Development, Validation, and Initial Outcomes of a Research-Based Assessment Instrument Probing Students' Proficiency with Measurement Uncertainty

by

Gayle Caryn Geschwind

B.S., Stony Brook University, 2017

M.S., University of Colorado Boulder, 2019

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Physics

2024

Committee Members:

Heather J. Lewandowski, Chair

Noah Finkelstein

Seth Hornstein

Thomas T. Perkins

Bethany Wilcox

Geschwind, Gayle Caryn (Ph.D., Physics)

Development, Validation, and Initial Outcomes of a Research-Based Assessment Instrument Prob-

ing Students' Proficiency with Measurement Uncertainty

Thesis directed by Prof. Heather J. Lewandowski

Physics education research in undergraduate laboratory courses is vital to ensure that these courses, which often require extensive resources, achieve their learning goals. These courses often provide students with valuable skills and knowledge not covered elsewhere in the physics curriculum, such as hands-on technical skills. The concepts and practices around measurement uncertainty are especially important and are frequently taught only in lab courses. These skills are valuable to students both for their own work and interpreting others' work. In this dissertation, I cover my role in developing a research-based assessment instrument called the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) aimed at examining student proficiencies and challenges with measurement uncertainty along 10 different axes, such as propagating errors, correct use of significant figures, estimating the size of uncertainty by considering instrument precision, and determining if two measurements with uncertainty agree with one another.

I then describe the creation of a novel scoring scheme using assessment objectives to score assessment items, which provides more information to instructors about the areas where students excel and those where students struggle. I also provide details of the validation of SPRUCE using classical test theory to present evidence for validity and reliability of the assessment instrument. Finally, I analyze preliminary results from SPRUCE, including a deeper dive into one specific set of questions addressing students' abilities with comparing measurements, as well as a broader look at the entirety of SPRUCE and how major and gender correlate with student performance.

This dissertation also describes my work towards creating a classification scheme, or taxonomy, of undergraduate laboratory courses globally. This scheme can be useful in providing more relevant information to instructors, as they can more easily compare their courses with other similar

courses. Further, this information can be used by physics education researchers in identifying which types of courses their research applies to. While the taxonomy is not yet complete, this dissertation covers the development of the survey, including interviews with instructors in 22 countries, and preliminary results from the survey, which involves a landscape of undergraduate lab courses based on 217 courses in 41 countries worldwide.

Overall, the work presented here can help future physics education researchers delve more into undergraduate lab courses to help bolster these and ensure they achieve their goals.

Dedication

To my mom, who spent countless hours building K'nex with me,

To my dad, who taught me math and statistics through football,

And to my husband, whose curiosity inspires my own.

Acknowledgements

First and foremost, I'd like to thank to my advisor, Heather Lewandowski. I never expected my journey to lead me from an REU under her guidance in 2016 to working with her in my graduate career, and it's been the most wonderful experience. Special thanks to the Lew Crew – Mike, Tori, Alex, Rachael, Kristin, and Joey – and the entire CU PER group. In particular, Mike helped me transition into PER seamlessly. Danny Caballero has been a wonderful collaborator on everything SPRUCE. I'm also so grateful to Tom Perkins, who helped mold me into the scientist I am today, as well as my coworkers from his group – Lyle, Devin, David, Arnulf, Marc-Andre, and Patrick.

Thanks to Kelly, Dick, Chris, and the Longmont High School SMART team for keeping me excited about science and making my Tuesday nights memorable.

My friends have been the best part of my grad school experience. From my cohort, Olivia, Dan, Keenan, Diego, Ryan, Charlie, and Garrison have made my time here fun, especially our trivia nights and office hangman before tests. My Thanksgiving co-host Mike helped me realize the full potential of my cooking abilities. Andrea has always been around for a venting session. Rachael listened to all the hockey gossip and made me laugh even on the worst days.

A special shout-out to my family for all of their love and support: Mom, Dad, Valerie, Marilyn, Steve, Dan, Spencer, Riley, Rowan, and Aunt Sue & Andy. Pennington is the best Zoom buddy and Curtis Martin provided endless cuddles and early morning wake-ups. Of course, my husband Keith has been the most patient and caring person; I love him more than anything.

Finally, I'd like to acknowledge the New York Jets. When faced with the challenges of grad school, their struggles served as a reminder: things can always be worse.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

1.1   The Importance of Laboratories in Physics

Physics is an experimental science; though theoretical models are useful, they need to be validated by experiment. In order to train the next generation of physicists, we need to include both theory and laboratory skills in our curricula to ensure they have the necessary tools for success [43].

While it would be ideal for all undergraduate students to participate in authentic physics research in professional and academic labs, the reality is that these resources are scarce and few students have this opportunity [88, 210, 228]. Thus, lab courses are extremely important as they might be the only taste of experimental physics the majority of undergraduate students experience before graduation. Students in these courses can learn unique skills, which can then prepare them for future careers in a wide variety of areas. However, undergraduate physics laboratory courses are often overlooked as less valuable learning experiences than lecture courses despite helping students learn about the nature of experimental physics, a topic not covered elsewhere in the curriculum. Further, lab instruction suffers due to personnel limitations, financial constraints, aging equipment, outdated experiments, and lack of advanced-level courses [81].

Lab courses frequently have many goals, including practicing scientific writing skills, learning how to visualize data, designing experiments, developing mathematical models of experimental results, constructing knowledge, and fostering technical and practical lab skills [138]. One important topic often emphasized in lab courses but not elsewhere in the physics curriculum is measurement

uncertainty, including both correct propagation of errors in students' own experiments as well as valid interpretation of others' reported measurements and their uncertainties [45, 50, 134, 146, 177, 206, 217].

1.2   Physics Education Research in Laboratory Courses

While lecture courses have traditionally been a large focus of physics education research (PER), studying laboratory courses is also important and is becoming more popular [164, 176]. Researchers have been analyzing the goals and outcomes of laboratory courses for decades, often finding that these classes fall far short of their intended goals. Some began to seriously question the effectiveness of these courses in the 1950s, 1960s, and 1970s [86, 117, 208, 211], leading to a call in 1982 for more research on the importance of laboratory courses and how best to utilize them [109]. Frequently, learning goals for lab courses are conflated with general science learning goals that are applicable to lecture courses, rather than being specifically focused on laboratory work and its aspects that uniquely contribute to learning science. Additionally, researchers have found a large gap between the outlining of best practices in labs and the implementation of these practices. For example, instructors often rely on practices shown to be ineffective, such as "cookbook" labs, which direct students on exactly how to interact with equipment and therefore don't allow students to engage in deeper thinking. Often, effective practices require resources that simply aren't available at many higher education institutions, or require more time and skill to implement than can reasonably be expected for instructors [110].

Prior research has highlighted some of the failures of current laboratory course frameworks. For example, some studies have found laboratory courses that focus solely on reinforcing physics concepts students have previously learned in lecture courses (as opposed to learning new concepts) often result in no net gain for student understanding in physics [113, 115]. This work encouraged departments to consider whether their lab courses are meeting their current goals (as well as to examine whether these goals are appropriate) and offered some suggestions to help improve these courses. Frequently, lab courses that focus on reinforcing content are much more structured

and guided due to the need for students to obtain the "correct" answer (for example, measuring gravitational acceleration by using a pendulum). The activities students engage in during these structured activities typically include following explicitly written step-by-step instructions to collect data and using templates for writing their results. Thus, students gain little from the actual hands-on aspect of the activities. Suggestions include having students participate in more open-ended projects rather than guided lab activities as well as integrating lecture and lab activities.

Further work also shows that students who are enrolled in laboratory courses that emphasize reinforcing physics concepts that are already familiar to students from lecture courses leads to students developing fewer expert-like attitudes and beliefs about the nature and importance of experimental physics as compared with students enrolled in courses that specifically focus on developing lab skills [213, 255]. Courses that focus on both of these goals also do not experience as positive a shift in students' attitudes towards expert-like as courses that focus solely on laboratory skills. Additionally, lab courses that focus on reinforcing physics concepts lead to less improvement in students' critical thinking skills than lab courses that emphasize developing experimentation skills [239].

Other research has shown little improvement in student performance in physics theory courses for students who took a laboratory course as part of an introductory physics course as opposed to those who did not take the lab [249]. In this study, students who took the lab course did not have extra benefits in mastering physics content as compared with their peers who elected not to take the lab, based on final exam performance in these courses at a large university. Again, this points to a need to both redefine the learning goals for lab courses and reform them to achieve these goals, rather than ascribing lab and lecture courses the same set of goals.

Overall, many suggestions for improving laboratory courses, including a careful examination of course goals and a shift to open-inquiry labs, might help these courses better prepare students for future experimental physics experiences.

1.2.1   Physics Laboratory Course Transformations

Large-scale transformations of laboratory courses have been heavily investigated in order to help them achieve learning goals more successfully. In transforming lab courses, it is essential to consider the desired learning goals for the course, which may (and ought to) be separate from learning goals for a related lecture course [269]. Any lab activities in the redesigned course should target these learning goals in some way. Overall, regardless of the transformation that takes place, transformed courses frequently lead to improved attitudes about science and experimental physics [187, 247, 251] and improved laboratory skills such as lab notebook practices [144], modeling [145], and critical thinking skills [239].

One major focus of transforming lab courses is trending away from reinforcing physics concepts students learn in lecture and moving instead towards allowing students to learn new physics concepts through the laboratory activities. As mentioned previously, prior research has shown that lab courses that heavily emphasize reinforcement often do not successfully accomplish their learning goals in that students do not gain more conceptual physics knowledge or traditional lab skills [113, 115, 249]. On the other hand, open-ended inquiry lab courses, where students are investigating unknown (to them) physics, leads to improvement in a variety of skills, such as critical thinking [239], use of many representations [40], measurement uncertainty [188], and engagement in expert-like practices [213], as well as improved attitudes about experimental physics [252].

Several models of teaching laboratory skills have emerged, including Modeling Instruction [244], Studio Physics [264], Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) [30], the Investigative Science Learning Environment (ISLE) [40, 78], and Thinking Critically in Physics Labs [207]. Each of these unique teaching methods aids students in learning laboratory skills in different ways.

In Modeling Instruction, lab and lecture are integrated and students must build, test, and revise structural models [36, 244]. This allows students to see physics as a coherent field where experiment and theory work together to mathematically model the world around them. This type

of instruction has been shown to improve conceptual reasoning [36, 38], attitudes [38, 39, 93], and retention of underrepresented students in physics [93, 205].

Studio Physics similarly integrates a lab and lecture environment that encourages student collaboration and group work. The defining characteristics of these courses are small class sections (typically with a maximum of 45 students), collaborative group work, the use of computers, and an emphasis on faculty-student interaction [56]. Teaching assistants circulate throughout the classroom while students are engaged in coursework, but there is no explicit way to train the teaching assistants detailed for this setup. This lack of structured training means that the effectiveness of studio physics can vary drastically. Early approaches to studio physics did not show significant learning gains, but as studio physics has evolved since its first introduction, it has become more effective. In fact, one study showed that it is more effective at teaching introductory Newtonian mechanics than traditional standalone lectures [56]. Additionally, it is important to note the extensive space and time requirements for a studio-based course over a traditional lecture and lab setup; because of the necessity of small class sizes and faculty involvement in all sections, this approach is often difficult to implement at large universities.

SCALE-UP also features an integrated laboratory environment where the space is specifically designed to facilitate interactions between small groups [30]. The lectures in this setup are class-wide discussions rather than a traditional lecture, and there is no delineation between the lab course and the lecture course. A large focus of SCALE-UP is the restructuring of the physical space — the entire course of students meets simultaneously (unlike in studio physics, where the course is split into smaller sections), but the classroom is structured into small tables to facilitate group work rather than a traditional large auditorium-style lecture hall [29]. SCALE-UP has been shown to lead to improved learning outcomes for students of all levels [29, 108], genders, and races, as well as improvement in attitudes, attendance, and retention [28]. One major drawback to implementing SCALE-UP is the space required. An entirely redesigned classroom must be implemented, and it must fit all of the students in the introductory physics course, while also providing groups of tables for students to work at, which is especially difficult at large universities.

ISLE encourages the use of authentic scientific practice while learning physics [40,77,78], with a large focus on the process of doing science authentically. It relies on the model of helping students build on their correct intuitions rather than the model of eliciting and confronting misconceptions implemented in many other instructional frameworks [162, 220]. These other methods – such as McDermott's elicit-confront-resolve formulation [162] and other conceptual change learning models – often do not work as intended, since cognitive conflict is often not enough to induce changes in student thinking [179]. Instead, ISLE does not rely on creating cognitive conflict, which might hinder learning due to negative emotions, but rather focuses on students expressing and exploring their own ideas without specifically being asked to make predictions [78].

ISLE can fit into a traditional physics lab course and can also be incorporated as part of a studio or SCALE-UP classroom. The two key features in the ISLE curriculum are student development of ideas and student representation of physical processes in multiple ways. This first feature involves students observing phenomena and looking for patterns, then developing explanations for these patterns and using these explanations to make predictions. Additionally, a strong emphasis is placed on revising explanations as students continue to experiment, thus encouraging students to engage in an expert-like iterative process. The second feature helps students develop representation skills for qualitative and quantitative problem-solving [78]. For example, students learn to draw pictures of experimental apparatus, record data in tables, and construct free-body diagrams to help see patterns in data [78]. This redesigned curriculum improves student performance in conceptual understanding and problem-solving skills [75, 76] and attitudes about physics and experimental physics [63]. Recent work has also found that the effects of learning through an ISLE curriculum persist years after the course ends [42].

The Thinking Critically in Physics Labs were designed in Fall 2021 at Cornell and guide activities for a lab course separate from lecture [207]. Research is still in progress on the effectiveness of these labs. They encourage open-ended exploration rather than guided activities, and they also emphasize critical thinking skills. Currently, lab activities exist for introductory mechanics and introductory electricity and magnetism. They focus on students' working to iteratively collect

data and revise the experimental procedure, evaluate the process and outcomes of an experiment, communicate about the experiment and outcomes, and work in a collaborative environment, all guided by an emphasis on scientific ethics. The laboratory activities are available on PhysPort [13].

Regardless of which method of course transformation is utilized — whether one of the branded approaches listed above or a separate approach — encouraging students to think critically, explore open questions in physics, and learn new content leads to better outcomes both in measurable skills, as well as attitudes towards experimental physics overall.

1.3   Research-Based Assessment Instruments (RBAIs) in Physics Education Research

To determine the success of the transformed course, instructors can implement a variety of research-based assessment instruments, or RBAIs. They can also be used to evaluate whether learning goals are being achieved more broadly. Madsen, McKagan, and Sayre define an RBAI as *"an assessment that is developed based on research into student thinking for use by the wider…education community to provide a standardized assessment of teaching and learning"* [153]. Essentially, these assessments are specifically designed to evaluate student learning in aggregate, rather than to evaluate individual students for the purpose of assigning grades, and are developed using specific research frameworks to ensure validity and reliability. Further, they are statistically validated to ensure that the research results they provide are meaningful.

RBAIs can help instructors evaluate courses in a variety of ways. A common tactic is to administer an assessment pre- and post-instruction in order to evaluate the impact of instruction. These assessments can also be used to evaluate course transformations, compare curricula at different institutions, and evaluate student learning across longer periods of time than a single course (i.e., longitudinal studies). These assessments are intended to help improve instruction rather than to assign individual student grades.

In order to design an RBAI rigorously, many researchers employ a theoretical framework during the development phase, such as evidence-centered design (ECD) [169], the three-dimensional

learning assessment protocol [141], or a framework such as the one described by Adams and Wieman [20]. These frameworks can help ensure that the assessment measures what it intends to (validity) and does so consistently (reliability). They typically begin with an exploratory research phase in order to determine the scope of the assessment and follow the development of the assessment through to the final validation and administration phase.

RBAIs have helped institute vital changes to the ways in which physics is taught at the university level. A particularly notable example is the Force Concept Inventory (FCI) [107]. Mazur's use of questions from this assessment to probe student understanding of Newton's third law in an introductory physics lecture at Harvard led him and others to transform instruction in the lecture setting due to students' lack of conceptual understanding [80, 160].

Currently, RBAIs in physics exist for a variety of topics such as mechanics [107, 230], electrostatics [155, 259], thermodynamics [195], and quantum mechanics [203], as well as other areas related to physics such as attitudes about science [19, 270], quantitative reasoning skills [248], and experimental skills [45, 58, 73, 202, 241].

1.3.1 RBAIs in Undergraduate Physics Laboratory Courses

A multitude of RBAIs exist specifically for use in undergraduate physics laboratory courses. Here, I briefly discuss some highly impactful assessments in the field of PER, including the topics they cover and the information we can learn from them. A large focus of this section surrounds assessments that probe student proficiency of measurement uncertainty, although others are also discussed.

First, the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) was originally developed by Benjamin Zwickl, Heather Lewandowski, and Noah Finkelstein at the University of Colorado Boulder [268, 270]. This assessment aims to examine how students perceive the nature of physics experiments in the specific context of physics laboratory courses, as well as in professional research labs. It is designed to be given pre- and post-instruction. Students answer paired Likert-style questions about their own perceptions of doing experiments, as well as

how they perceive an experimental physicist might respond about their own research.

Wilcox et al. found that students in both introductory and upper-level physics laboratory courses often leave their courses with ideas about experimental physics that are inconsistent with the views of professional physicists, but on a majority of items, they are able to predict the expert-like response even when their own views disagree [257]. While upper-division physics students' responses for their own beliefs are closer to expert responses than introductory students' responses, there is still a large gap between advanced students and experts. Additionally, instruction did not impact advanced students views over the course of a single semester, based on pre- and post-instruction results. However, instruction did impact introductory students responses to E-CLASS: post-instruction responses showed a slight negative shift. Thus, they conclude that laboratory courses generally have not been effective at engaging students and encouraging expert-like views on the nature of experimental physics. This can impact retention and recruitment efforts [257].

Additional research shows a significant improvement in E-CLASS scores in courses that used transformed curricula (including ISLE, Modeling Instruction, Studio Physics, and SCALE-UP) as compared to traditional laboratories at the introductory level [251]. The analysis showed that when accounting for pre-instruction score and major, students in courses with transformed curricula outperformed students in traditional courses on the E-CLASS post-instruction assessment. This effect was larger for women than for men, showing that these transformed curricula have an additional positive impact for women [251].

Researchers also determined that while E-CLASS post-instruction scores have little correlation with students' final course grades, these scores do correlate strongly with students perceptions of the grading structure, in particular, which elements they perceive as important for earning a good grade in their course [250]. Several recommendations to instructors have also been made as a result of E-CLASS, including the use of a transformed curriculum for introductory physics laboratory courses, focusing more on students' lab skills rather than reinforcing physics concepts, and the inclusion of open-ended activities at all levels of physics lab courses. These recommendations all correlate with higher E-CLASS scores post-instruction [258].

There are an abundance of other applications of E-CLASS, including investigating student perceptions of their courses during the COVID-19 pandemic [33,83], longitudinal studies and other curricula transformation [71,104,137,252,256], the impact of gender on perceptions of experimental physics [253], and correlating student performance on E-CLASS with lab course success [250].

Another RBAI, the Modeling Assessment for Physics Laboratory Experiments (MAPLE), examines how students model in the context of physics experiments [64,65,84,202]. This assessment begins with a "choose your own adventure" portion where students are given one experiment to virtually complete and are able to make decisions about which steps to take to investigate the problem they are tasked with solving. Once they finish, they are asked a series of coupled-multiple-response questions about this experimental activity. This RBAI is intended for use in upper-division electronics or optics laboratory courses. Statistical validation of MAPLE is an ongoing effort. Future research hopes to gather enough data to perform a clustering analysis and sort students into different personas of experimental physicists, such as "tinkerers" or "random walkers" [84].

Next, the Physics Measurement Questionnaire (PMQ) delves into measurement uncertainty in the context of physics laboratory courses [45,238]. This assessment presents students with decisions they might have to make in a laboratory context with different options for next steps. Students have to select which option they agree with (in a multiple-choice format) and then explain their reasoning (in an open-response format). Originally used in Cape Town, South Africa, it helped create a paradigm to sort students into pointlike reasoners, setlike reasoners, or mixed reasoners [41]. Within this paradigm, pointlike reasoning indicates students who believe that quantities that are measured have a true value that can be obtained with one perfect measurement. Setlike reasoners are considered to have more expertlike views, as they believe that measurements always have uncertainty, and a true value (if it exists) can never be obtained. Mixed reasoners exhibit some combination of these perspectives.

