

who did not complete at least two of the four experiments in the assessment, though by beta 3 (and in the final version of SPRUCE), we instead included a filter question (e.g., “please enter the number 175 into the text box below”) at the end of the third experiment and removed students who did not reach the filter question or who answered it incorrectly. Filter questions have been used in previous assessments [49] to ensure the quality of responses that are analyzed for research, and unlike the system used in betas 1 and 2, they allow us to remove students who complete the assessment by selecting or entering random responses.

When applied to our data from beta testing, our scoring scheme allowed us to conduct *preliminary* statistical validations of the instrument, specifically using classical test theory (CTT) with couplet scores (as opposed to item scores) as the unit of assessment [43]. In instances where CTT indicated poorly performing couplets, we investigated the couplet to determine if and how to modify the item prompt, the answer options, and/or the scoring scheme. Several specific examples of these changes are given in Sec. IV C, and a full CTT analysis is the focus of an upcoming paper.

### Beta 1

The first beta ran in the Spring of 2022, between interviews 1 and 2. This beta collected responses from students from eight courses at eight different institutions. In beta 1, almost all of the items were presented to students in a closed format (e.g., MC, MR as opposed to NOR), so that we could begin analyzing the distribution of students’ responses across expected common response options, though for many items we did include a “not listed” option that allowed students to enter a response in a text box.

However, one item, item 4.4, was presented in a NOR format despite being designed to be an MC item. Student responses to this item are shown in Fig. 3. The distribution of student responses had peaks at values that corresponded to our planned answer options and, critically, there were no unexpected peaks indicating an attractive distractor that we had not anticipated. This finding informed the development of our second beta, discussed below.

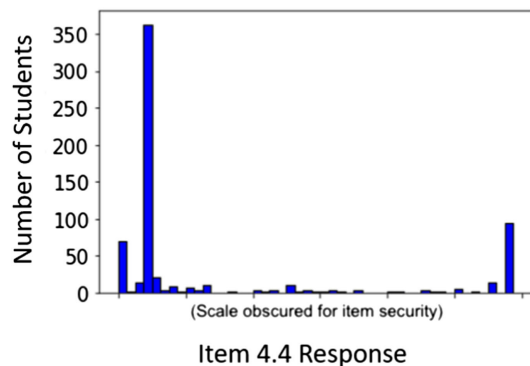


FIG. 3. Histogram showing students’ reported uncertainty values for item 4.4 on beta 1. These responses show peaks at the correct value and at our planned distractors.

### Beta 2

The primary purpose of beta 2 was to verify the reasonableness of our distractors for MC items, and so items were presented to students primarily in an open-response format (e.g., NOR). While this beta was administered only to students in one course, the responses gathered strongly indicated that our previously identified distractors covered the most frequent incorrect answers provided by students, with only a few new distractors being identified through this beta.

### Beta 3

The final round of piloting occurred during the beginning of the Fall 2022 term, where the assessment was administered prior to instruction (as SPRUCE is intended to be used in a pre-post modality). Items were primarily presented to students in their final format.

## C. Piloting-informed item refinement

As discussed above, evidentiary arguments allow for a mapping between student reasoning and student item responses. When this is not the case, items should be modified or discarded. As our evaluation scheme considered each of an item’s AOs independently, when modifying an item, we needed to consider each of the items’ AOs. In the following sections, we provide examples of how items were modified or removed based on our ability to develop sufficient evidentiary arguments. We do not discuss every evidentiary argument, item modification, or even every assessment item in these sections, rather we provide examples to represent the breadth of these arguments while highlighting the items for which establishing evidentiary arguments proved to be the most challenging. These sections are organized according to the four experiments that students work through on the assessment: brief descriptions of the four experiments are given in Table II and further detail is given in the following sections. In addition to changes informed by evidentiary arguments, many small formatting and wording changes informed by student interviews were made to ensure the items and answer options were clear and easily understood.

### 1. Experiment 0: Arrows on a target

The first item on SPRUCE is actually independent of the four experimental contexts and was added because of observed student difficulties during interviews in which students would consistently conflate accuracy and precision. The item presents four targets with different groupings of arrows and asks students to identify which grouping has high precision and low accuracy. This is a canonical scenario for discussing accuracy and precision in physics and was added to allow for the possibility of “calibrating” our interpretation of student responses in items (specifically items 1.1, 3.2, and 4.8) that require students to distinguish between concepts of accuracy and precision in more complex scenarios.

TABLE VI. Experiment 1 item types, descriptions, and AOs.