While the PMQ was largely successful in creating a new paradigm and investigating student proficiencies around measurement uncertainty, it has two large limitations. First, it probes very few ideas surrounding measurement uncertainty. Most questions are related to distributions of results

from repeated measurements. While this is an important aspect of measurement uncertainty, the PMQ is clearly limited in scope as it does not address comparison of measurements, significant figures, and determination of uncertainty based on instrument precision amongst other important skills. The other major limitation is the open response nature of the assessment — while it works well at small institutions, it is far too laborious to score in any large-scale administration due to the requirement of reading through and manually scoring paragraphs of student-written text.

In order to assess a different population of students, a portion of the PMQ was implemented at the University of Colorado Boulder in 2016, 2017, and 2018 [183, 188]. In this study, researchers observed a shift from mixed reasoning to set-like reasoning in pre- to post-instruction. These shifts were present both before and after a course transformation, with a stronger shift towards more expert-like reasoning in the transformed course. However, very few students, as compared with studies in South Africa, exhibited solely point-like reasoning, even pre-instruction. This can be ascribed to significant differences in student preparation between these two studies, and therefore shows that this pointlike and setlike reasoning paradigm may not be able to capture all of the nuances in student reasoning around measurement uncertainty.

Another RBAI in the lab context is the Physics Laboratory Inventory of Critical Thinking (PLIC), which was developed at Cornell University and Stanford University with the goal of assessing student proficiencies around experimental methods, data, and models [114, 194, 240, 241]. This assessment features multiple questions contextualized in a small number of experiments in both multiple choice and multiple response formats. While this test doesn't focus solely on measurement uncertainty, it does deal with many areas of measurement uncertainty extensively. This assessment builds on earlier research emphasizing the importance of teaching critical thinking in physics laboratory courses, including a focus on the ways that student use data and evidence to make choices about next steps [116].

Another important RBAI that addresses laboratory skills is the Laboratory Data Analysis Instrument (LDAI), which was developed in Israel to assess first year students' understanding of data analysis procedures. It consists of 30 multiple choice and true/false questions that are

contextualized in a real laboratory report detailing a step-by-step description of an experiment based on Newton's second law. The four objectives of this assessment are that students should (1) understand the meaning of, and ways to calculate, measures of central tendency, (2) understand the meaning of error and uncertainty, as well as how to compute this and distinguish between statistical and systematic uncertainties, (3) be able to choose graphical representations for their own data and decipher others' graphs, and (4) understand regression lines and how to fit them [73]. This assessment is therefore appropriate when these four areas are part of a course's curriculum. For the true/false questions, the LDAI requires students to write an open-response explanation to accompany their choice in order to receive credit, which increases the difficulty of widespread administration. Researchers implemented the assessment in this way because without the open response requirement, students have a 50% chance of getting the correct answer through random guessing. One implementation of the LDAI in Thailand found that introductory physics students struggled with uncertainty, including a lack of proficiency in distinguishing between systematic and statistical uncertainty. Additionally, most undergraduates in all years struggled with linear regressions, even after taking at least one laboratory course. However, first year students performed significantly worse than second and third year students in this study, showing that instruction might improve these skills to some extent [118], though there might also be selection bias: first year students who perform poorly are more likely to drop out.

Finally, the Concise Data Processing Assessment (CDPA) also examines student proficiencies with measurement uncertainty [58]. This assessment was designed at the University of British Columbia and emphasizes error propagation. It is a multiple choice test designed to be given pre- and post-instruction. This assessment was originally intended to complement an introductory physics lab course aimed at physics majors. As part of assessment validation, it was administered to graduate students who scored just over 50% on average and post-test scores for first year students average less than 40%. Thus, it might not be appropriate for the introductory level, due to high difficulty and less discriminatory power between introductory and graduate students than desired.

Further, the CDPA was used to investigate gender gaps in physics [59]. The CDPA did

reveal a significant gender gap at both the pre- and post-test level. While everyone does learn in the lab courses (as evidenced by improvement on the CDPA in this study), the gender gap remains unchanged post-instruction: men still outperform women. They posit that one reason women do worse on the CDPA is due to a lack of confidence. They further determined that there are gendered differences in student behavior in the lab. For example, they find that men tend to spend more time on the computer in the lab and women tend to spend more time on other activities such as writing or speaking with their peers. Another important result from this research is the use of analysis of covariance (ANCOVA) [97], which shows a significant gender effect and therefore indicates that a common measure of assessment analysis — the use of normalized gains to analyze pre-post shifts — might not be appropriate in many cases in PER research because this method does not account for confounding effects such as gender and major. They encourage other assessment developers to examine the statistical techniques being used to analyze results to ensure that all assumptions of tests are met.

Thus, a multitude of RBAIs in the context of physics laboratory courses exist, all surrounding a wide variety of topics important in these courses. I will discuss another such RBAI, the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE), which was developed to fill a void in the current RBAI landscape, in the next section and throughout this dissertation.

1.4   Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE)

In this section, I detail the important aspects of an assessment created to understand student proficiencies in measurement uncertainty, SPRUCE. I begin with a discussion about the goals and development of SPRUCE, followed by a brief description of the assessment itself. SPRUCE is discussed in more detail in Chapters 3, 4, 5, and 6.

1.4.1 SPRUCE Goals and Development

The main goal of SPRUCE initially was to fill a void in RBAIs: to create a widely administrable (and easily scorable) assessment aimed specifically at introductory undergraduate physics laboratory courses to determine student strengths and struggles with a wide variety of topics related to measurement uncertainty. No other RBAI targets all of these goals simultaneously.

SPRUCE was developed via the evidence-centered design (ECD) framework, which is described in more thorough detail in Chapter 3 as well as by Pollard et al. [184] and Mislevy [168,169]. The development began with interviews of current undergraduate physics laboratory instructors about their views on important components of measurement uncertainty for students to learn.

From these interviews, a list of Assessment Objectives (AOs) was developed; these inform the creation of the assessment items. AOs are *"concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment* [236]*."* Next, the AOs were honed until only ten were left in order to focus the assessment, and following this, assessment items were created to target these AOs. These ten were chosen to define the scope of SPRUCE to be wide enough to target areas instructors felt were important while not being overly broad.

Next, I iteratively created items, beta tested SPRUCE (including piloting the assessment in classrooms and conducting think-aloud student interviews), and revised items. I also revised the AOs during this time. Some AOs were dropped due to an inability to properly assess them. One example of this is an AO that targeted handling of outliers — because the treatment of outliers differs from course to course based on the field in physics and instructor preference, we realized we could not properly assess whether students were handling outliers "correctly" and therefore we removed this AO from SPRUCE.

Finally, with SPRUCE in its final form, wide-scale administration, statistical validation (see Chapter 4), and research into the results of SPRUCE (see Chapters 5 and 6) could occur.

1.4.2   Overall Characteristics of SPRUCE

SPRUCE is a fully online assessment (on Qualtrics) designed to be administered pre- and post-instruction. It consists of 19 questions and takes students a median of 18-20 minutes to complete. [1] The assessment is grounded in four authentic laboratory experiments, some of which students may have previously seen and interacted with, in order to reduce cognitive load and create a more authentic experience. The questions themselves have many formats: multiple choice, multiple response, numeric open response, coupled multiple choice, coupled multiple response, and coupled numeric open response.

Coupled-multiple-choice items ask students to select answers to two multiple-choice questions to give an answer to an overarching question. For example, in SPRUCE, one item directs students to report the final value for a measurement and the next item asks about the uncertainty on that final value. In this way, they are reporting a final result based on their answers to two different questions for a value and an uncertainty, hence why these items are considered coupled.

Coupled-numeric-open-response items work much the same way as coupled multiple choice items, but students are tasked with entering in any number they choose in response to two related questions, as opposed to selecting from a list of options provided.

Coupled-multiple-response items ask students to answer a multiple-choice question, followed by a multiple-response item to assess reasoning for their prior answer. This method of assessment has previously been shown to allow researchers to gather both a student response and reasoning elements in a closed-response format, therefore making assessments more scalable for large classes without requiring the manual coding of thousands of student written responses [259].

SPRUCE is then scored using the previously discussed 10 AOs in a couplet scoring scheme (see Chapter 2 for details on this novel scoring scheme). In short, we can extract single AO scores, as well as one overall score for the entire assessment. Additionally, the final overall score weights all of the AOs equally to give a broad overview of student proficiency at measurement uncertainty, while

---

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

the single AO scores allow for a more fine-grained view. These individual AO scores and overall score are then reported to instructors showing pre-post shifts, allowing instructors to determine actionable items to address in future iterations of their course. For example, they might see that students struggle with particular areas of measurement uncertainty and choose to spend more of the course time focusing on those areas.

1.5 A Worldwide Taxonomy of Physics Laboratory Courses

RBAIs are incredibly valuable tools for both researchers and instructors. Instructors typically receive a report with details about their students' performance on the RBAI, as well as comparison data that contains all historical data from all student responses to the assessment. However, this comparison data might not be applicable in many cases, as it is often dominated by larger R1 institutions in the United States. Further, for RBAIs that are appropriate for both introductory and advanced students, the data are largely dominated by introductory students (due to much higher enrollment), making the comparison data less useful for instructors of advanced courses.

Further, in terms of PER research in laboratory courses generally, it is often difficult to make widely generalizable claims due to research being done at specific institutions. Because there are currently few tools that allow us to compare courses at different institutions (both within the United States and worldwide), it can be a complex issue to determine whether research outcomes are applicable for other institutions.

In order to form a basis for comparing results worldwide, both for the purposes of creating instructor reports for RBAIs and for more widespread use within the PER community, a taxonomy — or a classification scheme — for physics labs can be created and applied appropriately. This will not just aid researchers in being able to compare studies, but it will also help instructors. Many instructors utilize a variety of RBAIs to improve their teaching methods. Instructor reports from these RBAIs often include comparison data so that they might be able to compare their teaching efforts to others'. However, it has been difficult to decide what comparison data should be included in these reports. For example, it seems unreasonable to include the data for a sophomore-level

course aimed at physics majors at an R1 institution as comparison data for a community college introductory-level course aimed at life science majors. The creation of a taxonomy for lab courses will help researchers provide only the appropriate comparison data in instructor reports moving forward.

My work in this regard encompassed the creation of a survey to probe undergraduate physics lab courses around the world, as well as analyzing the data from this survey.

From these data, we can begin to understand the worldwide landscape of undergraduate physics laboratory courses. These data can also be used as a guide to those who wish to transform their courses: they can analyze survey responses (which will be made publicly available) to determine best practices that exist in similar courses and implement these ideas in their own courses. Finally, these data will help PER researchers in ensuring claims are applicable to certain types of laboratory courses, based on the taxonomy scheme.

I discuss work towards creating such a classification scheme in more detail in Chapter 7.

1.6   Dissertation Outline

This dissertation covers my work on two major projects during my graduate studies: the development, implementation, and validation of an assessment aimed at measuring how introductory physics laboratory students handle various aspects of measurement uncertainty (SPRUCE) as well as the development and implementation of a survey aimed at probing undergraduate physics laboratory courses around the world in an effort to create a taxonomy of such courses.

Chapter 2 discusses a novel scoring scheme that can be applied broadly to many assessments; this scoring scheme was applied to SPRUCE. This scoring scheme is based on Assessment Objectives and provides deeper information to both researchers and instructors about student proficiencies in specific sub-areas rather than a single overall score for each student for the entire assessment. Chapter 3 discusses the development of SPRUCE, including the formation of assessment objectives, the development of assessment items, and the iterative process towards a first validation of those assessment items at the level of student reasoning and creation of evidentiary arguments to support

our understanding of student reasoning elements. Chapter 4 is a statistical validation of SPRUCE using classical test theory to provide evidence that it is both a valid and reliable instrument, which can be used to assess student understanding of measurement uncertainty. Chapter 5 discusses a specific Assessment Objective from SPRUCE, which probes handling of measurement comparison with uncertainty via two isomorphic items with different representations. Chapter 6 discusses more broadly students' proficiency with measurement uncertainty as probed by SPRUCE, including the impact of instruction, the importance of instructors' goals for a course and how these impact student learning, and the ways in which students' major(s) and gender(s) correlate with their performance. Chapter 7 discusses steps towards creating a taxonomy for undergraduate laboratory courses, including the development of a survey aimed to probe these courses worldwide. Finally, Chapter 8 provides conclusions from both my work on SPRUCE and the lab taxonomy project, as well as future work for these two areas.

Chapter 2

Couplet scoring for research-based assessment instruments

2.1 Contribution

This chapter is adapted from an article submitted to Discover Education [234].

I helped develop the scoring scheme discussed in this chapter, including its applications to SPRUCE. I was heavily involved in the writing of this chapter, especially the validation section. Further, I edited the paper version of this chapter, including a reorganization of the material before submission to the journal.

2.2 Introduction

Research-based assessment instruments (RBAIs) are surveys, questionnaires, and other tools that help educators (including instructors and education researchers) make informed pedagogical and curricular decisions [72, 152, 153, 157]. These instruments can be used to gather information on student beliefs, experiences, proficiencies, and other aspects of education that are of interest to educators. Unlike course quizzes and summative exams, which typically assess individual students, RBAIs are intended to identify trends in populations of students [72, 152]. Here, we focus on primarily RBAIs that measure student proficiency in specific content areas, which we refer to as *content RBAIs*.

Recently, we created a physics content RBAI for measurement uncertainty [235], employing assessment objectives (AOs) [236] throughout the development process. AOs are statements (similar in structure to learning objectives [24]) about the content the instrument aims to assess. For our

RBAI, these AOs are integral to the interpretation, scoring, and reporting of student responses, as each item is designed to align with one or more AO.

Our use of AOs supported developing an instrument that aligned with our assessment priorities. Indeed, the usefulness of RBAIs depends, in large part, on the degree to which an instrument measures what it purports to measure and how meaningfully these measures are reported to implementers [55, 72, 103, 148, 157, 167, 173, 200]. Typically, these measures are item scores, and they are often reported individually or as an overall assessment score [72, 167].

Here, we introduce and formalize our AO-aligned scoring paradigm for content RBAIs called *couplet scoring*, where a couplet is a scorable item-AO pair. In this paradigm, it is couplet scores, rather than item scores, that serve as the unit of assessment for reporting student proficiencies and validating the instrument. We posit that couplet scoring offers a number of affordances as compared to traditional item scoring, where each item has only one correct answer and the assessment score is generally the sum of the item scores.

The instrument for which we developed couplet scoring is the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) [235]. SPRUCE was developed in parallel with couplet scoring and is used throughout this chapter to demonstrate how this scoring paradigm might be applied to content RBAIs. However, the focus of this chapter is couplet scoring and not SPRUCE itself, and so we also incorporate examples from ongoing work to generate a *post hoc* couplet scoring scheme for the Force Concept Inventory (FCI).

The work presented here has the following research goals:

RG1 Introduce couplet scoring and explore its affordances and limitations; and

RG2 Demonstrate how couplet scoring can be employed in content RBAI development.

In Sec. 2.3, we provide background on assessments and scoring schemes. Sec. 2.4 then introduces our new scoring paradigm. Details of the implementation of couplet scoring are shared in Sec. 2.5, and Sec. 2.6 discusses possible implications for other types of evaluation. A summary and discussion of future work are presented in Sec. 2.7.

2.3   Background

Research-based assessment instruments (RBAIs) are tools used by educators and researchers to gather information from students about teaching, learning, student experiences, and other aspects of education to inform curricular and pedagogical decisions. These instruments are "developed based on research into student thinking...[to] provide a standardized assessment of teaching and learning" when administered to students [153]. Importantly, these instruments are not intended to help educators evaluate individual students or assign grades [72, 152].

In physics, most RBAIs are designed to measure student proficiencies in specific content areas, such as in mechanics [107, 230], electricity and magnetism [46, 155, 259], quantum mechanics [203], thermodynamics [195], or laboratory settings [45,58,65,235,241]. These *content RBAIs* (sometimes called concept inventories [152] or conceptual assessment instruments [148]) have proven valuable in identifying instructional weaknesses [152, 160] and evaluating the effectiveness of instructional changes [44, 94, 133, 152, 188] in physics education. These and other instruments can be found on the PhysPort website [11].

RBAIs can employ a wide variety of item types, including both closed-response (such as multiple choice) and open-response formats (such as providing a written explanation). Closed-response format items, especially multiple-choice items, are generally developed such that one response is considered the correct choice (typically determined by evaluating alignment with expert responses [72]). This convention supports the inclusion of multiple-choice items as opposed to multiple-response (sometimes called multiple-select or multiple-choice-multiple-response [214, 248]) and coupled multiple response items [259]. Single-correct-answer items require minimally complex scoring mechanisms, since the correct answer is given full credit and all other answer options are (typically) given zero credit. This scoring approach is often described as being the most objective [72, 173] and it generates scores that work well with validation algorithms [173]. Even when using item formats other than multiple-choice, having a single correct answer is common and aligns with instructor and student expectations around assessment.

It is worth noting that many existing instruments including scoring that goes beyond the strict "one correct answer per item" model. The Brief Energy and Magnetism Survey has several items in which the score depends not only on the response to that item but also to other, previous items [46]. The developers of the Force and Motion Conceptual Evaluation [230] advocate for an alternative scoring scheme that includes consolidating 3 different groups of 3 questions and awarding 2 points if all of the items in a group are answered correctly [229]. Many instruments employ various two-tier questions, with perhaps coupled multiple-response items (as discussed in Refs. [259], [195] and [196]) representing the greatest departure from conventional assessment items. However, these examples still represent a traditional scoring scheme, in which each item is intended to measure one proficiency or construct, even if there is variation in how that measurement is implemented.

To capture and report more information from existing RBAIs, researchers have worked to identify sub-scales related to different concepts covered on the RBAI [102, 167, 219], with sub-scale mean scores being reported in addition to an overall mean score. However, such sub-scale analyses are typically time and labor intensive to develop, external to the report provided to instructors, and not considered in the design and validation of the instrument [167]. For example, Stewart et. al. [219] and Hansen and Stewart [102] found the FCI and another popular physics content RBAI had a "lack of coherent sub-scales", limiting the usefulness of such post hoc sub-scale approaches.

Additionally, work has been done to learn about student reasoning and proficiencies from their incorrect answers [37, 214]. While much of this work is done post instrument development, some developers have worked to encode some information about the quality of incorrect answers through partial credit scoring schemes, where some distractors, rather than being worth no points, are worth a fraction of the points that are given for the correct answer. While this scoring model can be more sensitive in measuring student proficiencies, the exact fraction of a point earned for these answers is subjective and can restrict which validation algorithms one can use, removing one of the advantages gained by having single-correct items. Additionally, the nuance of exactly what part of a response earned a student credit is difficult, if not impossible, to convey in an overall assessment scores or even in the item score.

Finally, a related method, called choice modeling or choice analysis, has been used by researchers in economics, marketing, and transportation to explore choices made by survey respondents based on both selected and not selected options. Choice-level analyses can extract information beyond the final result, and are used to understand the opinions and rationale behind peoples' decisions [32, 151, 163]. However, this context is well outside of the sphere of RBAIs and physics education, being used to understand priorities and preferences rather than skills and proficiencies.

Here, we present another model of scoring an assessment, called couplet-scoring, that goes beyond scoring an item as either correct or incorrect and allows for additional information to be reported about student reasoning along several constructs. We will discuss how consideration of each of the above ideas contributes to the development of this scoring scheme.

2.4 Couplet Scoring for Research-Based Assessment Instruments

In this section, we present our new assessment scoring paradigm, couplet scoring. Central to couplet scoring is the use of assessment objectives (AOs), which we introduced previously [235, 236] and summarize below.

2.4.1 Assessment Objectives

AOs are "concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment" [236]. In the language of assessment development, AOs are the constructs the assessment aims to measure, only they are articulated as objectives. Table 2.1 includes some example AOs from SPRUCE and several theoretical principles developed by Stewart et. al for the FCI [219], which we employ as preliminary AOs in this chapter[1] to illustrate what couplet scoring might look like with the FCI.

In a previous paper [236], we outlined four broad affordances of AOs for instrument development: facilitating incorporating instructor priorities into the instrument, providing a means for

---

[1] These principles were developed as part of a multidimensional IRT analysis of the FCI and not as a set of AOs. However, they span a similar set of ideas as would AOs, and thus are adequate stand-ins for AOs for the example presented in this chapter.

Table 2.1: Examples of AOs from SPRUCE and theoretical principles from the FCI that serve as the item constructs in the examples below.

SPRUCE AO examples

S2 Identify actions that might improve precision

S3 Identify actions that might improve accuracy

H1 Propagate uncertainties using formulas

H2 Report results with uncertainties with correct significant digits

D1 Articulate why it is important to take several measurements during experimentation

D5 Determine if two measurements (with uncertainty) agree with each other

FCI theoretical principle examples [219]

C2 Objects moving in a curved trajectory will experience centripetal acceleration

L2 Newton's 2nd law

L4 Objects near the earth's surface experience a constant downward force/acceleration of gravity

F2 An object does not necessarily experience a force in the direction of motion

evaluating and scoring authentic items, providing a structure for feedback to implementers, and serving as a means for communicating the content of the instrument to potential implementers [236]. Many of these affordances complement those of couplet scoring, discussed in Sec. 2.4.3.

2.4.2   Couplet Scoring

Couplet scoring, as the name suggests, is a scoring paradigm in which item-AO couplets (or simply "couplets") are scored. This is in contrast to many traditional scoring paradigms in which each item is scored once.

Conceptually, a couplet is an assessment item viewed and scored in light of a particular AO. Multi-AO (i.e., multi-construct) items have a couplet for each AO, as depicted in Fig. 2.1 for an item that has two AOs and therefore two couplets. Each of these couplets is scored by considering

only that couplet's AO.
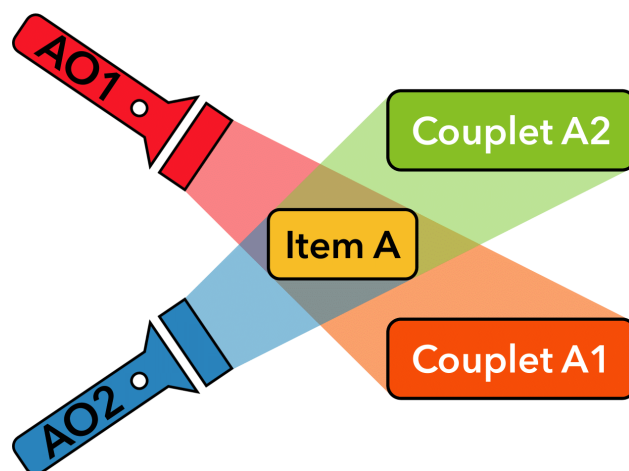
Figure 2.1: A simple graphic depicting how two AOs, represented as flashlights, can "illuminate" the same item and yet produce independent couplets.

2.4.2.1   Couplet scoring example - a simple model

Similar to how the image in Fig. 2.1 uses a simple graphic to illustrate how AOs and items combine to produce couplets, we now use a simple model of an instrument (comprised of only three multiple choice items and two AOs) to illustrate how couplet scoring is implemented and how the AO scores and the overall score (for one student) can be calculated. Table 2.2 shows the AO scores associated with each possible answer on this instrument. Figure 2.2 shows how to use student responses to the items in the model to compute couplet scores, AO scores, and, finally, an overall score.

Since many answer options for items on RBAIs are developed to be tempting distractors (i.e., results that may be correct in some ways but incorrect in others), it is possible (and, for SPRUCE, fairly common) for a couplet to award the maximum number of points to several different possible responses, even for closed-form items such as multiple-choice items.

To calculate an AO score for a student, each item that addresses that AO is scored to form a couplet score (See Fig. 2.2). The couplet scores for that AO are summed together, rounded to the nearest integer, and divided by the number of couplets for that AO. This produces an AO score

Table 2.2: Couplet scoring scheme for a sample instrument comprised of three items. Note that Item A does not probe AO2 and Item C does not probe AO1 in this example, similar to a real-world application where not every item will probe every AO.

Answer Option	Score	
Item A		
	A01	A02
a	1	-
b	1	-
c	0	-
d	0	-
e	0	-
f	0	-
Item B		
	A01	A02
a	0	0
b	1	1
c	0	1
d	1	0
e	1	1
Item C		
	A01	A02
a	-	1
b	-	0
c	-	1
d	-	0

with a minimum of 0 and a maximum of 1.

While the specifics of calculating scores should reflect the priorities of the developers and instructors, here we discuss an example based on the method used to calculate scores for SPRUCE. In Tab. 2.2, we can see a student who selected option b on all three of items A, B, and C would have AO scores of 1 on AO 1 (one point for couplet AO 1 - Item A, one point on couplet AO 1 - Item B, and item C is not scored in this couplet) and 0.5 for AO2 (item A is not scored for this couplet, one point on couplet AO 2 - Item B, and no points on couplet AO2 - Item C). Similarly, a second student who selected option c on item A, option b on item B and option d on item C would have AO scores of 0.5 on AOs 1 and 2.

Once all of the AO scores are calculated for a student, they can then be summed together and divided by the number of AOs to create an overall score, which has a minimum of 0 and a
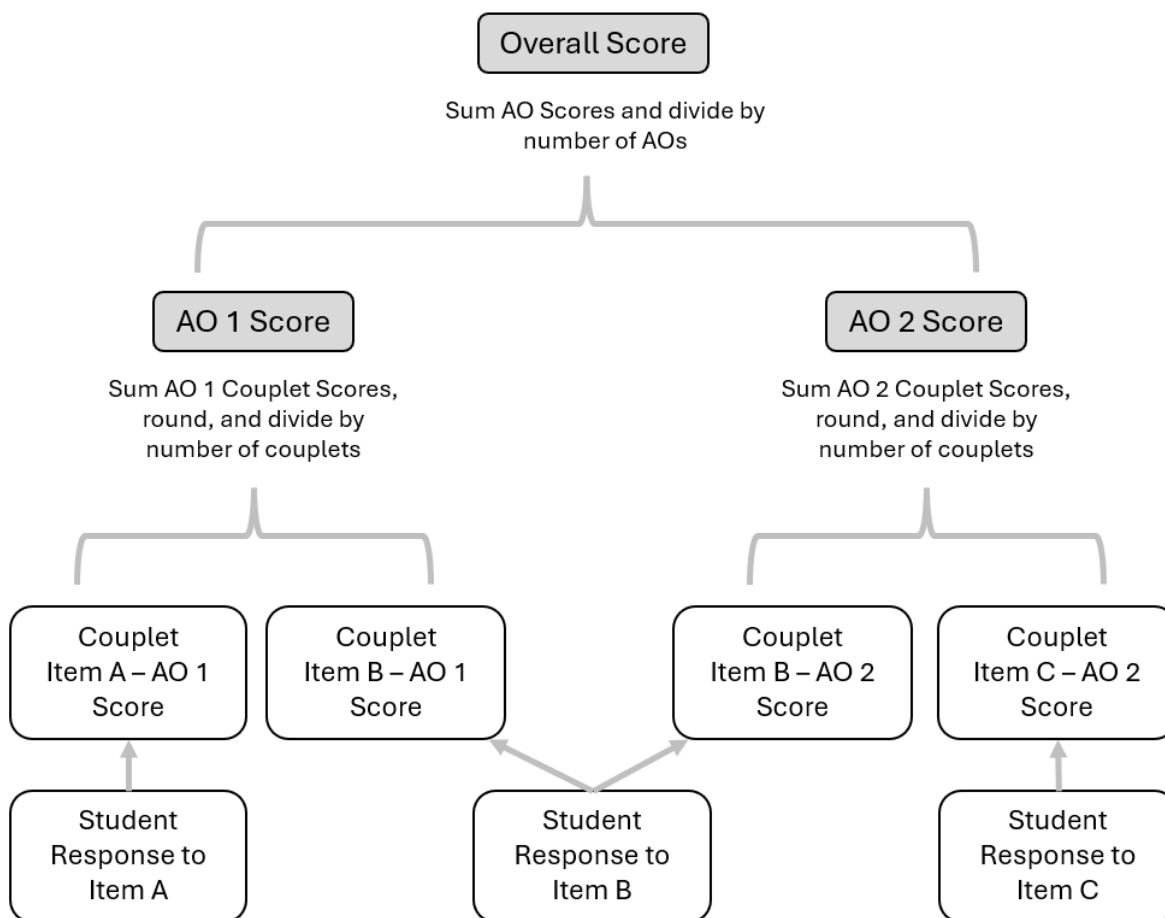
Figure 2.2: Flowchart showing how a student's responses to the model instrument items become couplet, AO, and finally overall scores. Generally,student response are scored along relevant AOs. These couplet scores are summed, rounded to the nearest integer, and normalized to 1 for each AO. Finally, these AO scores are summed and normalized to 1 in order to create an overall score that equally weights each AO. Gray shading indicates numbers reported to instructors and used in research analysis.

maximum of 1, giving an average of the AO scores (as opposed to an average of couplet scores), refleting a desire to equally weigh AOs. The average score for all students for each AO, as well as the average overall score, are all reported to instructors and can be used for answering research questions. While the overall score is reported to instructors to allow them to get a general sense of how well their students are doing on the broad concept of the topic covered by the RBAI (such as Newtonian mechanics or measurement uncertainty), the AO scores give more fine-grained and

actionable data on student performance.

The remainder of this section explores examples of couplet scoring from two RBAIs, SPRUCE and the FCI. Despite differences in these instruments, these examples show how specific items can be scored using a couplet scoring scheme to produce multiple couplet scores that contribute to various instrument AO scores.

2.4.2.2   Couplet scoring example - the Survey of Physics Reasoning on Uncertainty Concepts in Experiments

An example of a multiple-choice item scored using couplet scoring, item 3.3 from SPRUCE (shown in Fig. 2.3) tasks students with determining the period of oscillation for a mass hanging vertically from a spring. This item has two AOs, which we refer to as "3.3 H1" and "3.3 H2".

> You and your lab mates decide to measure 20 oscillations at a time. Using a handheld digital stopwatch, you measure a time of 28.42 seconds for 20 oscillations. You estimate the uncertainty in your measurement of 20 oscillations to be 0.4 seconds, based on an online search for human reaction time. What value and uncertainty do you report for the period of **a single oscillation**?
>
> ○ 1.421 ± 0.02 s   ○ 1.42 ± 0.02 s   ○ 1.4 ± 0.02 s
> ○ 1.421 ± 0.4 s   ○ 1.42 ± 0.4 s   ○ 1.4 ± 0.4 s

Figure 2.3: SPRUCE item 3.3 (with alternate numbers to protect test security), in which students are attempting to determine the period of oscillation for a mass hanging vertically from a spring. This item has two AOs, H1 and H2.

H1 Propagate uncertainties using formulas

H2 Report results with uncertainties with correct significant digits

The proficiencies represented by AOs H1 and H2 both impact the response a student would provide on item 3.3, but they are conceptually independent of one another: thus it is straightforward to

create an item and scoring scheme that independently determines if students correctly propagate uncertainty (H1) and if they report their value and uncertainty with correct significant digits (H2).

For couplet 3.3 H1, applying the appropriate uncertainty propagation formula simplifies to dividing the uncertainty in the time for 20 oscillations by 20, yielding a value of ± 0.02 s as the uncertainty in the period of a single oscillation. This value appears in three of the answer options, and so the selection of any of these three responses awards a full point for couplet 3.3 H1. Similarly, two of the six answer options are presented with appropriate numbers of significant digits for the value of the period (based on the value of the uncertainty) and thus receive full credit for couplet 3.3 H2.

Table 2.3: Example scoring for couplets of item 3.3

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

SPRUCE item 3.3's couplets may seem sufficiently independent that it would be straightforward to split the item into two independent items, one for each AO. However, such a separation will not meaningfully measure student proficiencies with H2 unless the response of the item targeting H2 is coordinated with the response of couplet 3.3 H1, as proper reporting requires coordinating between the significant digits of the result and the uncertainty. Additionally, splitting the item such that one item targets only H2 would likely involve prompting students to use proper significant digits (which SPRUCE does do with other couplets), rather than what the current item does, which is measure if students use proper significant digits without prompting. In these ways, couplet scoring allows these two AOs to be targeted in ways that two separate, traditionally scored items could not.

2.4.2.3   Couplet scoring example - the Force Concept Inventory

The process described in Sec. 2.5 for developing an instrument using the couplet scoring paradigm outlines how AOs and couplets impact nearly every stage of assessment development. While the previous example is from an instrument developed to employ a couplet scoring scheme, we believe at least some of the affordances of couplet scoring (described more fully in Sec. 2.4.3) can be illustrated by considering what a couplet scoring scheme for traditionally developed and scored items might entail. Specifically, we now consider item 18 from the FCI [107] and present a potential preliminary (i.e., un-piloted) couplet scoring scheme for this item. The purpose of this example is not necessarily to suggest that a couplet scoring scheme should be developed for the FCI (though ongoing work is investigating such a scoring scheme), rather, we aim to demonstrate how couplet scoring may (i) yield more actionable feedback than traditional item scoring; (ii) identify and inform potential modifications to items; and (iii) support or improve upon various instrument analyses. We also discuss the limitations of applying a couplet scoring scheme *post hoc* to an instrument designed with traditional scoring in mind. Importantly, many of the details of this section are intended to be exemplary and not prescriptive: the following depicts just one possible couplet scoring scheme for item 18 on the FCI.

The FCI has been extensively deployed and studied (e.g., [37, 70, 219]). To develop a couplet scoring scheme for item 18 on the FCI, we use Stewart et al.'s [219] principles as a starting point for developing our AOs. These principles are statements about the content to be assessed and were not intended to be used as AOs, however, one might reasonably interpret the principles as AOs by considering that the goal of the assessment is to measure student proficiencies regarding the ideas expressed in the principles.

Item 18 of the FCI, shown in Fig. 2.4, asks students to consider which of four possible forces are acting on a boy swinging on a rope. As there are four possible forces described in the item, there are theoretically $2^4 = 16$ possible responses, of which students are asked to choose between five.

18. The figure below shows a boy swinging on a rope, starting at a point higher than A.
    Consider the following distinct forces:
    1. A downward force of gravity.
    2. A force exerted by the rope pointing from A to O.
    3. A force in the direction of the boy's motion.
    4. A force pointing from O to A.

    Which of the above forces is (are) acting on the boy
    when he is at position A?
    (A) 1 only.
    (B) 1 and 2.
    (C) 1 and 3.
    (D) 1, 2, and 3.
    (E) 1, 3, and 4.
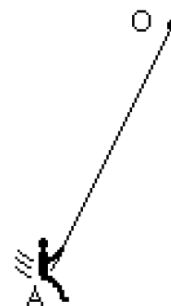
Figure 2.4: Item 18 from the Force Concept Inventory.

Stewart et al. categorized Item 18 as falling into all three broad categories: Kinematics, Dynamics, and Forces [219], and identified the following theoretical principles for this item:

C2 Objects moving in a curved trajectory will experience centripetal acceleration

L2 Newton's 2nd law

L4 Objects near the earth's surface experience a constant downward force/acceleration of gravity

F2 An object does not necessarily experience a force in the direction of motion

These theoretical principles remain associated with Item 18 in Stewart et al.'s optimal employed multidimensional item response theory (MIRT) model.

With these principles serving as our AOs, we consider each of the four forces described in the item independently in terms of each AO, with a potential numeric representation of how the selection of each force indicates proficiency with each AO shown in Tab. 2.4.

Values less than zero, equal to zero, between zero and one, and equal to one in Tab. 2.4 respectively represent incorrect, neutral, partially correct, and correct applications of the ideas expressed in the AOs. As can be seen in Tab. 2.4, for C2, 0 points are awarded for forces 1 and 3

Table 2.4: Independent scoring, by AO, of each force that is represented in the answer options of item 18 on the Force Concept Inventory. These scorings will contribute to the couplet scores for each answer option via a simple algorithm that sums the scores and then applies a floor of 0 and a ceiling of 1.

Force	Principle/AO			
	C2	L2	L4	F2
1 (gravity)	0	0.5	1	1
2 (radial in)	1	0.5	0	1
3 (along motion)	0	0	0	-10
4 (radial out)	-1	0	0	1

because they do not relate to centripetal acceleration, 1 point is awarded for force 2 because the forces is radial, and -1 point is awarded for force 4 because it is radial but not a force acting on the boy. For AO L2, half a point is awarded for each force 1 and 2 because, together, they comprise the net force acting on the boy. For AO L4, a point is awarded for force 1 (and no other forces) because force 1 is the force of gravity. Finally, for F2, a point is awarded for any force that does not point in the direction of motion, and -10 points are awarded for the selection of force 3 (to negate any points that could have been earned for this AO by selecting other forces).

These values are then fed into a simple algorithm that sums the score of each AO across the selected answer option and then applies a floor of 0 and a ceiling of 1 to that sum. In this way, a value of -10 will completely negate any positive values, ensuring a final score of 0 for that AO. We would again like to emphasize that this is not the only possible scoring system and algorithm that one could develop for tallying couplet scores for this or similar items.

Instead of having only 5 of the force combinations available, if students were allowed to select any or all of the 4 forces that they believed were acting on the boy, this algorithm could be applied directly to student responses. However, the FCI limits responses to 5 predetermined combinations of these 4 forces. Applying this algorithm to these 5 combinations of forces yields the (preliminary) couplet scores for this item that are shown in Tab. 2.5.

This scoring scheme produces five different results for the five different answer options, while traditional scoring for the FCI would produce a binary score, with the selection of any of the four

Table 2.5: Preliminary couplet scores for item 18 on the Force Concept Inventory.

	Principle/AO			
Selection	C2	L2	L4	F2
(A) 1 only.	0	0.5	1	1
(B) 1 and 2.	1	1	1	1
(C) 1 and 3.	0	0.5	1	0
(D) 1, 2, and 3.	1	1	1	0
(E) 1, 3, and 4.	0	0.5	1	0

distractors being equivalent in the final item score and overall score, which we believe represents a loss of potentially actionable information. It is notable that all of the available answer options earn a point for principle L4, as students are not allowed to select an option that does not include a downward gravitational force, and thus we note that this item, as written, cannot measure student proficiency with regards to L4. For this reason, L4 should be removed as an AO for this item, or else it will inflate L4's AO score for the FCI. This finding also suggests that Stewart et al's. MIRT analysis of the FCI might reasonably be modified to omit Item 18 from consideration for the factor containing L4.

While this scoring scheme has not been piloted in interviews to develop robust evidentiary arguments, we can, as an exercise, collapse the four couplet scores into a single normalized score for each answer option to approximate a partial-credit scoring scheme. Doing so creates a hierarchy of responses $B > D > A > E = C$ which is quite similar to Eaton et al.'s partial credit hierarchy of $B > D = A > E > C$ that was obtained from a two-parameter logistic nominal response model using polytomous item response theory [70]. While collapsing couplet scoring to a partial-credit model in this way is not aligned with the theoretical foundations of couplet scoring and would no longer be able to produce AO scores for the instrument as a whole, it does provide a convenient reasonableness check for the scoring schemes of individual items.

2.4.3 Couplet Scoring Affordances

We now describe the affordances of couplet scoring that we have observed in developing SPRUCE and creating a couplet scoring scheme for the FCI. These affordances, summarized in Tab. 2.6, benefit both instructors and education researchers and expand on the affordances provided by AOs discussed previously [236].

Table 2.6: Affordances of couplet scoring and related assessment design features

Affordance	Couplet-Scoring Design Feature
Item alignment with assessment priorities	Alignment is embedded into item development and scoring, as items are developed to be scored by AO, and verification of this alignment is supported by having concise and explicit AOs
Item alignment with instructional priorities	Instrument AOs (that inform and contextualize scoring) can be directly compared with course learning objectives
Authentic items	Scoring by AO allows for complex, authentic closed-form items with nuanced scoring
Scaffolded scoring	Developers creating a scoring scheme need only consider one AO at a time
Partial credit	Reduces need for partial credit scoring by resolving what is "partially" correct and incorrect into different couplets/AO scores
Data yield	Items with multiple AOs yield more data than items with just one AO
User experience	Indistinguishable from traditional instruments
Validation	Can use many traditional approaches with couplet scores as the unit of assessment
Reporting scores by AO	Scores are reported by AO in a manner that is clear and actionable

2.4.3.1   Item Alignment with Assessment Priorities

During instrument development, once items have been created, they undergo an iterative process of piloting and refining, which may change elements of the item such as the item context, the amount of information provided in the prompt, and the available answer options. Each of these changes has the potential to shift the item away from its intended objective. With couplet scoring, the scoring process itself requires that developers revisit the intended objective(s), rather than just consider if one of the answer options is correct and the rest incorrect. Thus, the chance that the final item shifts away from its intended objective(s) is minimized, and if the item does shift, then re-scoring it will require reconsideration of the item's AOs. This built-in reliance of scores on objectives supports alignment at all stages of assessment development, including even before piloting or other data have been collected.

In traditional item scoring, where the scoring schemes do not existentially depend on the item's objective(s), developers must be careful and take additional steps to ensure that the final products align with the intended assessment goals, or else the item may not provide a measure of the targeted and articulated proficiency.

Additionally, having the instrument constructs clearly articulated as AOs facilitates quick and effective verification of alignment between items and constructs by independent expert consultants [72, 103, 200], as was done with SPRUCE [235].