Item	Type	Description	AOs
1.1	CMR	Given a formula for acceleration, $a$ , in terms of distance, $d$ , and time, $t$ , students are asked what they would do next (and why) after taking one measurement for $d$ and $t$ .	S2, S3, D1, D2
1.2	MC	Students are presented with values of $d$ , $t$ , and their uncertainties and asked to reason about contributions to the uncertainty in the calculated value of $a$ .	H1, H4

While any such calibration would need to be supported by an empirical analysis of student responses, in theory, a student who conflated accuracy and precision on this item may still have a distinct, coherent, and largely correct understanding of these two concepts and may only be confusing the terms. Alternatively, this item may help identify if students who are able to correctly distinguish between accuracy and precision in this simple, likely familiar context are able to identify actions to improve accuracy and precision in more complicated, potentially unfamiliar situations.

### 2. Experiment 1: Cart acceleration

Experiment 1 presents students with an experiment to determine the acceleration of a cart rolling down a ramp. Specific assessment tasks are summarized in Table VI.

Item 1.1 (shown in Fig. 4) is largely modeled after the “Repeating Distance” item from the PMQ [11]. Early iterations of this item consisted of MC and CMC questions; however, the research team was unable to clearly establish evidentiary arguments because multiple explanations, some correct and some incorrect, would lead to different students selecting the same answer options. Eventually, the team decided to present the item as a single CMR item (as shown in Fig. 4), in which students select an answer and also the reasoning that supports their answer.

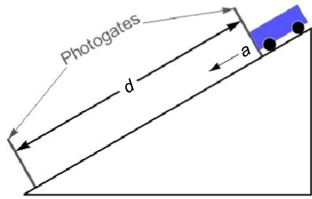
The reasoning elements in the MR part of the CMR item were initially derived from the codes used to score item RD on the PMQ [2,11] and refined based on interviews 2 and 3 and beta 3. Care was taken to ensure that answer options were generally mechanistic in nature: for example, one of the early answer options, “to improve accuracy,” was removed because evidentiary arguments for this answer option were somewhat tautological, as the answer option was redundant with one of the item’s AOs (“S3—Identify actions that might improve accuracy.”) Instead, answer options that explained how accuracy could be improved were added to the item.

### 3. Experiment 2: Mug density

In experiment 2, students are asked to measure the mass and volume of a coffee mug to determine the mug’s density (with uncertainty). Specific item tasks are summarized in Table VII.

For item 2.2, interviews revealed that many students were selecting the correct answer of “standard error (also known as the standard deviation of the mean)” because it contained the word “mean” (and most students had correctly calculated the mean in the previous item). However, when the parenthetical was removed for later interviews and betas, we observed that many students who knew the correct answer to be “the standard deviation of the mean” were unfamiliar with the term “standard error.”

You want to measure the magnitude of acceleration,  $a$ , of a cart rolling down a ramp (pictured below). Your setup includes two photogates that serve as high-precision timers.



To measure  $a$ , you release the cart from rest and measure the time,  $t$ , it takes to travel the distance,  $d$ , between the two photogates. You can then calculate  $a$  using the formula:

$$a = \frac{2d}{t^2}.$$

1.1 You first measure  $d = 1.31$  m. You then release the cart from rest, it passes through the two photogates, and the photogate display reads  $t = 0.987$  s. What do you do next before you calculate  $a$ ?

- Nothing, you use these values of  $d$  and  $t$  to calculate  $a$
- Measure  $t$  again one more time
- Measure  $t$  again multiple times

Please select **all** statements below that support your choice above, **but no others**.

- To find the most common value for  $t$  to use in my calculation
- If I am careful, I should get the exact same number each time
- To find a mean value for  $t$  to use in my calculation
- To be able to calculate an uncertainty
- To finish the experiment in a reasonable amount of time
- To reduce the impact of random fluctuations
- To reduce systematic uncertainties
- To reduce the impact of outliers
- To practice
- To remove outliers

FIG. 4. Item 1.1 asks students what they would do after taking a single measurement for time, then asks students to support that choice.



TABLE VII. Experiment 2 item types, descriptions, and AOs. \*H3 was eventually removed from item 3.2 due to our inability to establish clear evidentiary arguments.

Item	Type	Description	AOs
2.1	MC	Students are asked to report a value for the mass of the mug based on five measurements.	D3
2.2	MC	Students are asked to report an uncertainty for their value of the mass of the mug.	D5
2.3	CNOR	Students are shown before and after images of a graduated cylinder filled with water, where submerging the mug in the water has changed the level of the water line in the cylinders (and students are asked to report the values and uncertainties for the water levels before and after the mug is submerged).	S1, H3*
2.4	MC	Students are asked to propagate uncertainty in the water levels before and after submerging the mug through subtraction in order to determine the uncertainty in the measurement of the volume of the mug.	H1, H2
2.5	MC	Students are asked to propagate uncertainty in the mass and volume of the mug through division to determine the uncertainty in the calculated density of the mug.	H1, H2

This presented the research team with a dilemma as both of these findings threatened our ability to confidently make evidentiary arguments for this item. Ultimately, we decided to keep the parenthetical to avoid arbitrarily large discrepancies between classes based on the particular language used in the course. This decision also impacted item 4.7, where we use the same language.