2.4.3.2   Item alignment with Instructional Priorities

Prior to using a content RBAI, instructors and researchers need to ensure the instrument aligns with the content they wish to assess. By having the instrument constructs articulated as AOs (that are used in item development and scoring), direct comparison between the instrument objectives and instructional learning objectives is possible: if an instructor finds that AOs for an assessment match their own learning objectives, then it is likely that assessment will be of use to them. This alignment, known as curricular validity, is "how well test items represent the objective

of the curriculum" [161], and is an important consideration of instructors looking to use RBAIs in their course.

For many instruments, a clear list of instrument constructs is not articulated. With other instruments, the constructs are listed only in academic articles and not presented to implementers alongside the instrument. As such, an implementer would need to either attempt to interpret the intent of the developer and/or review published academic articles detailing instrument development in order to establish curricular validity. With couplet scoring, these objectives are a central part of the instrument reporting, and are straightforward to present to implementers along with the instrument.

2.4.3.3   Authentic Items

Though not unique to physics, the synthesis of multiple ideas is often valued in physics instruction and evaluation. This means that items that more authentically depict interesting and relevant physics scenarios often incorporate multiple concepts as a reflection of the interconnected nature of physics [123] and thus may include both composite and simple constructs.

While traditional assessment approaches, by design, typically have assessment items related to only a single construct, couplet scoring allows for more interesting and appropriate assessment items that can be scored and interpreted to produce meaningful feedback to implementers.

As an example, in SPRUCE, the items are all contextualized within four experiments (such as a cart rolling down an inclined plane); this context may help students more easily understand the items without overwhelming them with a different physical context for each item. Couplet scoring then facilitates taking advantage of this design to ask rich, authentic questions about these experiments without the limitation that the items must target only a single construct.

2.4.3.4   Scaffolded Scoring

With couplet scoring, the development of a scoring scheme is scaffolded by considering which item responses indicate proficiency in a particular AO. This feature is especially important for scor-

ing schemes with more complex item types, such as multiple-response items and coupled multiple response items.

Anecdotally, when developing couplet scoring schemes for SPRUCE, several members of the research team expressed that this scaffolding made developing a scoring scheme for SPRUCE's coupled multiple-response items feel faster, easier, and less subjective than for coupled multiple response items developed during previous projects.

2.4.3.5   Partial Credit

For closed-form items, students select a response from a list of options that are generally made up of a correct answer and several tempting distractors. These distractors are most effective when they represent an answer that one would arrive at by employing mostly correct reasoning, but also a common misunderstanding or an easy mistake [72]. However, educators often wish to distinguish between different incorrect responses, such as between a response resulting from a simple mistake and one resulting from a fundamental misunderstanding or misapplication of a core concept. As a result, researchers will sometimes employ partial credit scoring schemes (e.g., Ref [70]).

In couplet scoring, each of the lines of reasoning one must employ to obtain the correct answer can often be represented by an AO, and so various distractors may be completely correct in terms of one AO while being incorrect in terms of others. As the item is scored by AO, it is possible for multiple responses to receive full credit for one AO, while not receiving credit for another. This can substantially reduce the need for partial credit, which requires arbitrary weights for partially correct responses. It also better captures and reports the elements of desired reasoning that students employ, since two mostly correct responses will result in meaningfully different couplet scores that do not get obscured by representing the measures of student reasoning with a single number.

For SPRUCE, couplet scoring eliminated the need for partial credit on all items except coupled multiple-response items. In instances where, if using item scoring, we might have considered awarding partial credit for a particular response, we instead were able to award full credit for the one AO and zero credit for another. For example, this can be seen by the couplet scores in Tab. 2.3

all being zero or one.

2.4.3.6   Data Yield

Items that align with more than one AO will have more than one couplet, and, as the couplet is the unit of assessment in couplet scoring, this means that using couplet scoring allows researchers and instructors to get more data from the same number of items, as compared to traditional item scoring. This feature can help to reduce the overall number of items in an instrument, making it easier for students to complete.

As an example, for SPRUCE, numeric-open-response items allowed us to evaluate students use of significant digits independent of the numeric value they chose to report, and we were able to do so without presenting students with additional items.

2.4.3.7   User Experience

As couplet scoring is essentially a back-end feature, the process of completing an RBAI that uses couplet scoring is virtually indistinguishable from completing an RBAI using traditional item scoring. The only perceivable difference for students might be that the items may be more complex than with traditionally scored RBAIs.

2.4.3.8   Validation

The effectiveness of an RBAI is largely contingent on how well it "measures what it says it measures," a property known as validity [72]. The investigation of various types of validity and their metrics is a major topic of scholarship [55, 72, 103, 141, 148, 167, 173, 200]. Commonly accepted methods for establishing the validity of an item generally include statistical approaches such as Classical Test Theory and Item Response Theory, though other measures of validity (including through consultation with content experts and the establishment of evidentiary arguments) also exist.

As couplets replace items as the unit of assessment, couplet scores replace item scores in statistical validation procedures. Many of the common statistical validation approaches can be easily adopted to work with couplets instead of items, and, for SPRUCE, this is the focus of Chapter 4. It is also worth noting that the validation of instruments using couplet scoring should primarily focus on ensuring that AO scores from couplets are meaningful measures of those constructs, and that common statistical metrics and thresholds may need to be reevaluated when used with couplet scores. Validation considerations for couplet scoring are discussed further in Sec. 2.5.4.

2.4.3.9   Reporting Scores by AO

Central to couplet scoring is the idea that couplets produce scores that are aligned with specific AOs. By reporting scores independently for each instrument AO, and by having the AOs be clear and concise statements about student proficiencies, we argue that the results of an instrument that uses couplet scoring could be better contextualized and more actionable for instructors than are single, overall assessment scores. While many instruments could report scores by objective, with couplet scoring, producing and presenting AOs scores is straightforward and also in line with the theoretical foundation of couplet scoring.

2.5   Developing and implementing an RBAI that uses couplet scoring

Now that we have introduced the concept of a couplet and described many of the affordances of couplet scoring, we now discuss details and limitations of implementing a couplet scoring scheme while designing a physics content RBAI. This section draws on primarily our experience developing SPRUCE, which employed and expanded on the assessment development framework of evidence-centered design (ECD) [169].

The five layers of assessment development as defined by ECD and modified for couplet scoring, are:

- *Domain Analysis*: the collection of information about the topic to be assessed from texts,

research literature, interviews with experts, etc.

- *Domain Model*: the distillation of information from the *domain analysis* into AOs and potential item contexts, including detailing acceptable evidence of proficiencies and the methods for gathering this evidence.

- *Conceptual Assessment Framework*: the determination of appropriate and desirable assessment features and item formats to support gathering evidence of student proficiencies based on the domain model.

- *Assessment Implementation*: the writing and piloting of items (and couplets), and the revising of items, AOs, and couplets, to establish evidentiary arguments linking student responses to student reasoning.

- *Assessment Delivery*: the implementation of the finalized items, scoring scheme, and feedback for implementers.

Many of these layers are similar to steps described in other assessment development frameworks [20, 141]. In the following sections, we discuss these steps and important considerations for assessment developers aiming to implement a couplet-scoring scheme while employing their choice of assessment development frameworks.

2.5.1 Developing Assessment Objectives

The process of collecting information on the topic to be assessed and processing that information into objectives and proto-items (the *domain analysis* and *domain model* in ECD) begins similar to any effort to determine the priorities and objectives of an assessment. Such efforts include consulting the education research literature and commonly used textbooks, identifying and reviewing existing assessments on similar topics, and soliciting priorities and feedback from instructors and other content specialists.

However, as the name "assessment objective" suggests, the distillation of this information into constructs should be expressed in terms of desired student proficiency, not just the name

of topic to be addressed. For example, SPRUCE contains the AO "S2 - Identify actions that might improve precision." Articulated in this way, AO S2 describes a measurable objective (much like a course learning objective [24]), as opposed to just stating that the instrument assesses the construct of "precision," which is ambiguous as to what specific knowledge or skills around precision will be assessed. As discussed in the previous section, these AOs can have conceptual overlap and need not be wholly independent, since they will be evaluated and reported independent of one another. AOs may be added, removed, split, consolidated, or otherwise refined throughout instrument development, as long as they continue to align with the information gathered in the first stages of development (i.e., in the *domain analysis* for ECD).

2.5.2   How to develop items with couplet scoring

The process of creating multi-construct items that align with the instrument AOs involves many of the steps of traditional item creation. However, these steps are necessarily iterative for instruments using couplet scoring, which is not always the case for instruments using a traditional scoring scheme. For example, as the development of items and a couplet-scoring scheme will likely necessitate expanding, narrowing, splitting, consolidating, or otherwise changing specific AOs, other items that target those AOs will need to be revisited to ensure they still align with the updated AOs. If significant enough modifications are made to items, additional piloting may be appropriate.

Initially, in the first stage of instrument development (the analysis of the topic to be assessed), the assessment priorities of instructors and other experts are being assembled into AOs. At the same time, instrument developers can begin to gather ideas that will inform the creation of specific tasks, as well as logistical considerations that may inform the types of items that are viable and reasonable for these tasks. These steps, for traditionally scored items, are described in many assessment development frameworks, such as in layers one through three of ECD [169].

In the next steps of instrument development, when items are being crafted and before they have been piloted, development for instruments using couplet scoring will differ from more traditional instruments in that the items can reflect a level of complexity representative of instruction,

as they can address more than one concept. For closed-form items, initial distractors can be developed by considering the responses that respondents might provide if they were to employ correct reasoning along some of the item's AOs, but incorrect reasoning along other, and the scoring of an item with such distractors should vary between different couplets of the same item.

Once the items have been drafted, the process of item refinement continues much like that of traditionally scored items: an iterative process of piloting the items and refining the item prompts, answer options, and scoring. However, if it is found that a particular couplet is inappropriate or too difficult to assess, it may be possible to remove that particular couplet from the instrument without discarding the entire item. This happened with SPRUCE where, for example, a task asking students to report a best estimate of a value based on a set of repeated measurements dropped a couplet relating to removing outliers, but the item remained in SPRUCE because other couplets for this item were still viable.

2.5.3   Scoring Couplets

For single-AO items (i.e., items with just one couplet), the process of developing a scoring scheme is much the same as with traditional items, with the added benefit of having an explicit AO guiding scoring to help ensure that the item measures what it was intended to measure.

It can be tricky initially for content experts, who are used to coordinating many different ideas at once, to look at a multi-AO item and consider how each response relates to only one AO at a time. Such compartmentalization is relatively straightforward for multiple-choice items with AOs that have no conceptual overlap, such as with SPRUCE item 3.3 and the scoring scheme shown in Tab. 2.3, however, it can be more challenging for multiple-response or coupled-multiple-response item formats and AOs that are more closely related. Fortunately, having the instrument and item constructs clearly articulated as objectives provides an easy reference for developers, and the process of scoring by AO can quickly become intuitive. In fact, for SPRUCE, the developers ultimately found the process of scoring by AO made scoring easier for complex items types such as coupled-multiple-response items, as the AOs themselves productively narrowed the idea-space the

developers were considering at any one time.

Additionally, couplet scoring lends itself to an intuitive check of the scoring scheme early on in item development, as demonstrated with the example scoring scheme from item 18 on the FCI discussed in Sec. 2.4.2. This approach involves considering a (temporary) "collapse" of the scoring scheme into a partial credit scheme to gauge possible item responses in terms of the fraction of the total points—across AOs—they would receive: a response aligning with only one of three AOs would receive a third of the total points for the item, for example. Developers can then determine if the relative scores of various answer options seem reasonable, and such checks can help identify potential issues with items, which may be especially useful before piloting with students has occurred.

2.5.4   Statistical Validation

As discussed in Sec. 2.4.3, by replacing items with couplets as the unit of assessment, instruments employing couplet-scoring schemes can use many common statistical validation approaches. However, just as a considering a 'total item score' for a multi-AO item can serve as a check to identify large issues in the scoring of couplets for an item, validation metrics that use an overall assessment score can be used as a check to identify large issues in an instrument. As couplet scoring is designed to score and report proficiencies on an AO-by-AO basis, it may not be reasonable to expect that metrics based on a total score necessarily adhere to the conventional thresholds of single-construct items and instruments. For example, pure guessing on SPRUCE item 3.3 (Fig. 2.3) would result in an average score of 50% for AO H1 (Tab. 2.3), which is higher than what is generally desired for item scores with ~ 4 answer choices [55]. However, it is important to keep in mind that this item also serves as a measure of AO H2, and that the AO score for AO H1 also takes into account other items, contributing to an overall more reliable measure of the AO.

Additionally, CTT metrics may be calculated using traditional methods by simply replacing items with couplets. The calculations remain the same, although the interpretations may differ slightly. For example, Doran suggests that difficulty should be between 0.30 and 0.90 for each item [62]. However, this assumes each item has a four-option multiple-choice format, where a

difficulty of 0.25 would represent random guessing. Clearly, for some couplets, this assumption no longer holds — for example, in the previously mentioned SPRUCE item 3.3 example, for AO H1, 50% is random guessing.

Similarly, Englehardt [72] discusses that difficulty per item should be about 50% to obtain the best discrimination for each item. Again, this interpretation is not necessarily true when couplet scoring is used and multiple answer options can get full credit. Discrimination should be calculated for each individual couplet, and there may not be the typical correlation with difficulty that one might expect. This is because when using couplet scoring, multiple answers in a multiple-choice question might be scored as correct, so a 50% difficulty score might not lead to the most ideal discrimination.

Fortunately, the raw numbers for other forms of CTT validation still follow the usual conventions, such as interpretations of discrimination, and reliability in the forms of test-retest stability, as well as internal consistency (Cronbach's alpha).

In addition to performing a CTT validation of the entire test (using statistics such as Cronbach's alpha and Ferguson's delta), as well as on a couplet-by-couplet basis (using statistics such as couplet difficulty and couplet discrimination), some CTT statistics can be applied on an AO basis. For example, one could examine the Cronbach's alpha within a particular AO to determine whether all of the items assigned to that AO are consistent with one another. Additionally, we can calculate difficulty and discrimination at the AO level, answering questions about how difficult certain concepts are for students, as well as how well specific AOs discriminate between high and low performers. Instead of performing CTT on only a couplet and entire assessment level, calculating CTT statistics on AOs allows for an intermediate level of validation, in that they are finer-grained than an overall score but coarser-grained than individual couplet scores.

2.5.5 Instructor Reports

Instruments that employ couplet scoring will produce a score for each of the instruments AOs, the mean score of several couplets that all target the same AO. These AO scores can be

presented to instructors with minimal elaboration, as long as the AOs are clearly written. Thus, AO scores should provide specific and actionable feedback for instructors compared to instruments that use traditional item scores to present instructors with a single number and/or a number for each item. When the instrument is used in a pre-instruction then post-instruction modality, the instrument user can see how proficiency with each AO changes between the beginning and end of their course.

We have previously presented an example figure from an instructor report for SPRUCE that highlights how AO scores are the primary mode of conveying student proficiencies to instructors for an instrument that uses a couplet scoring scheme [235].

2.6   Implications for Other Assessment and Evaluation

While couplet scoring was developed in parallel with SPRUCE, a physics content RBAI, we believe this scoring model can be employed in a variety of assessment settings, including in other fields and in formative and summative assessment within a course. Alignment between assessment items and objectives, and having scores reported by objective, has been heralded as a valuable aspect of course assessment [24, 160, 161]. While content RBAIs are not intended to provide instructors with feedback about individual students, instructors may find some of the details of couplet scoring discussed in this chapter helpful in efforts to ensure that course instruction and evaluation are truly aligned. Additionally, our descriptions of scoring and reporting by objective may prove useful for instructors who are interested in providing students scores or feedback in terms of specific course objectives for other assignments in the course.

2.7   Summary

In this chapter, we introduced a new scoring paradigm, couplet scoring, in which each instrument item is scored potentially multiple times, once for each of the assessment objectives (AOs) that the item aims to measure. We explored how couplet scoring and the use of AOs produce meaningful measures of student proficiency. We then discussed some of the nuances and challenges

of implementing a couplet-scoring scheme for a research-based assessment instrument (RBAI), using examples from our work with both developing the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) [235] and exploring a *post hoc* couplet-scoring scheme for the FCI. Finally, we discussed how couplet scoring might inform physics assessment outside of formal content RBAIs.

Related future work includes investigating a full *post hoc* couplet scoring scheme for the FCI and a statistical validation of SPRUCE that uses couplet scores and AO scores, rather than item scores, as the statistical units for validation.

Chapter 3

Survey of physics reasoning on uncertainty concepts in experiments: the development of an assessment of measurement uncertainty for introductory physics labs

3.1   Contribution

This chapter is adapted from Vignal, et. al., 2023 [235].

I joined the team working on SPRUCE in the middle of the development phase of the assessment. When I began working on this project, a set of items grounded in four distinct laboratory experiments had been developed and piloted through one round of student interviews and one round of online test administration (see Interviews 1 and Beta 1 in Table 3.5 in the text of this chapter). An initial draft of the Assessment Objectives was already in place when I joined the project. Much of my work on this project began with various beta data, which is detailed more clearly throughout this chapter.

I began work with the Beta 1 data, which had not yet been analyzed at all. Through the analysis of these data, I helped determine potential problem items, including a revision of multiple items on the assessment, as well as the addition of several items. In addition, I participated in interviewing students during Interviews 3 and analyzing these interview data to implement further revisions to the assessment items. Additionally, through these interviews, I helped to refine our Assessment Objectives.

Further, I proposed implementing Beta 2 testing as open response to ensure that our distractors covered all possible common student responses. I then analyzed the Beta 2 data to corroborate

our distractors.

I also analyzed the data received from Beta 3 piloting including some qualitative coding of open response items for several hundred student responses in order to transform multiple choice + open response items into multiple choice + multiple response (also known as coupled multiple response).

I further helped develop the scoring scheme (couplet scoring), which had not been explored before I joined the project. Scoring is more heavily discussed in Chapter 2 of this dissertation as a novel scoring method was developed by myself and Michael Vignal in scoring SPRUCE.

I also created the instructor reports that are shared with all instructors that participate in SPRUCE to help guide them in their students' understanding of measurement uncertainty. I wrote the initial draft of the piloting section of this chapter and created Figures 3.3 and 3.5.

Finally, the factor analysis section does not appear in the version of this chapter which is currently published as a paper in Physical Review Physics Education Research; I performed this analysis and wrote this section for this dissertation.

3.2 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 [184]) to teach basic MU techniques in introductory science labs. In experimental physics, MU informs comparisons of multiple measurements [188] or between measurements and values predicted by models [64], and so instruction around MU can help students better understand the nature of experimentation. This and other important features of MU has led to policies and recommendations to include MU in introductory science courses [14, 138]. 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 [24]—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 (RBAIs), which Madsen, McKagan, and Sayre define as "an assessment that is developed based on research into student thinking for use by the wider...education community to provide a standardized assessment of teaching and learning" [153]. It is important to note that RBAIs 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 RBAIs often employ a theoretical framework during assessment development, such as Evidence Centered Design (ECD) [169], the Three-Dimensional Learning Assessment Protocol [141], or the framework described by Adams and Wieman [20]. Such frameworks "facilitate communication, coherence, and efficiency in assessment design and task creation" [169], 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 on the assessment: these claims are grounded in evidentiary arguments (a major focus of this work) and contribute to the validity of the assessment instrument.

Of the RBAIs 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 point-like and set-like reasoning paradigms [45] and in measuring the success of course transformations aimed at helping students shift towards more set-like reasoning [146, 183, 186, 188]; the Physics Lab Inventory of Critical Thinking (PLIC) [241] has been used to assess the effectiveness of a scaffold and fade approach to teaching critical thinking in a physics course [116]; and the Concise Data Processing Assessment (CDPA) [58] has been used to identify changes in student

performance around MU [135] and to look at student performance across genders [59].

While each of these assessments deals with MU in some way, there is not currently a widely-administrable RBAI 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 [169]. 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 work, we present SPRUCE's assessment questions (hereafter referred to as "assessment items" or simply "items") and discuss their development. The goals of this work are to demonstrate:

(1) A need for a widely-administrable assessment of measurement uncertainty and how SPRUCE will satisfy that need,

(2) The assessment item development and refinement process, as guided by ECD;

(3) 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

(4) An example of feedback for instructors and how that feedback might be interpreted.

We begin by discussing, in Sec. 3.3, the need for a new MU RBAI and how the framework of ECD can facilitate the development of such an assessment. In Sec. 3.4, 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 is discussed in Sec. 3.5. In Secs. 3.6 and 3.7, we briefly discuss components of validity and how instructors might interpret and use the results of the instrument. In the final section, Sec. 3.9, we summarize the work discussed in this work and provide information for instructors and researchers who may

be interested in using SPRUCE.

3.3   Background

Over the last 30 years, research-based assessment instruments (RBAIs) have been used in physics classrooms to probe areas of interest and import for physics education researchers and physics instructors. Particularly notable examples of RBAI use include Mazur's use of assessment questions from the Force Concept Inventory [107] to probe student understanding of Newton's third law in his introductory physics lecture at Harvard [160], which led him (and others) to rethink what instruction in a lecture setting should look like [80] and Eblen-Zayas' use of the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS [270]) 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 [71].

More broadly, RBAIs have been developed and deployed in the areas of mechanics [107, 230], electrostatics [155, 259], quantum mechanics [203], and thermodynamics [195]. In addition, assessments have been used to probe quantitative reasoning [248], beliefs about physics and physics courses [19], experimental research and lab courses [270], modeling [65, 202], and concepts and practices used in laboratory courses [45, 58, 241]. These and other assessments can be found on the PhysPort website [11], and many of them are also accessible on Learning About STEM Student Outcomes (LASSO) [10].

In this section, we provide more detail about RBAIs 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) [168] informed this work.

3.3.1   Research-Based Assessment Instruments in Physics Labs

The following sections discuss three existing RBAIs that include some assessment of MU topics. While each of these RBAIs 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.

3.3.1.1   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 [165] for use with students at the University of Cape Town, South Africa [45]. 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" [45].

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 point-like, indicating students believe that quantities measured have a true value that can be obtained with a single, perfect measurement; set-like, a typically more expert-like 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 point-like and set-like 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 [188].

3.3.1.2 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" [11]. 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 [241]. 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 [116]. This approach involves 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 [116].

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 [241].

3.3.1.3   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 [58]. 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 [135]. While there were not enough matched pre- and post-instruction data to make comparisons of improvement on these two assessments, no correlation was found between CDPA scores and E-CLASS pre-instruction 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 pre- to post- assessment (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.

3.3.2 Assessment Development Framework: Evidence Centered Design

To help guide the development of SPRUCE, we employed the assessment development framework of Evidence Centered Design (ECD) [169] 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 work, 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" [169]. 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 [195,236]), 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*: finalized implementation of assessment, scoring scheme, and instructor reports.

The first layer of ECD, *domain analysis*, is the topic of a previous paper [184] 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 only briefly discussed in this work.

3.4   SPRUCE Development

3.4.1   *Domain Analysis*

The first steps towards developing an RBAI on MU were presented in a previous paper [184]. In that work, we conducted and analyzed interviews with 22 physics lab instructors at institutions that spanned a range of sizes, 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 semi-structured 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.

3.4.2   *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" [169]. 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 assess-

ment arguments in terms of assessment objectives [195]. Assessment objectives (AOs) are "*concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment*" [236]: essentially, the AOs are the instrument's constructs. AOs are similar in concept and grain size to learning objectives [24], but they are designed to "span the space of feasible, testable outcomes" of an assessment [195]. As discussed in [236], AOs also provide a number of additional benefits for assessment development beyond organization of ideas collected in the domain analysis.[1]

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 was developed in Ref. [184] 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 [193]:

- 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 [270], that remain as integral parts of the other categories. Additionally, as described in [236] and discussed further in Sec. 3.5, some individual AOs in the other categories

---

[1] In [236], 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*.

were also removed because of difficulties in establishing clear evidence of student reasoning. The finalized AOs are presented in Tab. 3.1.[2]

Table 3.1: SPRUCE assessment objectives, organized by assessment objective category.

	Sources of Uncertainty
S1	Estimate size of random/statistical uncertainty by considering instrument precision
S2	Identify actions that might improve precision
S3	Identify actions that might improve accuracy
	Handling of Uncertainty
H1	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

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

---

[2] Using the AOs outlined in Tab. 3.1, 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).

in Sec. 3.5.1.

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.

3.4.3 *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.

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 Tab. 3.2 and described in more detail in Sec. 3.5.3.

Table 3.2: 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 uncertainty of a single measurement, make comparisons between results, and answer questions about taking many more measurements.

To develop an assessment that is easy to administer to a large number of students—twice, as SPRUCE is intended to be used pre-instruction and post-instruction—we opted for an online format for the assessment [233, 261] 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" [72]) 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 evaluating 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) items that have a MC part followed by a MR part, and coupled-numeric-open-response (CNOR) items that have two NOR parts. These item types are examples of two-tier questions [231] in which, rather than considering the answer options selected in either question independently, it is the combination of selections from the coupled questions that is evaluated.

For SPRUCE's CMR items, the multiple-response answer options are *reasoning elements* that allow students to compose a justification to their response to the multiple-choice question, a design used in other physics assessments [181,196,259] 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 [263]). 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 pre-instruction assessment and

suggest or auto-fill them during the post-instruction assessment, and so quantities that factor into student responses on NOR and CNOR items are slightly different between pre- and post-instruction versions of the assessment.

3.4.4   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 Chapter 2, 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. 3.1) is presented in Tab. 3.3. 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 information given in the prompt about a measurement of the time it takes the mass to complete 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 3.3. 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. 3.7.

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

○ 1.421 ± 0.02 s   ○ 1.42 ± 0.02 s   ○ 1.4 ± 0.02 s
○ 1.421 ± 0.4 s   ○ 1.42 ± 0.4 s   ○ 1.4 ± 0.4 s

Figure 3.1: 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 3.3.

Table 3.3: 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 ± 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

3.5   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*.

3.5.1   Evidentiary Arguments

As stated above, the key focus of this work, and a key component of ECD, is the establishment of evidentiary 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, [188] and researcher expertise and experience also informed these arguments.

Data from pilot testing (discussed in the next section) was 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 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 3.2) in Tab. 3.4. 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 established evidentiary arguments.

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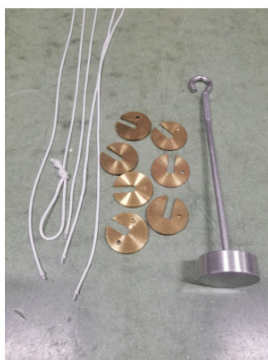
Your lab instructor asks the class to find the "breaking mass," m_{breaking}, for the string. They describe m_{breaking} by saying: "The string can support m_{breaking}, but the string will break if you add even a grain of sand more than m_{breaking}."

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_{breaking}?
  ◯ 520 g   ◯ 521 g   ◯ 530 g   ◯ 539 g   ◯ 540 g

What uncertainty do you report for your best estimate of m_{breaking}?
  ◯ 0 g   ◯ 1 g   ◯ 5 g   ◯ 10 g   ◯ 19 g   ◯ 20 g

Figure 3.2: Item 4.1 went through iterations informed by multiple rounds of pilot testing with students.

3.5.2 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 are 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

Table 3.4: Example Evidentiary Arguments for item 4.1. The top and bottom halves of the table includes 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 non-zero 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	Non-Inclusively 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."

additional information about student reasoning and student responses. These formatting decisions, as well as other priorities of the various pilots, are described in Tab. 3.5.

Table 3.5: 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 Pre-instruction Implementation Expand Evidentiary Arguments Refine Scoring Scheme	1048
Interviews 3	(Primarily Final Item Formats) Finalize Evidentiary Arguments	27

Even with fairly robust evidentiary arguments (as exemplified in Tab. 3.4) 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 [184] 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.

Information about these courses and numbers of student participants is shown in Tab. 3.6 and student demographics are shown in Tab. 3.7.

Table 3.6: 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 US. 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. We provide information about the three interview stages (Int.) as well as Beta testing.

Inst. Num.	Inst. Type	Course Num.	Int. 1 N	Int. 2 N	Int. 3 N	Beta 1		Beta 2		Beta 3	
						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 %

3.5.2.1   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

Table 3.7: 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≈1970)
Gender		
Man	51%	59%
Woman	41%	39%
Non-Binary	8 %	2 %
Not Listed	0 %	1 %
Race or Ethnicity		
White	72%	75%
Asian	23%	16%
Hispanic/Latino	10%	10%
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 %

refinements, and to begin developing evidentiary arguments.

Through course instructors, we solicited interview participants who had completed a 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 screen-sharing with the interviewer via Zoom. Interviews lasted between 30 minutes and 1 hour 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.

3.5.2.2   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 minutes 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 Tab. 3.6, where we report the number and percentage only of students who consented to allowing 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 [257] 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 [72]. 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. 3.5.3, and a full CTT analysis is presented in Chapter 4.

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
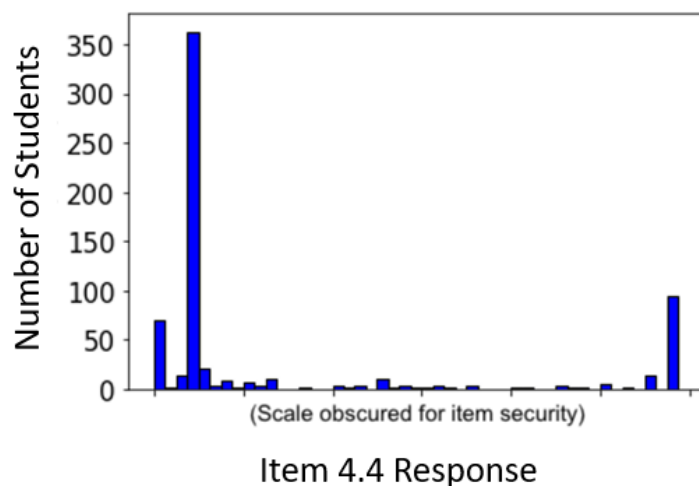
Figure 3.3: Histogram showing students' reported uncertainty values for item 4.4 on beta 1. These responses shows peaks at the correct value and at our planned distractors.

a response in a text box.

However, one item, item 4.4, was presented in a NOR format despite being designed to be a MC item. Student responses to this item are shown in Fig. 3.3. 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.

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.

3.5.3 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 Tab. 3.2 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 are clear and easily understood.

3.5.3.1 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.

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 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.

3.5.3.2 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 Tab. 3.8.

Table 3.8: 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	Student 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

Item 1.1 (shown in Fig. 3.4) is largely modeled after the "Repeating Distance" item from the PMQ [45]. 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. 3.4), 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 [45, 188] and refined based on interviews 2 and 3 and

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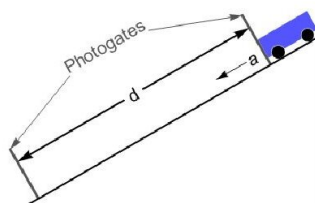
To measure a, you release the cart from rest and measure the time, t, it takes to travel the distance, d, between the two photogates. You can then calculate a using the formula:

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

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

Figure 3.4: Item 1.1 asks students what they would do after taking a single measurement for time, then asks students to support that choice.

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.

3.5.3.3 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 Tab. 3.9.

Table 3.9: Experiment 2 item types, descriptions, and AOs.

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

*H3 was eventually removed from item 3.2 due to our inability to establish clear evidentiary arguments.

For item 2.2, interviews revealed that many students were selecting the correct answer of "standard error (also known as 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." 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.

3.5.3.4   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 Tab. 3.10.

Table 3.10: 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 10 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

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 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., [66, 92, 131, 199, 237], as well as in other STEM fields and more generally [53, 101, 178]), and these items provide researchers and instructors an opportunity to observe the impact of representation on student reasoning around comparing data.

3.5.3.5 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. 3.2, and the specific experiment tasks are summarized in Tab. 3.11. 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. 3.2) 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 g 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 g and 19 g for the uncertainty estimate. The evidentiary arguments for this item are presented in detail in Tab. 3.4.

Item 4.2 (also shown in Fig. 3.2) was initially developed, in part, to address an AO of identifying and removing outliers, as one of 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

Table 3.11: Experiment 4 item types, descriptions, and AOs.

Item	Type	Description	AOs
4.1	CMC	Students are asked to identify $m_{breaking}$ and the uncertainty for a single measurement.	S1
4.2	NOR	Students are asked to report a value for the breaking mass based on 10 measurements.	D3
4.3	MR	Students are asked to compare their value (and an 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

notice the outlier, 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.

3.6 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 [23,58,72,103,148,173,200]. 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 work. Table 3.12 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 presented in Chapter 4.

Table 3.12: 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. Further evaluated in Chapter 4 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 [184] and Tables 3.6 and 3.7.	Established in Chapter 4 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.	

As outlined in Tab. 3.12, 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 Tab. 3.12 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 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 [241], and, eventually, item response theory (IRT) [265] or multidimensional IRT [219].

3.7 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 SPRUCE so far. The main graphic from such a report is shown in Fig. 3.5. The graph represents pre- and post-instruction 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 non-significant 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 [147]. This analysis produced similar qualitative effect sizes (small, medium, and large); thus, we report Cohen's D for simplicity. Not shown in Fig. 3.5 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 [261], and will be refined as feedback from instructors whom implement SPRUCE continues to be collected.
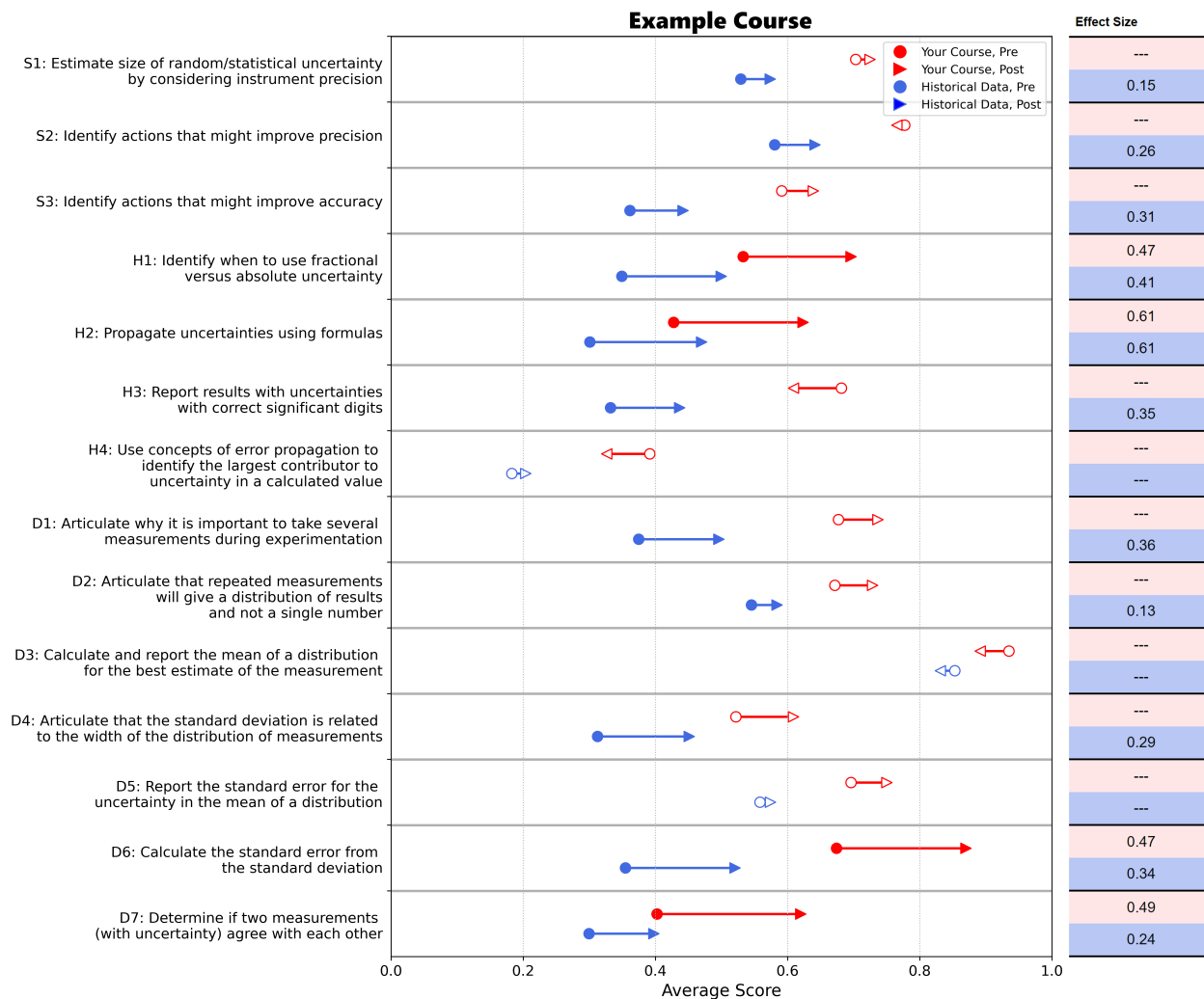
Figure 3.5: A portion of an instructor report showing pre- and post-instruction scores for the course and all historic data. Statistically significant shifts (as determined by a Mann-Whitney U test) for each AO are shown with solid circles and non-significant shifts are shown with open circles. Effect sizes for the statistically significant items are calculated using Cohen's d and shown on the right side of the chart.

For the course represented in this report, one can identify several important features. First, there are four AOs that show statically significant shifts from pre to post. Those include: (1) "H1 - Identify when to use fractional verses 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 post scores higher than pre scores) 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 casual 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.

3.8   Factor Analysis

Another potential avenue of analysis for SPRUCE data is factor analysis [215]. Factor analysis is a data reduction technique that partitions variance in the data into common variance and specific (or unique) variance. Common variance is shared amongst a set of indicators, which can be grouped

together into factors (in the case of SPRUCE, the indicators are the individual couplets). One major assumption of factor analysis is that this common variance is an effect of the underlying factors themselves. The goal of factor analysis is to identify and interpret a smaller number of factors that explain most of the common variance by assuming there are some latent variables that can't be directly measured but can instead be explored through the relationships they cause in the variables we can measure. On the other hand, the specific variance is not explained by the common factors but is instead due to characteristics of individual indicators [128]. Thus, this allows us to explore the different dimensions that might be present in the data. In the case of SPRUCE, we examine how the different indicators (couplets) might be combined in various ways to explain the common variance in the data. We use post-test data only to perform this analysis, including 1,923 total responses from the Spring23 and Fall23 semesters, since it is desirable for students to have instruction in these areas before trying to group responses.

There are two main methods of performing factor analysis – exploratory and confirmatory. In exploratory factor analysis, the latent variables are not set in advance, but rather determined by the data itself. In confirmatory factor analysis, the researcher sets the latent variables and analyzes how well the results match with these expected variables [128, 201].

In our case, we use factor analysis to try to reduce the dimensionality of our data from all of the AO-item couplets on the assessment to a few factors that might explain the results on these couplets, as described in detail by Ding and Beichner [61]. Our initial hope was that these factors would align with the AOs and provide further evidence for our scheme. We choose factor analysis over principal component analysis (PCA) because, despite being similar, PCA doesn't allow for causal relationships between the the components and observed variables, while factor analysis does allow for this [61], and we believe this therefore makes factor analysis the correct technique for our purposes.

Assuming we have an assessment with i AO-item couplets (or items, in a more conventional assessment) where we hope to determine j common factors, we can solve equations such as:

$$C_i = b_{i1}F_1 + b_{i2}F_2 + ... + b_{ij}F_j + U_i, \hspace{3cm} (3.1)$$

where we define the couplets (the observed variables) C_i in terms of regression coefficients or weights b, common factors F_j, and a unique factor for each couplet, U_i. Factor analysis calculates optimal weights by solving eigenvalue equations for the observed correlation matrix. The diagonals of this correlation matrix will be the variances that are accounted for by common factors - thus, the portion of the variance that is explained by the unique factors is not used in this analysis. After the factor analysis is done, the loadings of each couplet onto each factor are analyzed [61].

I started first with an exploratory factor analysis, where the underlying factors themselves are not assumed (though the number of underlying factors is assumed). Because the AOs are all related in many ways, and all are sub-components of measurement uncertainty, we decided to choose an oblique rotation rather than an orthogonal one. An oblique rotation allows the factors to correlate with each other, whereas an orthogonal rotation does not allow for such correlation [52,96]. In general, rotations can be imagined as multiplication of the data with some rotation matrix. If we imagine reducing our data to two factors, we can imagine these factors as x and y axes. An orthogonal rotation will rotate the usual x and y axes to some new positions that are still perpendicular to one another, while an oblique rotation will rotate these to some new positions that are not perpendicular to one another [17]. Then, the loadings of the variables on the factors is their projection onto these axes. Ideally, each variable will load heavily onto only one of these axes, and the axes themselves will be appropriate for all of the variables. Examples of orthogonal rotations include varimax, oblimax, and quartimax [17]. Examples of oblique rotations include promax, oblimin, and quartimin [96].

Specifically, a promax rotation was applied in this analysis. This is one of the more common oblique rotation methods, and it uses an orthogonal solution as a starting point; this orthogonal solution is then altered in specific ways to give an ideal oblique solution. [17, 96]. It generally assumes a varimax orthogonal rotation as the initial solution. Varimax rotations minimizes the

number of variables that have high loadings on each factor, which thus simplifies the interpretation of factors [96,119]. In other words, the solution means that each factor has a small number of large loadings and large number of very small (or zero) loadings, thus simplifying interpretation as each original variable tends to be associated with only one or few factors and each factor represents only a small number of variables [17]. This is the most popular orthogonal rotation technique and is the starting point for the oblique promax rotation.