#### 4. Experiment 3: Spring constant

In experiment 3, students are asked to determine the spring constant of a spring by first measuring the value of a mass and then the period of oscillation of that mass when it oscillates up and down while hanging from the spring. Specific task summaries are presented in Table VIII.

Item 3.2 asks students to select and then justify the number of trials, and the number of oscillations per trial, they would use to obtain a measurement of the period of oscillation. Interviews revealed that different students were employing the same reasoning (e.g., wanting to minimize how much the period changed throughout the measurement) to justify different answers, and conversely, other

TABLE VIII. Experiment 3 item types, descriptions, and AOs.

Item	Type	Description	AOs
3.1	CNOR	Students are asked to identify the mass uncertainty in a single digital scale measurement.	S1, H3
3.2	CMR	Students are asked how many measurements or trials, and then how many oscillations per trial, they would use to measure the period of oscillation for a mass hanging vertically from a spring. Follow-up questions ask for justifications.	Trials: S2, S3, D1, D2 Oscillations: S2, S3
3.3	MC	Students are asked how they would report a value and uncertainty for a single oscillation based on a measurement of ten oscillations and a given uncertainty estimate.	H2, H3
3.4	MR	Students are asked to identify means and uncertainties from other groups (represented numerically) that agree with their mean and uncertainty.	D7

students were using different, often opposing, reasoning to justify the same answer. The research team ultimately elected to present this item in a double CMR format, with one MR follow-up asking students to justify the number of trials and the other to justify the number of oscillations per trial. Student interview responses to this item and to item 1.1 (which targets the same AOs), as well as a qualitative coding of student responses to a “justify your answer” free-response follow-up question on beta 3, informed the development of CMR reasoning element answer options. This item, in its CMR format, was then piloted in the third round of interviews, in which interviewers asked targeted follow-up questions to understand why students did or did not select specific answer options.

Item 3.4 asks students to determine if their measured value with uncertainty (reported numerically) agrees with the measurements of other groups. This item is isomorphic to item 4.3, which presents the exact same relative relationships between measurements using graphs. There is an abundance of research in the physics education literature regarding the use of various or multiple representations in physics (e.g., [50–54], as well as in other science, technology, engineering, and mathematics fields and more generally [55–57]), and these items provide researchers and instructors an opportunity to observe the impact of representation on student reasoning around comparing data.

### 5. Experiment 4: Breaking mass

Experiment 4 intentionally asks students to consider measurement uncertainty in a novel situation: determining how much mass one must hang from a string before the string breaks. This situation is presented in item 4.1 as shown in Fig. 2, and the specific experiment tasks are summarized in Table IX. This item was intended to be novel for students while still being tractable, allowing us to evaluate student proficiency with various AOs in a novel context.

Item 4.1 (shown in Fig. 2) is a somewhat unusual question for an experimental setting in that the resolution of the measurement is quite large (20 g), even for introductory physics labs. During interviews, this feature revealed interesting insights into student reasoning and led to the refinement of the prompts and the inclusion of 521 and 539 g (as the string was able to hold 520 g but broke when an additional 20 g was added) for the mass estimate and 1 and 19 g for the uncertainty estimate. The evidentiary arguments for this item are presented in detail in Table IV.

TABLE IX. Experiment 4 item types, descriptions, and AOs.

Item	Type	Description	AOs
4.1	CMC	Students are asked to identify $m_{\text{breaking}}$ and the uncertainty for a single measurement.	S1
4.2	NOR	Students are asked to report a value for the breaking mass based on ten measurements.	D3
4.3	MR	Students are asked to compare their value (and uncertainty we provide for them, both represented graphically) with the value and uncertainties of other groups.	D7
4.4	MC	Students are asked to calculate the standard error given the mean, number of measurements, and standard deviation.	D6
4.5	CNOR	Given the mean, number of measurements, standard error, and standard deviation, students are asked to report their value and uncertainty with appropriate significant digits.	H3, D5
4.6	MC	Students are asked what the impact on the standard deviation would be when going from 200 to 1000 measurements.	D4
4.7	MC	Students are asked what the impact on the standard error would be when going from 200 to 1000 measurements.	S2
4.8	MC	Students are asked what the impact on accuracy and precision would be when going from 200 to 1000 measurements.	S2, S3

Item 4.2 (also shown in Fig. 2) was initially developed, in part, to address an AO of identifying and removing outliers, as one of the measurements given was substantially different from the rest. However, fewer than 10% of students removed the outlier in beta piloting, and in interviews, students described not removing the outlier for many different reasons, including that they did not notice the outlier, noticed the outlier but did not think it was enough of an outlier to justify removal, or thought it was a substantial outlier but did not feel comfortable removing it without being able to explain why it was an outlier. For these reasons, we removed this AO from this item (and from the assessment as a whole), but, as this was not the only AO addressed by this item, the item remained in the assessment.