Then, to make a solution that is better than the one given by this initial orthogonal rotation, it assumes that the mid- to low-level loadings need to be lower in the oblique solution than the orthogonal solution, while the high loadings must remain high. This is only possible if the factors are allowed to be oblique, rather than orthogonal. To find these oblique factors, we take advantage of the fact that the loadings are between -1 and 1: if these loadings are raised to any power greater than 1, they become smaller (ex., $0.9^2 = 0.81 < 0.9$). Further, raising these loadings to a power also emphasizes the difference between small and large loadings (ex., $0.3^2 = 0.09$ and $0.9^2 = 0.81$, and while $0.9 > 0.3$, $0.81 \gg 0.09$). If we increase the power to which the loadings are raised (ex., 3 instead of 2), these differences become further emphasized. The mid- and low-level loadings thus approach zero quickly as the power they are raised to increases. By varying the power, the degree of obliqueness can be varied as well: higher powers lead to more obliqueness (meaning, more overlap between factors) because they reduce the smaller loadings to a greater extent. The correct power to choose is the one that gives the simplest structure with the least correlation between factors. Typical powers to try are 2, 4, and 6, while ensuring that the signs of the loadings are restored after raising them to these powers such that negative loadings remain negative [96,105]. Then, to find the factors from this new matrix, we determine a new matrix Λ:

$$\Lambda = \left(F'F\right)^{-1} F'PD, \tag{3.2}$$

where Λ is the transformation matrix, F is the orthogonally rotated loadings matrix, P is the matrix obtained from raising the loadings matrix F to some power, and D is the diagonal

matrix which normalizes the product $(F'F)^{-1}F'P$. The matrix Λ is the transformation matrix from the orthogonal factors to the oblique reference vectors. Once Λ is obtained, it can be used to determine the correlations between factors as well as the primary loadings of the data onto these new correlated factors:

$$F\Lambda = V_P, \tag{3.3}$$

where V_P are the desired promax loadings [57]. As a simple example, consider an assessment with ten items and three factors. Then F is the 3×10 matrix of loadings determined via the varimax method (where the columns are factors and the rows are items) and F' is the transposed F matrix and is therefore 10×3. P is the matrix F raised to some power, and is therefore also 3×10 and D is the diagonal matrix to normalize the product and is 10×10. Thus, we find a matrix Λ which is 10×10 and when we multiple $F\Lambda$, we find the promax loadings as a 3×10 matrix, where again each column is a factor and each row is an item. Thus, the matrix Λ has transformed the varimax loadings F into the promax loadings V_P.

To be thorough, we also attempted to perform the factor analysis with an orthogonal varimax rotation.

In order to determine whether the results are robust and accurate, the analysis was done in both R and Python and then the loadings were compared. While both R and Python have packages available to do factor analysis, they use slightly different methods. If the data and results are robust enough, the loadings will agree between these two. If not, they will be significantly different and therefore, it is likely the analysis is producing results that are not meaningful.

The first step is to begin with making a scree plot to determine how many factors we should use in the exploratory factor analysis. This type of plot examines the eigenvalue versus number of eigenvalues in the correlation matrix, and should have an elbow that indicates the number of factors that might work well in factor analysis. This type of plot is a starting point to determine the number of factors to retain in factor analysis. The scree plot for SPRUCE post-test data is

shown in Figure 3.6, and indicates that either one or five factors would be appropriate as a starting point.
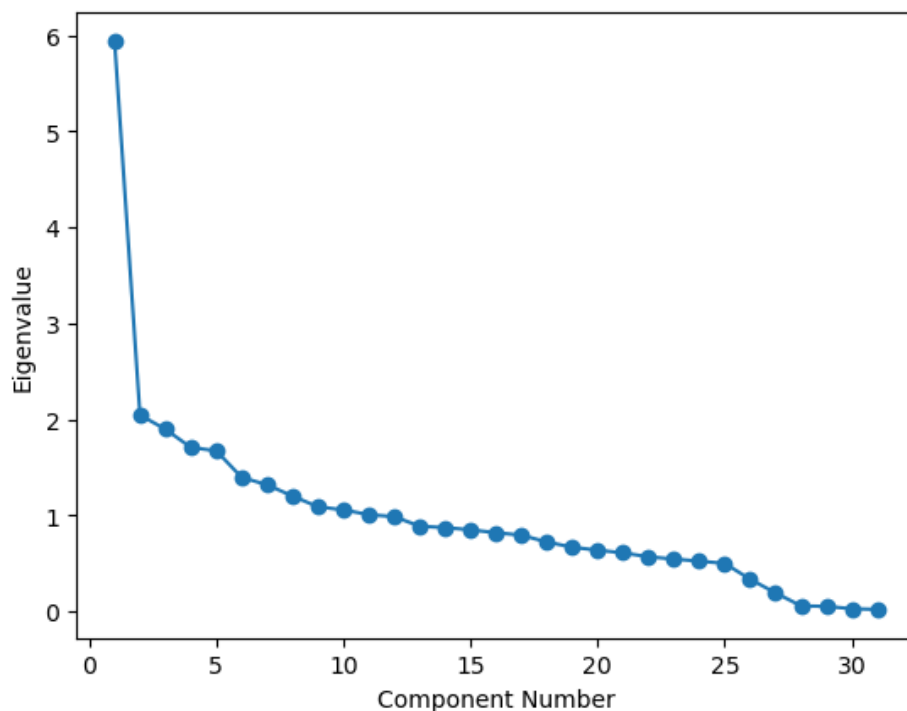
Figure 3.6: Scree plot generated from 1,923 post-test responses to SPRUCE, showing that either one or five factors might explain the data

This number of factors did not align with our initial hypothesis that the factors might correlate with the AOs. However, fourteen (and later, ten) AOs, and therefore factors, is a large number for an assessment of this length, so it is not surprising that fewer factors are appropriate. We do split our AOs into three categories (sources of uncertainty, handling of uncertainty, and distributions and repeated measurements), so three factors is another reasonable guess.

From this analysis in both R and Python (which aligned well with one another), we find no conceptually relevant factors in our data as shown by Tab. 3.13. That is, the factors did not align with either the AOs or the AO categories using either varimax or promax rotation, and any number of factors between two and five (with the three factor promax rotation loadings shown in

Tab. 3.13). Further, none of the factors revealed any theoretically sound groupings.

Table 3.13: Exploratory factor analysis with three factors for SPRUCE post-test data. This includes 1,923 post-test responses to 31 SPRUCE couplets (scored with the ten AO couplet scoring scheme). Couplets are labeled by their item number (denoted by E) and an AO number (using the scheme provided in Tab. 3.1). They are organized by AO. Ideally, we would have liked to see all sources of uncertainty couplets load into one factor, all handling of uncertainty couplets load into one factor, and all distributions and repeated measurements couplets load into one factor with little crossover between these factors. We define high loadings as about 0.3 or greater (with loadings in this range colored in yellow), while the low or zero loadings defined as less than 0.10 [226]. We do not observe this in our data, and the loadings do not reveal any meaningful factors; analyses with other rotations and number of factors provided similar results. This analysis was done using an oblique promax rotation.

Couplet	Factor 1	Factor 2	Factor 3
E2.3B-S1	0.24	0.10	0.08
E3.1-S1	0.64	0.03	0.01
E4.1-S1	0.17	0.06	0.06
E1.1-S2	0.37	0.47	0.11
E3.2B-S2	0.37	0.09	0.50
E3.2C-S2	0.38	0.21	0.00
E4.7-S2	0.34	0.16	0.09
E4.8-S2	0.06	-0.02	0.02
E1.1-S3	0.20	0.86	0.17
E3.2B-S3	0.13	0.26	0.87
E3.2C-S3	0.38	0.20	-0.01
E4.8-S3	0.07	0.08	0.08
E1.2-H1	0.14	0.08	0.02
E2.5-H1	0.19	0.16	0.13
E2.6-H1	0.36	0.18	0.11
E3.3-H1	0.02	0.02	-0.02
E3.1-H2	0.64	0.03	0.01
E3.3-H2	0.07	0.10	0.04
E4.5-H2	-0.01	0.12	0.03
E1.1-D1	0.21	0.95	0.06
E3.2B-D1	0.13	0.26	0.91
E1.1-D2	0.28	0.83	0.01
E3.2B-D2	0.22	0.02	0.61
E2.1-D3	0.42	0.12	0.09
E4.2-D3	0.47	0.14	0.13
E2.2-D4	0.27	0.11	0.11
E4.4-D4	0.43	0.22	0.16
E4.5-D4	0.41	0.10	0.11
E4.6-D4	0.10	0.17	0.08
E3.4-D5	0.37	0.13	0.09
E4.3-D5	0.35	0.08	0.07

Thus, we turn to confirmatory factor analysis instead. In this version of factor analysis, we pre-determine the factor structure and then verify this structure. Because we specify the factors, we do not have to use rotations. We tried several factor structures in this analysis. First, we allowed each AO to be its own factor. Next, we collapsed AOs into their three categories (sources, handling, and distribution) as factors.

The first attempts, using ten factors (one per AO in the updated AOs – see Chapter 4 for details about these updated AOs), failed to converge in both R and Python. Next, we attempted this analysis instead splitting into three factors: sources, handling, and distribution, as related to the AO categories. In this case, the analysis did converge in both R and Python. However, the loadings were significantly different in these two programs as shown in Tab. 3.14. Further investigation revealed that while it did converge, it was likely an edge case that lead to extremely different loadings in the two programs based on how they handle edge cases, and therefore, this can be treated the same as if it did not converge.

Therefore, neither exploratory nor confirmatory factor analysis was able to reduce the dimensionality of SPRUCE post-test data in a way that conceptually matches the assessment. To some extent, this is expected based on the scree plot indicating one factor might be appropriate. One possible explanation for the failed factor analysis is the novel scoring method applied to SPRUCE. Because of the use of couplet scoring, some items are scored multiple times. These couplet scores are clearly correlated, and therefore, factoring them into distinct dimensions without significant overlap will not work appropriately. However, because of the use of this method of scoring, we can provide instructors information on a more fine-grained scale than simply an overall score on SPRUCE, therefore reducing the need for factor analysis.

3.9   Summary and Ongoing Work

In this work, 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) [169] to create the Survey of Physics Reasoning on Uncertainty

Table 3.14: Confirmatory factor analysis with three factors for SPRUCE post-test data. This includes 1,923 post-test responses to 31 SPRUCE couplets (scored with the ten AO couplet scoring scheme). Couplets are labeled by their item number (denoted by E) and an AO number (using the scheme provided in Tab. 3.1). They are organized by AO. Results are shown from both Python (using the FactorAnalyzer package) and R (using the lavaan package) but are significantly different when they should be nearly identical, indicating an issue with the convergence.

Couplet	Python (FactorAnalyzer)			R (lavaan)		
	Factor 1: Sources	Factor 2: Handling	Factor 3: Distributions	Factor 1: Sources	Factor 2: Handling	Factor 3: Distributions
E2.3B-S1	0.88	-	-	0.12	-	-
E3.1-S1	0.91	-	-	0.15	-	-
E4.1-S1	0.88	-	-	0.12	-	-
E1.1-S2	0.93	-	-	0.23	-	-
E3.2B-S2	0.88	-	-	0.19	-	-
E3.2C-S2	0.93	-	-	0.98	-	-
E4.7-S2	0.91	-	-	0.15	-	-
E4.8-S2	0.87	-	-	0.01	-	-
E1.1-S3	0.92	-	-	0.24	-	-
E3.2B-S3	0.91	-	-	0.13	-	-
E3.2C-S3	0.93	-	-	0.97	-	-
E4.8-S3	0.89	-	-	0.07	-	-
E1.2-H1	-	0.81	-	-	0.24	-
E2.5-H1	-	0.84	-	-	0.34	-
E2.6-H1	-	0.85	-	-	0.45	-
E3.3-H1	-	0.84	-	-	0.11	-
E3.1-H2	-	0.85	-	-	0.40	-
E3.3-H2	-	0.86	-	-	0.24	-
E4.5-H2	-	0.83	-	-	0.18	-
E1.1-D1	-	-	0.93	-	-	0.94
E3.2B-D1	-	-	0.90	-	-	0.33
E1.1-D2	-	-	0.94	-	-	0.93
E3.2B-D2	-	-	0.86	-	-	0.17
E2.1-D3	-	-	0.87	-	-	0.26
E4.2-D3	-	-	0.91	-	-	0.29
E2.2-D4	-	-	0.90	-	-	0.19
E4.4-D4	-	-	0.93	-	-	0.33
E4.5-D4	-	-	0.92	-	-	0.23
E4.6-D4	-	-	0.86	-	-	0.16
E3.4-D5	-	-	0.90	-	-	0.26
E4.3-D5	-	-	0.91	-	-	0.20

Concepts in Experiments (SPRUCE) to meet this need. While a previous paper [184] discussed background research (*domain analysis*), the layers of ECD discussed in this work deal with: the

creation of assessment 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*).

While the ECD documentation depicts a fairly linear progression through the ECD layers, we found iteration across layers (outlined in Sec. 3.3.2) 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, items formats, item prompts and answer options, and evidentiary arguments.

Chapter 4

Evidence for validity and reliability of SPRUCE

This chapter is adapted from an article submitted to Physical Review Physics Education Research [90].

4.1   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 [107, 230], thermodynamics [196], electricity and magnetism [47, 260], and quantum mechanics [158, 266], as well as for lab courses regarding critical thinking [240], handling of measurement uncertainty [45, 188], handling of data [58], modeling [181], and views about experimental physics [270]. 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 [184, 235].

Prior research on these topics has revealed challenges, such as students' lack of understanding

regarding the importance of taking multiple measurements during experimentation [225] and the belief in a singular "true" value [50, 142]. 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 [112]. 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 Chapter 3.

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 via Evidence-Cented Design, or ECD [169], 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.

Evaluating the external validity of assessments is a well-established practice [26], 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 [236]."* 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 chapter 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.

4.2   Background

4.2.1   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 [153]. 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) [107] and the Force and Motion Conceptual Evaluation (FMCE) [230], both designed to evaluate introductory physics students' understanding of simple Newtonian mechanics; the Physics Measurement Questionnaire (PMQ) [45], the Physics Lab Inventory of Critical thinking (PLIC) [114,194,240,241], and the Concise Data Processing Assessment (CDPA) [58], intended to evaluate students' handling of measurement uncertainty and general experimental skills; and the Colorado Learning Attitudes about Science Survey (CLASS) [18] and Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) [270], 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 [107, 189].

RBAIs are often created through a rigorous development process. One such development process is Evidence-Centered Design, or ECD [169]. This framework was used to guide the design and implementation of SPRUCE. Other examples of assessment frameworks include the Three Dimensional Learning Assessment Protocol [141] and the framework described by Adams and Wie-

man [20]. 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.

4.2.2   SPRUCE

SPRUCE aims to measure students' proficiency with measurement uncertainty concepts and practices. While some assessments, such as the Physics Measurement Questionnaire (PMQ) [45], the Physics Lab Inventory of Critical thinking (PLIC) [114, 194, 240, 241], and the Concise Data Processing Assessment (CDPA) [58] 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 in Chapter 3. 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 [259], and coupled numeric open response. More information about the development of the items and the types of items on SPRUCE is discussed in Chapter 3.

SPRUCE consists of 19 items grounded in four experiments, which probe 10 AOs. The SPRUCE AOs are shown in Tab. 4.1 and are organized into three categories of measurement uncertainty: sources of uncertainty, handling uncertainty, and distributions & repeated measurements. These were designed based on findings from instructor interviews [184]. AOs provide many affordances, as detailed in Chapter 2.

For clarity, we note that these AOs are slightly different than those previously reported for SPRUCE (in both Chapter 3 and [235]). During the process of scoring and validating the

Table 4.1: SPRUCE assessment objectives, organized by assessment objective category.

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

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 Tab. 4.1). 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.

SPRUCE was designed by iterating through the ECD framework [169]. As described in previous work [184, 235], 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 chapter 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 [235]. This important step of determining evidentiary reasoning is a critical part of the ECD process.

4.2.3   Classical Test Theory

Classical test theory (CTT) 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 [61]. 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 [72]. 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 chapter. 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 measures (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 [198]. 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.

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

Figure 4.1: 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.

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 [198], 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 [61], CTT is an important first step before further validating the assessment using other methods. Additionally, CTT requires considerably fewer data than IRT.

4.2.4   Scoring By Couplet

As discussed in Chapter 2, 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 Tab. 4.1). 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. 4.1 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 4.2 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.

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

Table 4.2: Example scoring for couplets of item 3.3

Answer Option		Score	
		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

of scoring and are used in the statistical validation presented in this chapter. Chapter 2 presents 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 Chapter 2. 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 for them to take. Additionally, 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.

4.3 Methods

4.3.1 Data Collection and and Cleaning

Instructors who teach 36 different courses at 22 institutions administered SPRUCE in the Fall 2022, Spring 2023, and Fall 2023 semesters, using a pre-post format. We recruited instructors who had previously expressed interest in this project, as well as by posting advertisements on the ALPhA (Advanced Laboratory Physics Association) 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

Table 4.3: 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.

	Num. Institutions	Num. Students
Highest Degree		
PhD	7	2,152
Master's	4	325
Bachelor's	9	86
Associate's	2	33
Minority Serving Status		
HSI	4	48
AANAPISI	1	29

intended for physics majors and courses intended for other science majors). Table 4.3 shows brief information about the institutions that administered SPRUCE.

In the work presented here, we use only the post-test data from all three semesters. Validation results do not take into account pre-test data because students have not yet learned the material they are being tested on. From these three semesters of post-test data, we received 3,644 responses, of which 2,596 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 non-consent 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 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.

4.3.2   SPRUCE Scoring Scheme and CTT

The scoring scheme we used for SPRUCE as discussed in Sec. 4.2.4 takes into account the fact that SPRUCE is a multi-construct assessment. Similar to the CDPA [58], the PLIC [241], and the FCI [219], 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. 4.2 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. 4.2. 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. 4.2. 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 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 Tab. 4.4.

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. 4.2. 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 don't 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

Table 4.4: 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]

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. 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. 4.2. 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 sub-scores 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
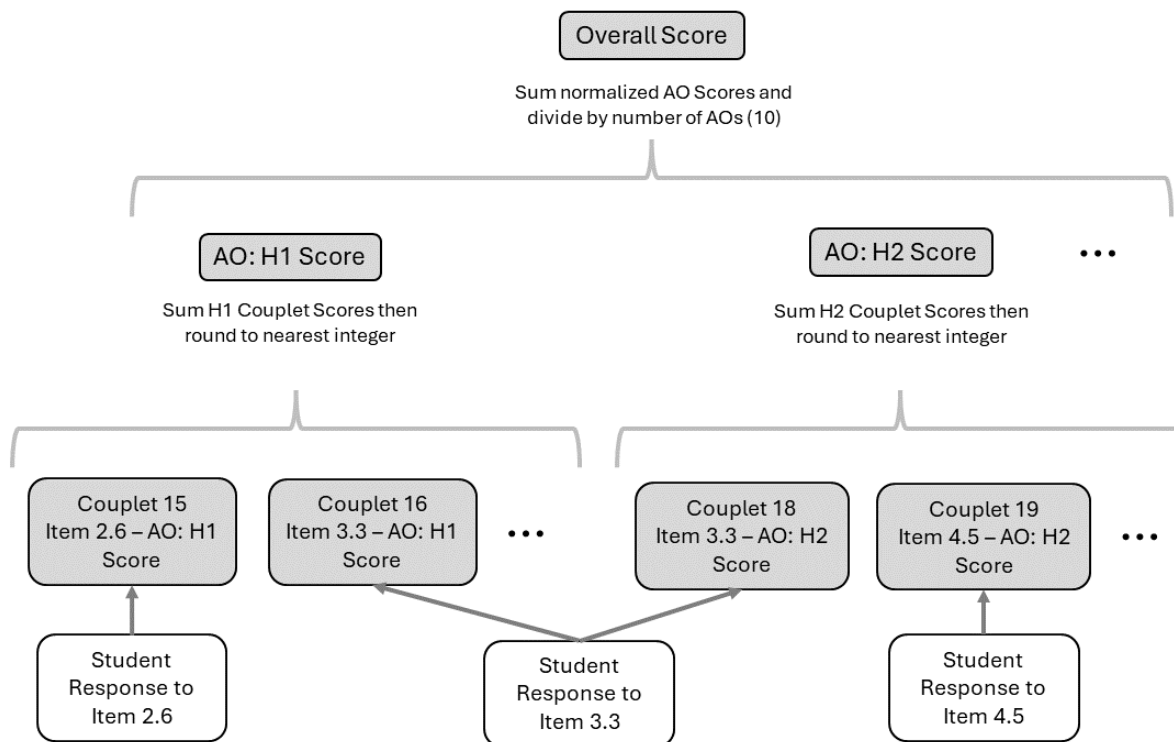
Figure 4.2: Flowchart indicating how to proceed from student responses to an item to couplet scores, AO scores, and an overall score. Students respond to items, which are then paired with AOs. These item-AO pairs are scored as couplets. The couplet scores for each AO are then summed, rounded, and normalized 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.