## V. DESIGNING FOR, AND ESTABLISHING EVIDENCE FOR, VALIDITY

A valid instrument is one that measures what it says it measures and produces scores that are meaningful measures of the content assessed [17,43,58–62]. Considerations of validity were a primary focus of the development team and led us to use ECD and create AOs, which in turn guided every step of instrument development discussed in this paper. Table X details several types of validity along with design features that support developing a valid instrument. The table also outlines evidence for each of these types of validity, though establishing evidence for validity is the primary goal of an upcoming paper.

As outlined in Table X, design decisions made throughout SPRUCE's development were intended to contribute to SPRUCE's content, face, and external validity. Many of these decisions center on our use of AOs and our extensive piloting.

The types of validity presented in Table X are primarily qualitative in nature. Preliminary quantitative evidence of validity was established through statistical analyses of student responses from pilot phase data using CTT. A full suite of statistical validation statistics using a broad range of student responses to the final assessment will be presented in future work, with such analyses using couplets and couplet scores, rather than items and item scores, as the units of assessment. Such analyses will include CTT, factoring, differential functioning, and pre-post results (i.e., concurrent validity [15], and, eventually, item response theory (IRT) [63] or multidimensional IRT [64].

## VI. INSTRUCTOR REPORTS

One of the main goals of developing an RBAI is to give instructors direct feedback about the impact of their course on student learning along the dimension measured by the assessment. For centrally administered RBAIs, instructors are often provided a report of the analysis of their students' performance. For SPRUCE, we provide an instructor report that not only provides the results from their students but also comparison data from all other courses that have used

TABLE X. Several types of validity, including design features intended to support that type of validity and the evidence needed to show that SPRUCE has that type of validity.

Validity type	Definition	Design features to support validity	Evidence of validity
Content validity	The instrument measures the intended content domain.	AOs derived from instructor interviews. AOs reviewed by instructors throughout development.	Independent matching of items to AOs by two physics education researchers with experimental backgrounds: initial and final agreements with developers of 93% and 99%, respectively. To be further evaluated in upcoming validation paper using statistical methods, though preliminary statistical validations were performed on piloting data and used to guide item refinement.
Face validity	Items appear to measure their intended construct.	Items were created and refined to align with specific AOs.	Established during piloting interviews. Items were also reviewed by instructors at various stages of development.
External validity	Results are generalizable beyond piloting population.	Instructor interviews and student piloting drew from many different institutions, as shown in [1] and Tables XI and XII.	To be established in an upcoming validation paper comparing results between piloting institutions and other institutions.
Criterion validity	Scores correlate with other metrics.	Not explored due to limitations in our institutional review board protocol.	

SPRUCE so far. The main graphic from such a report is shown in Fig. 5. The graph represents preinstruction and postinstruction scores for both the course and all historic data, with statistically significant shifts (as determined by a Mann-Whitney  $U$  test) for each AO shown with solid circles and nonsignificant shifts shown with open circles. Effect sizes for the statistically significant items are calculated using Cohen’s  $d$  and shown on the right side of the chart. Because our data were not normal, which Cohen’s  $d$  relies on for interpretation, we checked our findings using modified forms of Cohen’s  $d$  [65]. This analysis produced similar qualitative effect sizes (small, medium, and large); thus, we report Cohen’s  $d$  for simplicity. Not shown in Fig. 5 are several paragraphs intended to support instructors in interpreting the graphic and effect size. These reports are based on the reports for the E-CLASS that were developed through interviews with instructors [42] and will be refined as feedback from instructors who implement SPRUCE continues to be collected.

For the course represented in this report, one can identify several important features. First, there are four AOs that show statistically significant shifts from pre to post. Those include (1) “H1—Identify when to use fractional versus absolute uncertainly,” (2) “H2—Propagate uncertainties using formulas,” (3) “D7—Determine if two measurements (with uncertainty) agree with each other,” and (4) “D6—Calculate the standard error from the standard deviation.” All of these shifts are positive (with postscores higher than prescores) and have small effect sizes in the range of 0.11–0.41. All four of these AOs align with course learning goals. However, there are many other AOs that also align with the course goals that show no statistically significant shifts. These observations can lead to actionable items for

the course instructor. For instance, “S3—Identify actions that might improve accuracy” is a main goal for the course but shows no improvement over the semester. In this case, the instructor might consider interventions to target that goal, such as allowing for time in the lab for students to iteratively refine their apparatus, model, or data-taking procedure.

The second trend to note is that the students in this course score higher (even in the pretest) on average than students in all other courses combined. However, there are larger gains (effect sizes) for many of the AOs for the historical data courses than for the target course. As we develop a large database of SPRUCE results, we can explore, as researchers, courses that succeed in having larger positive gains for specific AOs to understand possible causal effects using additional qualitative data. Additionally, we can explore many research questions using just the quantitative data. For example, we can determine correlations between the AO scores and the activities in the course (collected on the course information survey) or demographic information collected. The results of these studies can then be used by instructors broadly as they make changes to improve their courses.

## VII. SUMMARY AND ONGOING WORK

In this paper, we discussed the need for a widely administrable assessment of measurement uncertainty for introductory physics labs and how we are using the assessment development framework of evidence centered design (ECD) [8] to create the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) to meet this need. While a previous paper [1] discussed background research (domain analysis), the layers of ECD discussed in this paper deal with the creation of assessment



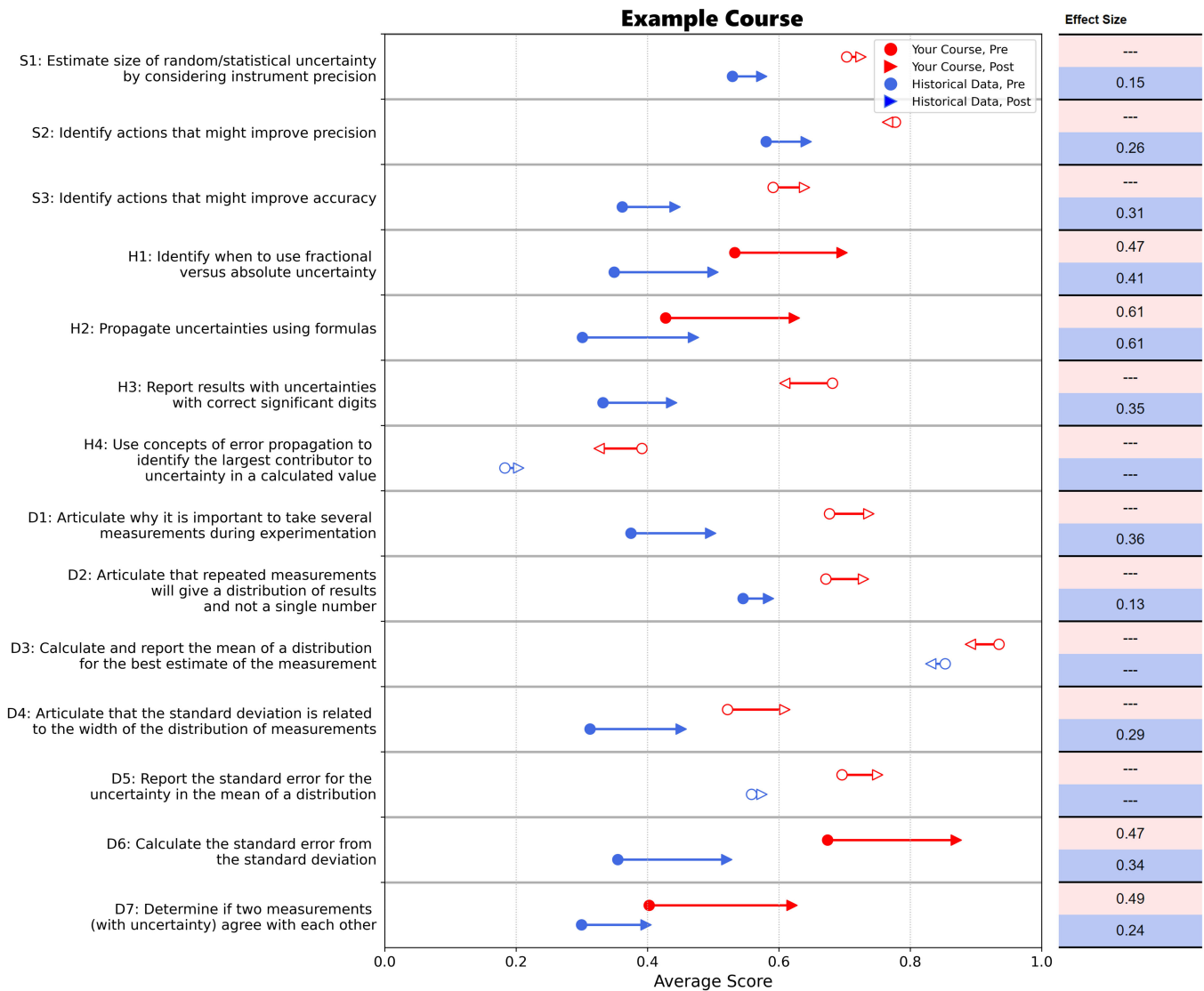


FIG. 5. A portion of an instructor report showing preinstruction and postinstruction scores for the course and all historic data. Statistically significant shifts (as determined by a Mann-Whitney  $U$  test) for each AO are shown with solid circles and nonsignificant shifts are shown with open circles. Effect sizes for the statistically significant items are calculated using Cohen’s  $d$  and shown on the right side of the chart.

objectives and assessment arguments (domain model); instrument design, including the selection of a new scoring paradigm (conceptual assessment framework); item development, piloting, and refinement with a focus on developing evidentiary arguments (assessment implementation); and a portion of an example instructor report (assessment delivery). Future papers will focus on additional aspects of assessment implementation (e.g., scoring) and assessment delivery (e.g., statistical evidence supporting validity).