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 Tab. 4.5.

4.4   Results and Discussion

4.4.1   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

Table 4.5: 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	✓	-	-

implementing a system to exclude responses based on incorrect answers to the most straightforward questions. For example, one such item presents 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 Sec. 4.5 for couplet numbering).

Couplet 12 asks 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 requires 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 Chapter 5. The expert average on the pictorial version of this couplet was 81%. Therefore, we believe that the low average of 59% on the nu-

merical 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 $\pm$ 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 $\pm$ 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 and the most common incorrect answer does not align with conventional instruction around this topic, we chose to keep this couplet as-is in the assessment.

4.4.2   Overall Score

From three semesters of data collection, we analyzed a total of 2,596 post-test responses to SPRUCE. In Fig. 4.3, 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 $\pm$ 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
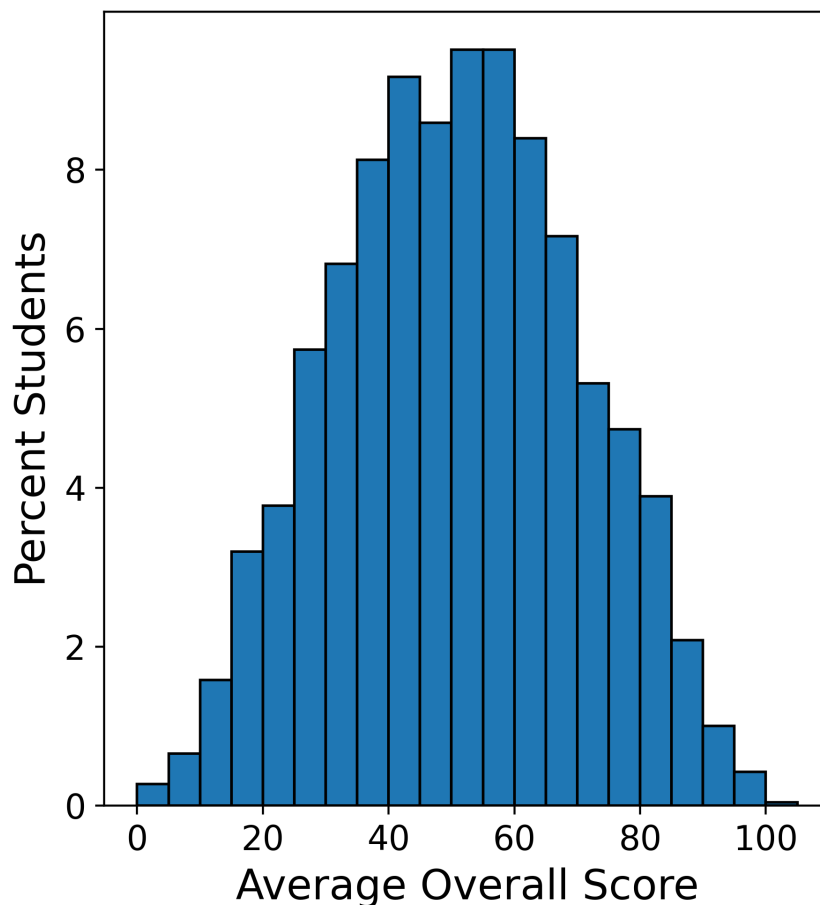
Figure 4.3: Histogram showing the distribution of overall post-test scores on SPRUCE [N = 2,596]. This distribution is normal, as determined by the Anderson-Darling test, skewness, and kurtosis.

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 [25, 172], as well as determining 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 overall distribution of the overall score data is very close to normal as the skewness and kurtosis are between -1 and 1 [98] [1] .

4.4.3   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 inter-rater reliability (IRR) testing on assignment of the couplets. We provided two independent (i.e., had never worked on SPRUCE) physics education researchers with PhDs 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 inter-rater reliability was performed due to too small numbers of 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.

4.4.4   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

---

[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

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 [62]. 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 statement about item difficulty is not always applicable, especially because the the main rationale for removing very easy and very difficult items is because they are generally poor for discrimination [61, 62].

Doran describes an additional schema for determining "good" values of difficulty, where a distribution of difficulties amongst the items is ideal [62]. This distribution should be tailored for the intent of the assessment and the level of instruction. Instead of hard cutoffs, Doran et al. recommends having questions of varying difficulty at all levels. To this end, we show the distribution of couplet difficulties in Fig. 4.4, 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 Sec. 4.5 in Tab. 4.8. 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
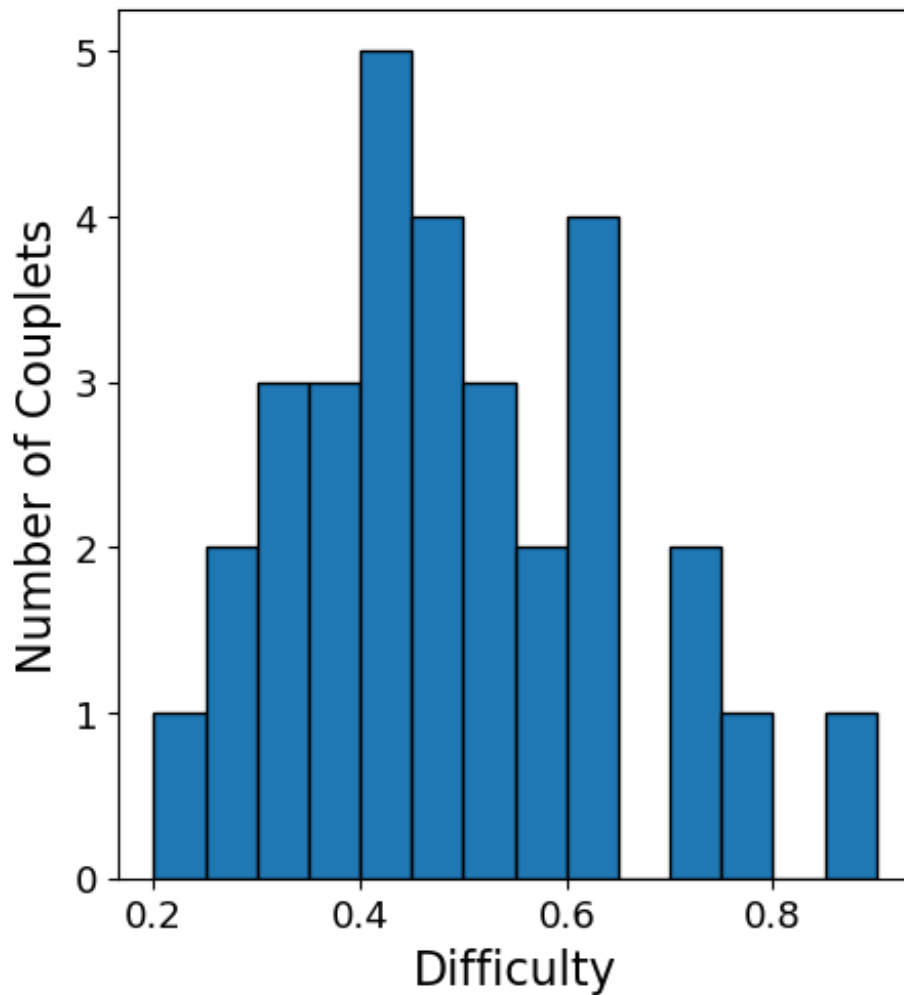
Figure 4.4: Individual couplet difficulties. The histogram shows the distribution of couplet difficulties. A large spread of difficulties, as seen here, is a sign of a robust assessment.

couplet difficulty was 0.49 ± 0.03, again showing an acceptable level of whole-test difficulty.

Next, we present AO-level difficulty in Tab. 4.6. 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.

Table 4.6: 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 ± 0.01)

AO	Difficulty	Disc. Index	Pearson Coef.
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)

4.4.5   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.

4.4.5.1   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 [61]. To determine this measure, we use the equation outlined in Ding and Beichner [61]:

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

(4.1)

where N represents the total number of students in the dataset (in our case, 2,596), K is the number of AOs (10), and f_i is the number of students whose overall score is i [61]. The Ferguson's delta for SPRUCE is $\delta = 0.947$. Because this is above 0.90 [127], 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 (ex, [0,1), [1,2), etc.). This is necessary for the statistic to be calculated.

4.4.5.2 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 calculate the discrimination index, D. This is done by using data from only the top and bottom 27% of performers [72] 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 [62] 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 Sec. 4.5.

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

$$r = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \tag{4.2}$$

where the x_i refers to the score on a particular couplet by the i-th 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 i-th 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 $r \geq 0.2$ [125]. 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 Sec. 4.5.

Further, we calculate the minimum critical Pearson coefficient value [54] for each couplet:

this is defined as being two standard deviations above zero, where the standard deviation is given by:

$$\sigma_r = \frac{1}{\sqrt{N-1}}, \tag{4.3}$$

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 $r_{\min} = 0.039$. 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 4.5 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 exceed these ranges.

Couplet 13 (see numbering in Sec. 4.5) 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 [184], 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.

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). 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 maybe cause of the low discrimination for this couplet.

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 and thus need measures of student performance in it in order to improve instruction which would be loss with removal of the couplet.

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.
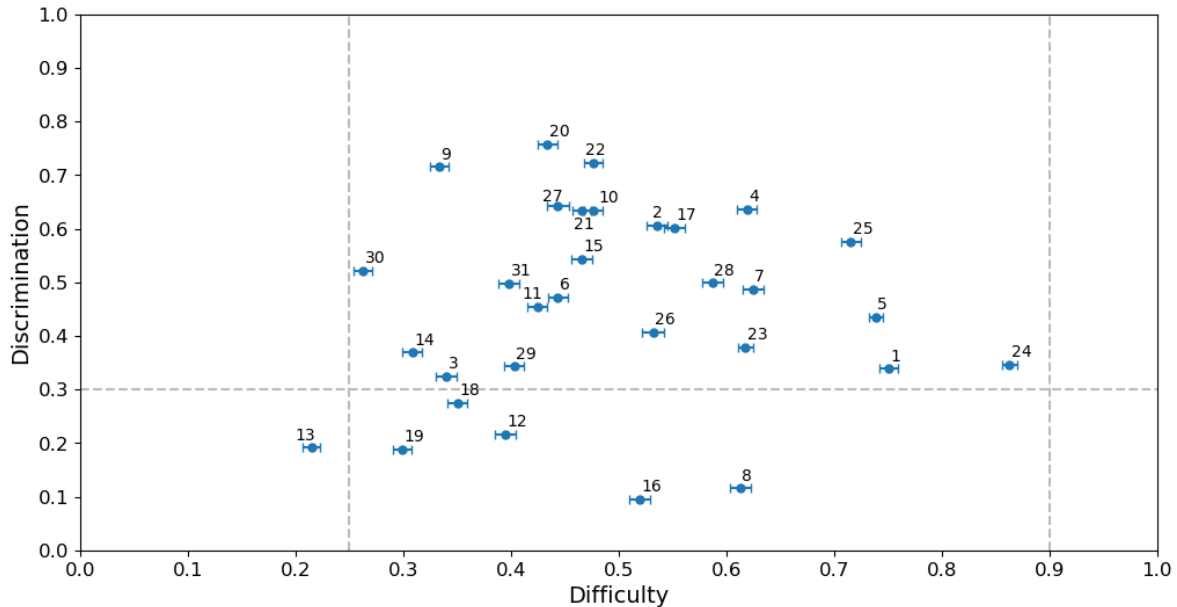

Figure 4.5: Discrimination vs. Difficulty for each item-AO couplet on SPRUCE. Couplet number labels match the corresponding couplet number in Sec. 4.5 Gray dashed lines indicate reference cutoff values.

4.4.5.3 AO-level discrimination

We present AO-level discrimination indices and Pearson coefficients in Tab. 4.6. 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.

4.4.6   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 [72]. 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.

4.4.6.1   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, for example, 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 [72] do not apply. Instead, we used an alternate approach as described in Day and Bonn [58]. We find a Pearson coefficient between two administrations of the same test in different semesters. We calculate the Person 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 [126]. For the Fall 2022 and Fall 2023 semesters, we find the stability of SPRUCE to be 0.98 ± 0.01 with with $p \ll 0.01$. We choose 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. [49]). In either case, the stability is both significant and high, showing that SPRUCE conforms to test-retest stability when calculated in this manner.

4.4.6.2   Internal consistency

To determine the internal consistency of the entire assessment, we use two methods. First, the assessment is fully scored, and then randomly split into two halves, each with five of the ten AOs. The Pearson coefficient is then calculated between the average scores on the two sets of five AOs. Next, the Spearman-Brown prophecy equation is applied as a correction due to each half of the assessment having fewer items than the whole [72] and is given by:

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}}. \tag{4.4}$$

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 [62, 72]. 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 is calculated using the following equation [72]:

$$\alpha = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum_{i=1}^{K}\sigma_i^2}{\sigma_t^2}\right). \tag{4.5}$$

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 $\alpha = 0.83\pm0.01$, showing that the AOs are internally consistent with one another, since this is above the generally accepted cutoff of about 0.70 [62, 72].

We further calculate 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.

Table 4.7: Summary of statistical test results for whole assessment, SPRUCE (N = 2,596). 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 (Fall'22/Fall'23)	$[-1, 1]$	≥ 0.70	0.98 ± 0.01
Test-retest stability (Fall'22/Spring'23)	$[-1, 1]$	≥ 0.70	0.985 ± 0.009

4.5 Individual Couplet Statistics

Here, we present statistics for all individual couplets on SPRUCE in Tab. 4.8.

4.6 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-

Table 4.8: Summary of statistical test results for each SPRUCE couplet [N = 2,596]. See Tab. 4.1 for full text of each Assessment Objective (AO). 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

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

is collected to make this feasible. In addition, future work is forthcoming regarding student learning gains compared between pre- and post-instruction 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.

Chapter 5

Representational differences in how students compare measurements

This chapter is adapted from Geschwind, et. al. 2023 [92].

5.1   Introduction & Background

All measured quantities have associated uncertainties, making measurement uncertainty a crucial aspect of experimental physics. Using measurement uncertainty correctly is essential for interpreting measurements, presenting results, and drawing reliable conclusions based on those results. The Effective Practices for Physics Programs (EP3) Guide [7] also emphasizes the significance of learning measurement uncertainty techniques as taught in physics laboratories. Despite its critical role, students frequently struggle with concepts and practices surrounding measurement uncertainty, including propagation of error, comparison of measurements, calculating standard deviations and standard errors, and taking several measurements to get a distribution of results, even after taking a course emphasizing these areas [45, 112, 116, 134, 194, 216].

As part of efforts to improve student learning of measurement uncertainty, we have developed a new research-based assessment instrument (RBAI) called the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) [184, 235]. SPRUCE is an online assessment intended to be utilized in a pre-post format allowing instructors to measure the impact of a course on students' proficiency with concepts and practices of measurement uncertainty. We developed SPRUCE using the framework of Evidence-Centered Design (ECD) [169], a robust method of creating and validating an RBAI. Although validation is an ongoing project (see Chapter 4), SPRUCE

still offers a wide variety of insights into how students handle measurement uncertainty. Its design provides instructors with their students' progress along 14 dimensions referred to as Assessment Objectives (AOs) [236] after one term of a laboratory class.

AOs are *"concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment* [236].*"* AOs are similar to course learning goals and are essentially the constructs the assessment aims to measure. We developed the SPRUCE AOs with input from introductory laboratory instructors to determine which aspects of measurement uncertainty they find important and want their students to learn in their courses [184]. These AOs then aided in writing the SPRUCE assessment items: each item on SPRUCE addresses at least one of these objectives. In this way, we focused the scope of SPRUCE to topics instructors frequently deem important to their introductory laboratory courses.

Here, we examine one particular SPRUCE AO: *Determine if two measurements (with uncertainty) agree with each other.* SPRUCE has two isomorphic questions for this objective. First, the assessment presents students with numerical measurements and asks about agreement between these measurements. Then, later in the assessment, with several questions in between, a similar question appears with the data represented pictorially, as symbols with error bars. Students are not explicitly informed about the relationship between these two items. This allows us to probe how students are able to compare measurements when presented with the same data with two different representations.

Existing literature has explored the use of multiple representations while students problem solve [129–132, 224]. For example, Kohl et al. found that students frequently view a mathematical problem and a pictorial problem as 'opposites,' where students consider pictorial problems as more aligned with "concepts," which are frequently treated distinctly from numerical problems. Further, they found statistically significant differences in performance based on different representations of isomorphic problems on homework and quizzes. Students tended to perform worse on problems in a mathematical or numerical format than with problems in other formats (e.g., pictorial, verbal, or graphical) [129].

The work presented here aims to identify whether student performance in comparing measurements similarly depends on representation. To do this, we will answer the following research questions.

- Do students respond differently to questions about comparing measurements when presented with different representations?

- How do students reason about comparing measurements when presented with different representations?

5.2   Methodology

We use a mixed methods approach, as the data collection and the analysis involve qualitative and quantitative components. To study students' handling of measurement uncertainty, we administered SPRUCE in a pre-post online format during the Fall 2022 semester in 12 courses at eight institutions (See Tab. 5.1). We received 670 valid post-instruction responses after we removed responses from students who did not consent to have their data used for research, did not correctly answer the filter question, or did not answer both items of interest.

Table 5.1: Institutions and student responses in the dataset after removing student answers for incorrect filter/nonconsent to research

Number of Institutions	Institution Type	Number of post responses
1	2 Year	7
1	4 Year	7
1	Master's	39
5	PhD	617

We also conducted interviews during the Fall 2022 semester. These interviews aimed to determine whether students interpreted all of the items on SPRUCE as intended, as well as to probe student reasoning for each answer option on the assessment. Students were recruited from seven courses at four institutions (2-year, Master's and PhD granting) already participating in the administration of SPRUCE during this semester. Each of the 27 interviews conducted lasted

approximately one hour and students were compensated for their time. Interviewers (Michael Vignal and myself) observed as students completed SPRUCE and inquired about students' reasoning for each answer selected, as well as about why they did not select certain answer options. The interviews were audio/video recorded for future reference. Analysis of these interviews consisted of taking notes during interviews and transcribing student quotes as needed.

For the analysis, we focused on the responses to two isomorphic multiple-response items, focusing on both the difficulty [61, 72] of these items and student reasoning for their answers to both. The first item, as shown in the upper half of Fig. 5.1, presents students with their 'own' numerical data (with uncertainty) for a measurement of a spring constant; they are then asked to select all answer choices of numerical data (means with uncertainties) that agree with their measurement. The second item presents these similar data in pictorial form, as shown in the lower half of Fig. 5.1. For brevity, we will refer to the numerically represented item as NRI and the pictorially represented item as PRI for the remainder of this chapter. Students receive credit on these multiple response items by answering with the combination 'ABCD' or 'ABCDF,' based on expert responses. The uncertainties in both items represent the standard error; therefore, overlap or near overlap of the error bars is required for agreement. No other answer combinations earn credit, and no partial credit is awarded for these items. Note that we changed the order of answer options for the PRI for this chapter to make discussion of the items easier.