While the ECD documentation depicts a fairly linear progression through the ECD layers, we found iteration across layers (outlined in Sec. II B) to be extremely valuable and, in our view, necessary to gain the insights that informed the finalized products of the earlier layers. For example, item and AO development informed one

another as we narrowed in on exactly what proficiencies we wanted to measure. Additionally, multiple rounds of piloting allowed us to present the same items to students using different formats (e.g., open response formats where students could input any answer and closed response formats where students selected from a list of possible answer options), which allowed us to gather different types of data on student responses to create more robust evidentiary arguments. All together, these data informed our refinement of AOs, item formats, item prompts and answer options, and evidentiary arguments.

Instructors and researchers who are interested in using SPRUCE in their teaching and/or research can visit the SPRUCE website at Ref. [66] for more information about how to use it in their own classes and studies.

**ACKNOWLEDGMENTS**

This work is supported by NSF DUE 1914840, DUE 1913698, and PHY 1734006. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund. We would also like to thank Robert Hobbs for his work contributing to the domain analysis, as well as the instructors and students

who contributed to the body of data upon which this assessment was built and refined.

**APPENDIX: PILOTING INSTITUTIONS, RESPONSES, AND STUDENT DEMOGRAPHIC DATA**

TABLE XI. The number ( $N$ ) and response rates (RR) of student participants in all six pilots, organized by course and institution.  $N$  is all of the students who consented to participate in the research study and who correctly answered a filter question located at the end of experiment 3: RR is this  $N$  value divided by the total number of students in the course. All courses were introductory laboratory courses at institutions in the United States. R1 and R2 refer to Ph.D. granting institutions (with very high and high research intensity, respectively), M1 and M2 refer to master’s granting institutions (with M1s being larger), BS and AS refer to bachelor’s and associate’s degree granting institutions (respectively), and MSI stands for minority serving institution.

Institution number	Institution type	Course number	Interview 1	Interview 2	Interview 3	Beta 1		Beta 2		Beta 3	
			$N$	$N$	$N$	$N$	RR	$N$	RR	$N$	RR
1	R1	1	4	1	6	180	58%	74	35%	155	37%
2	R2	2	3	...	...	123	40%	...	...	218	31%
3	R2	3	1	...	...	...	...	...	...	...	...
4	R2	4	1	...	...	9	50%	...	...	...	...
5	R1	5	...	...	7	390	75%	...	...	321	74%
5	R1	6	...	2	8	...	...	...	...	112	91%
6	R1	7	...	...	...	...	...	...	...	57	71%
6	R1	8	...	...	1	...	...	...	...	10	67%
6	R1	9	...	...	2	...	...	...	...	29	53%
6	R1	10	...	...	2	...	...	...	...	19	66%
7	AS	11	...	...	1	...	...	...	...	10	40%
8	R1	12	...	...	...	128	31%	...	...	...	...
9	R1	13	...	...	...	33	85%	...	...	35	81%
10	M2	14	...	...	...	25	76%	...	...	...	...
11	M1, MSI	15	...	...	...	23	71%	...	...	17	35%
12	AS	16	...	...	...	...	...	...	...	54	93%
13	R1	17	...	...	...	...	...	...	...	20	95%
14	BS/AS, MSI	18	...	...	...	...	...	...	...	16	73%
15	BS, MSI	19	...	...	...	...	...	...	...	9	100%
16	BS	20	...	...	...	...	...	...	...	6	67%

TABLE XII. Aggregate student demographics of students who participated in SPRUCE piloting and who elected to complete each of the optional demographic questions at the end of the survey.

Demographic category	Interviews ( $N = 39$ )	Betas ( $N \approx 1970$ )
Gender		
Man	51%	59%
Woman	41%	39%
Nonbinary	8%	2%
Not listed	0%	1%
Race or ethnicity		
White	72%	75%
Asian	23%	16%
Hispanic or Latino	10%	10%

(Table continued)

TABLE XII. (Continued)

Demographic category	Interviews ( $N = 39$ )	Betas ( $N \approx 1970$ )
Black or African American	0%	4%
American Indian or Alaska Native	3%	1%
Native Hawaiian or other Pacific Islander	0%	1%
Not listed	3%	3%
English as a first language		
Yes	87%	87%
No, but I am fluent in English	8%	10%
No, and I sometimes struggle with English	5%	2%
No, and I often struggle with English	0%	1%

- [1] B. Pollard, R. Hobbs, R. Henderson, M. D. Caballero, and H. J. Lewandowski, Introductory physics lab instructors' perspectives on measurement uncertainty, *Phys. Rev. Phys. Educ. Res.* **17**, 010133 (2021).
- [2] 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).
- [3] D. R. Dounas-Frazer and H. J. Lewandowski, The Modelling Framework for experimental physics: Description, development, and applications, *Eur. J. Phys.* **39**, 064005 (2018).
- [4] Analyzing and Interpreting Data | Next Generation Science Standards (2022).
- [5] 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.
- [6] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Longman, Harlow, England, 2001).
- [7] 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).
- [8] R. J. Mislevy and M. M. Riconscente, Evidence-centered assessment design: Layers, structures, and terminology, SRI International Center for Technology in Learning, Technical Report, 2005.
- [9] 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).
- [10] 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).
- [11] B. Campbell, F. Lubben, A. Buffler, and A. Sallih, *Teaching Scientific Measurement at University: Understanding Student's Ideas and Laboratory Curriculum Reform* (Southern African Association for Research in Mathematics, Science and Technology Education, Windhoek, Namibia, 2005).
- [12] B. Pollard, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab course on students' understanding of measurement uncertainty, presented at PER Conf. 2018, Cincinnati, OH, [10.1119/perc.2017.pr.073](https://doi.org/10.1119/perc.2017.pr.073).
- [13] 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. 2018, Cincinnati, OH, [10.1119/perc.2017.pr.056](https://doi.org/10.1119/perc.2017.pr.056).
- [14] B. Pollard, R. Hobbs, D. R. Dounas-Frazer, and H. J. Lewandowski, Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses, presented at PER Conf. 2019, Provo, UT, [10.1119/perc.2019.pr.Pollard](https://doi.org/10.1119/perc.2019.pr.Pollard).
- [15] 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).
- [16] N. G. Holmes, C. E. Wieman, and D. A. Bonn, Teaching critical thinking, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11199 (2015).
- [17] J. Day and D. Bonn, Development of the Concise Data Processing Assessment, *Phys. Rev. ST Phys. Educ. Res.* **7**, 010114 (2011).
- [18] 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.
- [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] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [21] E. Mazur, Farewell, lecture?, *Science* **323**, 50 (2009).



- [22] A. P. Fagen, C. H. Crouch, and E. Mazur, Peer instruction: Results from a range of classrooms, *Phys. Teach.* **40**, 206 (2002).
- [23] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, *Phys. Rev. ST Phys. Educ. Res.* **10**, 010120 (2014).
- [24] M. Eblen-Zayas, The impact of metacognitive activities on student attitudes towards experimental physics, presented at PER Conf. 2016, Sacramento, CA, [10.1119/perc.2016.pr.021](https://doi.org/10.1119/perc.2016.pr.021).
- [25] 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).
- [26] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, *Am. J. Phys.* **69**, S12 (2001).
- [27] 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).
- [28] H. R. Sadaghiani and S. J. Pollock, Quantum mechanics concept assessment: Development and validation study, *Phys. Rev. ST Phys. Educ. Res.* **11**, 010110 (2015).
- [29] K. D. Rainey, M. Vignal, and B. R. Wilcox, Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage, *Phys. Rev. Phys. Educ. Res.* **16**, 020113 (2020).
- [30] S. White Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics, *Phys. Rev. Phys. Educ. Res.* **17**, 020129 (2021).
- [31] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010101 (2006).
- [32] D. R. Dounas-Frazer, L. Ríos, B. Pollard, J. T. Stanley, and H. J. Lewandowski, Characterizing lab instructors' self-reported learning goals to inform development of an experimental modeling skills assessment, *Phys. Rev. Phys. Educ. Res.* **14**, 020118 (2018).
- [33] L. Ríos, B. Pollard, D. R. Dounas-Frazer, and H. J. Lewandowski, Using think-aloud interviews to characterize model-based reasoning in electronics for a laboratory course assessment, *Phys. Rev. Phys. Educ. Res.* **15**, 010140 (2019).
- [34] PhysPort Assessment Resources, <https://www.physport.org/>.
- [35] LASSO: Learning About STEM Student Outcomes, <https://learningassistantalliance.org/>.
- [36] R. J. Mislevy, G. Haertel, M. Riconscente, D. W. Rutstein, and C. Ziker, Evidence-centered assessment Design, in *Assessing Model-Based Reasoning using Evidence-Centered Design*, SpringerBriefs in Statistics (Springer International Publishing, Cham, 2017), pp. 19–24.
- [37] R. Millar, F. Lubben, R. Got, and S. Duggan, Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance, *Res. Pap. Educ.* **9**, 207 (1994).
- [38] 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](https://doi.org/10.1119/perc.2022.pr.Vignal).
- [39] B. Priemer and J. Hellwig, Learning about measurement uncertainties in secondary education: A model of the subject matter, *Int. J. Sci. Math. Educ.* **16**, 45 (2018).
- [40] B. M. Zwickl, N. Finkelstein, and H. J. Lewandowski, Incorporating learning goals about modeling into an upper-division physics laboratory experiment, *Am. J. Phys.* **82**, 876 (2014).
- [41] B. Van Dusen, M. Shultz, J. M. Nissen, B. R. Wilcox, N. G. Holmes, M. Jariwala, E. W. Close, H. J. Lewandowski, and S. Pollock, Online administration of research-based assessments, *Am. J. Phys.* **89**, 7 (2021).
- [42] B. R. Wilcox, B. M. Zwickl, R. D. Hobbs, J. M. Aiken, N. M. Welch, and H. J. Lewandowski, Alternative model for administration and analysis of research-based assessments, *Phys. Rev. Phys. Educ. Res.* **12**, 010139 (2016).
- [43] P. V. Engelhardt, An introduction to classical test theory as applied to conceptual multiple-choice tests, <https://www.per-central.org/items/detail.cfm?ID=8807>.
- [44] D. F. Treagust, Development and use of diagnostic tests to evaluate students' misconceptions in science, *Int. J. Sci. Educ.* **10**, 159 (1988).
- [45] 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](https://doi.org/10.1119/perc.2020.pr.Pollard).
- [46] K. D. Rainey, M. Vignal, and B. R. Wilcox, Validation of a coupled, multiple response assessment for upper-division thermal physics, *Phys. Rev. Phys. Educ. Res.* **18**, 020116 (2022).
- [47] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, *Phys. Rev. Phys. Educ. Res.* **18**, 010141 (2022).
- [48] M. Vignal, G. Geschwind, R. Henderson, M. D. Caballero, and H. J. Lewandowski, Couplet scoring for research based assessment instruments, *Phys. Rev. Phys. Educ. Res.* (to be published).
- [49] B. R. Wilcox and H. J. Lewandowski, Students' views about the nature of experimental physics, *Phys. Rev. Phys. Educ. Res.* **13**, 020110 (2017).
- [50] R. J. Dufresne, W. J. Gerace, and W. J. Leonard, Solving physics problems with multiple representations, *Phys. Teach.* **35**, 270 (1997).
- [51] D. Rosengrant, A. Van Heuvelen, and E. Etkina, Case study: Students' use of multiple representations in problem solving, *AIP Conf. Proc.* **818**, 49 (2006).
- [52] 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).
- [53] M. Vignal and B. R. Wilcox, Investigating unprompted and prompted diagrams generated by physics majors during problem solving, *Phys. Rev. Phys. Educ. Res.* **18**, 010104 (2022).
- [54] G. Geschwind, M. Vignal, and H. J. Lewandowski, Representational differences in how students compare measurements, in *Physics Education Research Conference 2023* (to be published).
- [55] R. Cox and P. Brna, Supporting the use of external representations in problem solving: The need for flexible learning environments, *J. Artif. Intell. Educ.* **6**, 239 (1995).
- [56] O. Pamafes and A. Disessa, Relations between types of reasoning and computational representations, *Int. J. Comput. Math. Learn.* **9**, 251 (2004).
- [57] B. Hand and A. Choi, Examining the impact of student use of multiple modal representations in constructing arguments in organic chemistry laboratory classes, *Res. Sci. Educ.* **40**, 29 (2010).
- [58] J. K. Hemphill and C. M. Westie, The measurement of group dimensions, *J. Psychol.* **29**, 325 (1950).
- [59] R. J. Rovinelli and R. K. Hambleton, On the use of content specialists in the assessment of criterion-referenced test item validity, *Tijdschrift voor onderwijsresearch* **2**, 49 (1977).
- [60] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory* (McGraw-Hill, New York, 1994), 3rd ed.
- [61] AERA, APA, and NCME, Standards for Educational & Psychological Testing (2014 Edition).
- [62] R. Lindell and L. Ding, Establishing reliability and validity: An ongoing process, *AIP Conf. Proc.* **1513**, 27 (2013).
- [63] F. M. Yang, Item response theory for measurement validity, *Shanghai Arch. Psychiatry* **26**, 171 (2014).
- [64] 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).
- [65] J. C.-H. Li, Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data, *Behav. Res. Meth. Instrum. Comput.* **48**, 1560 (2016).
- [66] SPRUCE for Instructors | JILA—Exploring the Frontiers of Physics, <https://jila.colorado.edu/lewandowski/research/spruce-instructors>.