5.3   Results & Discussion

5.3.1   Overall difficulty scores

While laboratory instruction commonly focuses on measurement comparison [184], low scores on both of these items at the end of the term indicate persistent student difficulties in handling comparison with uncertainties. Students score an average of $(25 \pm 3)\%$ on the NRI and an average of $(40 \pm 4)\%$ on the PRI on the post-test, with the error indicating 95% confidence interval. These scores indicate that, while not many students answered these items correctly, students answered

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

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

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

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

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

☐ (A) $3.71\frac{N}{m} \pm 0.06\frac{N}{m}$ ☐ (E) $3.91\frac{N}{m} \pm 0.06\frac{N}{m}$

☐ (B) $3.71\frac{N}{m} \pm 0.17\frac{N}{m}$ ☐ (F) $3.91\frac{N}{m} \pm 0.17\frac{N}{m}$

☐ (C) $3.76\frac{N}{m} \pm 0.06\frac{N}{m}$ ◯ (G) None of these agree with my data

☐ (D) $3.76\frac{N}{m} \pm 0.17\frac{N}{m}$

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

☐ (A)

☐ (B)

☐ (C)

☐ (D)

☐ (E)

☐ (F)

◯ (G) None of these agree with my data

Figure 5.1: Two Isomorphic Items on SPRUCE. These items probe student understanding of measurement comparisons with uncertainty by presenting the same data in two different representations - a numerically represented item (NRI) and a pictorially represented item (PRI). The students first encounter the NRI and then, after answering several unrelated questions, they encounter the PRI. Note that the answer options on the PRI are in a different order when presented to students (DAEBFCG) than shown here; we present them in the same order as the answer options for the NRI in this chapter for ease of understanding.

the PRI correctly more often. We conducted a Mann-Whitney U test (a nonparametric test for independent measures) [156] to determine if this represents significant statistical difference, and

Table 5.2: Number of students who answered the NRI, PRI, or both correctly [N = 670]; error shown as 95% confidence interval

	Only NRI	Only PRI	Both Correct
Number of Students	38	134	131
Percent of Students	6 ± 1	20 ± 3	20 ± 3

found the p-value for these items as $p = 2.1 \times 10^{-8}$, indicating a statistically significant difference in student performance on these items. Additionally, we calculated the effect size to compare these two items using Cohen's d [51, 124], finding $d = 0.31 \pm 0.05$, showing a moderate effect size.

We also calculated the Pearson coefficient to determine the correlation between the two items. The Pearson coefficient varies between $r = -1$ and $r = 1$, where a more positive coefficient indicates a stronger positive correlation [72]. Anything above $r \approx 0.30$ indicates a fairly significant positive correlation. For these items, we find $r = 0.45 \pm 0.04$, which shows a fairly significant correlation in that if a student correctly answered one item, they are more likely to have correctly answered the other. However, the correlation is not perfect ($r = 1$): many students correctly answer only one of these items. The number of students who answered each question correctly is presented in Tab. 5.2. Only about half of the students who correctly answered the PRI also correctly answered the NRI, but about 75% of the students who correctly answered the NRI also correctly answered the PRI. This suggests that students who are able to reason through the numerically presented data seem better equipped to handle the pictorially presented data, but the reverse is not true on average.

We turn to the qualitative interview data to help to understand these results. During interviews, some students describe mentally switching from a numeric to a pictorial representation easily and using this skill to solve the numeric item:

I just looked at the values and saw it – like I kind of picture if they have that little bar with their error bars to see if they overlap.
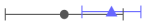
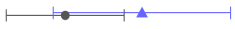This student essentially converted the numeral data into pictorial data in their mind and then used that representation to reason about the comparisons. Using this skill of mentally changing representations, they were able to answer both items correctly. This finding is similar to ones from

Kohl et al. [129] and Weliweriya et al. [243], in which students were often able to switch between different representations when forming mental models of data.

5.3.2 Individual answer analysis

In addition to comparing how well students scored on each question, we want to look at which answer options students choose to gain more insight into student reasoning. We determined how many students selected each of the seven answer options (due to the multiple response nature of the question, students could select multiple options, hence we do not expect these numbers to add up to 100%). Table 5.3 shows these data with 95% confidence intervals.

Table 5.3: Percentage of students who selected each answer option with 95% confidence interval [N = 670]

Numeric Representation (NRI)	Percent of Students	Pictorial Representation (PRI)	Percent of Students
A. 3.71 ± 0.06	58 ± 4	**A.**	72 ± 3
B. 3.71 ± 0.17	66 ± 4	**B.**	79 ± 3
C. 3.76 ± 0.06	45 ± 4	**C.**	46 ± 4
D. 3.76 ± 0.17	55 ± 4	**D.**	69 ± 4
E. 3.91 ± 0.06	10 ± 2	**E.**	1.3 ± 0.9
F. 3.91 ± 0.17	15 ± 3	**F.**	7 ± 2
G. None of these agree with my data	6 ± 2	**G.** None of these agree with my data	1.5 ± 0.9

For both the PRI and the NRI, students most commonly select B, in which the means of both measurements lie within each other's error bars. The second most common choices were A and D, in which the error bars of only one of the measurements overlaps with the mean of the other measurement. This shows that, frequently, students require one of the means to be within the error bars of another measurement, as opposed to accepting error bar overlap as agreement between two measurements with uncertainty.

Again from Tab. 5.3, many more students selected answer option E for the NRI than the PRI; this answer option is the only one where the two measurements definitely do not agree. Students identify this disagreement more frequently when presented with the data pictorially, where it is

clear that the error bars are very far from one another, rather than when presented with this same data numerically. During interviews, one student selected all answer options (aside from "None of the above") on the NRI, and said:

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

However, this student provided expert-like reasoning regarding overlap of the full range of each measurement when correctly answering the PRI, showing a clear difference in thinking about measurement comparison between the two representations.

Knowing the most commonly selected answer options allows us to delve further into common incorrect answer *combinations* and reasonings for these choices. Figure 5.2 shows a heat map of the most common answer combinations to each of the two questions (representing 409 of the 670 total student answers). The diagonal represents students who chose the same answer options for both the NRI and the PRI; the off-diagonal elements are students who selected different answers for each of these items.

One of the more common incorrect combinations on both items is 'AB' [NRI: $54/670 = (8\pm 2)\%$, PRI: $79/670 = (12\pm 2)\%$]. This incorrect response aligns with students who consider their measurement more important in some sense, and therefore believe that the other groups' mean must be within their own error bars in order the measurements to agree with one another as compared to the other way around (requiring their mean to be within the other measurement's error bars), as would be indicated by the selection of 'BD' [NRI: $45/670 = (7\pm 2)\%$, PRI: $63/670 = (9\pm 2)\%$].

For example, one student who selected only 'AB' on the numeric item said:

For the other four groups, the uncertainties for their values did not put them in the same range as my values with its uncertainty so I don't believe they agree with my value.

In other words, when comparing numeric measurements with uncertainty, they placed more weight
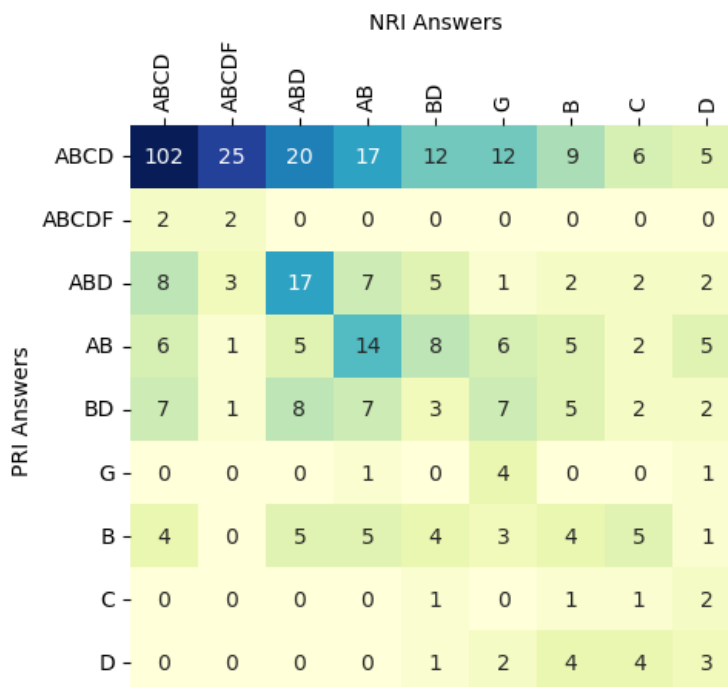
Figure 5.2: Heat map showing the most common answer combinations for the NRI and PRI [N = 409]. Answer combinations ABCD and ABCDF were marked as correct; no other combinations earned credit. Diagonal elements indicate students who answered identically to both the NRI and PRI, and off-diagonal elements indicate students who answered the items differently.

on their own measurement — in order for agreement to occur, the uncertainty of the other measurement had to encompass their own mean. When solving this problem, they only added and subtracted their uncertainty to their own value and then selected the two answers whose means fell within that range; they ignored the uncertainties in the measurements in the answer options. However, we note that when answering the PRI, this same student selected a correct response of 'ABCD', and provided expert-like reasoning. Thus, their reasoning changed with representation.

This theme of placing more importance on their own measurements frequently appeared in student interviews. Occasionally it was present when students provided reasoning for the PRI, but it was more typically found in student answers to the NRI.

Another common incorrect answer for both items is ABD [NRI: $57/670 = (9 \pm 2)\%$; PRI: $60/670 = (10\pm2)\%$]. In this line of incorrect reasoning, students did not consider answer option C

to be correct, in which just the error bars overlap: they required at least one of the means to be within the error bars of the other measurement in order for agreement between measurements to occur.

Student reasoning from interviews supports this interpretation. For example, one student interviewed selected 'ABD' on the PRI because:

Not only do a large portion of their error bars overlap, it also contains the measurement itself,

when referring to answer option B and D. They then chose A because:

I would include [A] because now that measurement is included in mine, but [C] I am not sure about because... I don't necessarily know for sure they agree.

In this example, the student did not consider answer option C to show agreement despite the error bar overlap - instead, they placed additional emphasis on requiring the mean to be included in at least one of the uncertainties of the other measurement.

Figure 2 also shows that very few students chose 'ABCDF' for the PRI [4/670, or $(0.60\pm 0.06)\%$], but many more students chose this for the NRI [the heat map shows 32 of the $35/670 = (5 \pm 2)\%$ students who chose this option]. In the PRI, answer option F is one in which the error bars do not overlap, but are very close to each other, showing that agreement might be possible, hence why selection of F was not considered when scoring this item – this option's correctness largely depends on which guidelines instructors teach students. Additionally, interview data showed mixed reasoning for students who selected this option.

5.4   Conclusions & Takeaways

Overall, students performed better on the PRI than the NRI, showing a more expert-like understanding of measurement comparison when presented with a pictorial format. However, students' did not perform as well as desired on either item, indicating room for improvement in teaching this important skill to students. Only about 40% of students correctly identified whether measurements with uncertainties agree with one another in a pictorial format, and this drops to only about

25% when presented numerically instead. Since many scientific papers generally provide numbers with uncertainties for measurements, this is a valuable skill needed in their future scientific careers to interpret experimental results. It is also vital for students to be able to work with many representations of data and convert between them. This study suggests that having students work with multiple representations, and convert between them, could be beneficial for developing expertise with measurement uncertainty and comparing measurements.

In future work, we will examine pre-post gains across this objective by examining scores prior to, and after, instruction in introductory laboratory courses. Additionally, we will explore other research directions using SPRUCE data, such as students' ideas around accuracy and precision and their ability to propagate errors to obtain an uncertainty in a calculated quantity. Finally, we will examine the alignment of student performance on SPRUCE with a variety of variables, including race, gender, institution type, and instructional methods.

Chapter 6

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

This chapter is adapted from an article submitted to Physical Review Physics Education Research [91].

6.1   Introduction

Measurement uncertainty is a core concept in physics experiments, as all measured quantities have associated uncertainties. Knowledge of uncertainty and how it affects interpretation of the outcomes from an experiment is crucial for both presenting results from experiments and understanding others' work. The importance of measurement uncertainty has led to recommendations for including this topic in introductory science laboratory courses [7, 15, 138]. However, instruction in this area could often be improved, with students frequently struggling to understand many of the important aspects of measurement uncertainty, including error propagation, taking several measurements to get a distribution of results, and comparing measurements with uncertainty, even after taking a course emphasizing these topics [41, 45, 73, 112, 114, 116, 134, 194, 216].

To facilitate improved learning of measurement uncertainty in laboratory courses, we previously developed a research-based assessment instrument (RBAI), the Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE). SPRUCE was developed to measure student proficiency with measurement uncertainty practices along 10 dimensions that were identified as important to undergraduate physics laboratory instructors [184, 235]. It was designed for first- and

second-year lab courses and is to be given in a pre-post instruction format via an online survey platform. The results of the surveys are analyzed by the development team and are presented to the instructor in an easily interpretable report, where their course's data is shown in comparison to aggregate data from other courses. These data can be used by the instructor to improve the course as well as by researcher to learn about student understanding of measurement uncertainty.

In the work presented here, we produce the first research results from an analysis of students' responses to SPRUCE from many courses and institutions to provide a broad landscape of student understanding in this area by answering the following research questions:

(1) How proficient are students at demonstrating the practices and concepts of measurement uncertainty as a general topic and on sub-topics such as comparing measurements or propagating uncertainties? Where do they excel and where do they need additional support?

(2) How does instruction impact student proficiency with the practices and concepts of measurement uncertainty?

- How does instruction impact this proficiency for students with different majors and genders?

- How does the importance an instructor places on specific learning objectives about measurement uncertainty impact this proficiency?

We choose to investigate both major and gender for several reasons. First, because students' major and gender are often intertwined, choosing to explore one of these variables while ignoring the other could lead to inaccurate results for the variable included in the analysis [21]. Additionally, prior studies have noted correlations with both of these variables on students' scores on physics assessments, leading to recommendations for removing gender bias from both instruction and assessment [59, 239, 253]. Because an important goal for the creation of SPRUCE is data literacy for all students, not just physicists, we are interested in examining the correlations between students' major and their performance on SPRUCE.

To answer these questions, we present results from 1,576 students enrolled in 31 courses at 20 institutions in which SPRUCE was administered during the Spring 2023 and Fall 2023 semesters. We first provide an analysis of post-instruction responses, including a deeper look into student reasoning along several of the different areas of measurement uncertainty measured by SPRUCE. We then analyze pre-post shifts, including the impact of students' major(s) and gender on the results. Finally, we examine, in detail, three separate areas of measurement uncertainty — Sources of uncertainty, Handling of uncertainty, and Distributions & repeated measurements — including both results from statistical analysis and example student reasoning provided during think-aloud interviews.

6.2   Background

6.2.1   Previous Work on student learning of Measurement Uncertainty

Previous studies have explored students' handling of measurement uncertainty in undergraduate physics laboratory courses. For example, students frequently have misconceptions about uncertainty, some of which do not improve even post-instruction [45, 73, 114, 139, 154, 225, 238, 240, 241]. In addition to these studies, much has been learned about students' use of measurement uncertainty through previous RBAIs on this topic. These misconceptions and the use of RBAIs are discussed in detail below.

The first of these RBAIs is the Physics Measurement Questionnaire (PMQ) [45, 238]. This RBAI aims to examine students' ability with measurement uncertainty, specifically looking at repeated measurements and measurement comparison with uncertainty. It consists of multiple choice questions followed by open-response questions allowing students to provide their reasoning for their multiple-choice selections. Nominally, only the open-response is coded for the analysis. This makes it difficult to perform a large-scale administration of this RBAI. Through use of this survey, researchers in South Africa found that they could separate student thinking into two categories: point-like and set-like reasoning. Students who fall into the point paradigm believe there is a

"true" experimental value, whereas students who use set-like reasoning understand that experiments provide incomplete information about a measured quantity and all data must be combined to obtain a best value. Additionally, some students were found to use mixed reasoning (a combination of both point- and set-like reasoning). [41]. Many students in this study retained their point-like views of experimental physics after instruction, and only about 20% of physics majors were found to exhibit a more set-like view of measurement uncertainty post-instruction [238].

A partial version of the PMQ was implemented at the University of Colorado Boulder to examine the impact of a transformed laboratory course. Researchers observed a shift from mixed reasoning to set-like reasoning after instruction for both the traditional and transformed courses. Very few students exhibited solely point-like reasoning in any case (even before instruction), though mixed reasoning was not uncommon. They also found that the course transformation had a positive impact in that students shift towards more sophisticated reasoning in the transformed version of the course [146, 182, 185, 188, 247]. Overall, these student responses differed significantly from the student responses from South Africa, which could be ascribed to significant differences in the student populations surveyed. Thus, while the point-like and set-like reasoning paradigm might be a useful classification scheme for student reasoning in some cases, it does not capture the full range of students' ideas and skills with measurement uncertainty.

Other work has built on the PMQ, using some of the same probes and adding additional fine-grained probes to determine student proficiencies, as well as examining the point-like and set-like paradigm, specifically relating to data processing [154]. This work found that students face challenges with specific areas of measurement uncertainty, such as using the mean as the best approximation in a repeated measurement experiment. On this open-response assessment, students were often unable to articulate why the mean might be used beyond providing a definition of the mean.

Some PMQ probes were also recently used in exploring upper-division students' views of measurement and uncertainty at multiple institutions in the United States [222]. They found that while both introductory and advanced students frequently used set-like reasoning, advanced

students often provided more sophisticated reasoning, especially on a question pertaining to two sets of data with different means and the same spread. Advanced students were more likely to correctly identify sources of uncertainty in this situation. Overall, very few students at any level discussed uncertainty as an inherent property of experimentation. The researchers conclude that, while advanced students do perform better than introductory students on some of the probes, there is still a significant amount of improvement that could be made in helping students master certain concepts in measurement uncertainty, especially related to the shapes of data distributions.

Another RBAI, the Laboratory Data Analysis Instrument (LDAI), was developed in Israel to assess first-year students' understanding of data analysis procedures. It consists of 30 multiple choice and true/false questions that are contextualized in real laboratory reports. The LDAI requires students to write an open-response explanation to accompany their choice on true/false questions in order to receive credit, which increases the difficulty for widespread administration due to the open-response nature of the assessment. The four objectives of this assessment are that students should (1) understand the meaning of, and ways to calculate, measures of central tendency, (2) understand the meaning of error and uncertainty, as well as how to compute this and distinguish between statistical and systematic uncertainties, (3) be able to choose and decipher graphs, and (4) understand regression lines and how to fit them [73].

One implementation of the LDAI in Thailand found that introductory physics students, in particular, faced challenges in fully understanding uncertainty, while undergraduate students of all levels struggled with linear regressions, even after taking at least one laboratory course. However, first-year students performed significantly worse on this assessment than second and third year students, indicating that instruction does improve skills to some extent since first-year students have not had as much learning experience with data analysis [118].

Another important RBAI in the laboratory space is the Physics Laboratory Inventory of Critical Thinking (PLIC), a 10-question assessment designed to examine student learning in physics lab courses [240, 241]. The PLIC is aimed at analyzing students' laboratory skills as a whole rather than focusing on measurement uncertainty specifically. It examines four skills: evaluating

data, evaluating methods, evaluating conclusions, and proposing next steps. The PLIC shows that students have not fully mastered measurement uncertainty, including conflating systematic error, random uncertainty, and human mistakes [241]; this broad study includes matched pre-post responses from several thousand students at 29 institutions and includes both first-year and beyond-first-year courses. In a large-scale administration of the PLIC, there were no observed statistically significant shifts in performance from pre- to post-instruction. However, students enrolled in a lab course specifically designed to teach skills measured by the PLIC do show statistically significant improvements on this assessment [240], showing the importance of aligning laboratory instruction with the desired learning goals.

Finally, the Concise Data Processing Assessment (CDPA) was also developed to probe student ideas related to measurement uncertainty, focusing mainly on error propagation [58]. Research using this assessment has found curriculum-dependent student challenges dealing with measurement uncertainty. For example, many students excelled at questions involving measurement error in linear fits, but struggled with questions involving power laws; an examination of the curriculum for the course in which students were surveyed noted an emphasis on the former and no instruction on the latter [136].

Further, the CDPA was used to investigate gender gaps in physics [59]. The CDPA did reveal a significant gender gap at both the pre- and post-test level. While all students did improve on the CDPA post-instruction, the gender gap remains unchanged: men still outperform women. The authors posit that one reason women do worse on the CDPA is due to a lack of confidence due to previous work showing that women generally report lower confidence in themselves in terms of their physics knowledge than their male counterparts.

Research about students' understanding of measurement uncertainty also exists outside of the space of RBAIs. One study found that students enter university courses frequently believing they must take exactly three measurements. Post-instruction, this belief was often corrected in that students understood that three trials may not always be sufficient. However, many students did not improve from pre- to post-instruction in other areas, including an understanding of the

importance of reporting uncertainties and using uncertainty to determine whether measurements agree with one another [139].

Another study determined that, even post-instruction, students tend to establish a hierarchy of measurements and do not fully understand the need to take several measurements. Instead, students judge their first measurement as the most important and use subsequent measurements as a check of their first one. They are also unable to distinguish between random and systematic errors. Students tend to state that the more measurements that a person makes, the better the result is, without fully understanding how or why more data is better [225].

Most of this prior work examines student proficiencies with measurement uncertainty in the context of non-quantum courses, but the performance of students in quantum courses is also important to investigate and is the subject of several recent papers. For example, one collaboration between researchers at Cornell and California State University Fullerton looked at student responses to questions about measurement uncertainty in both classical and quantum contexts and found that an updated definition of the point-like and set-like paradigm might help advance understanding of student views on this topic, especially due to the binary nature of this paradigm and the prominence of mixed-reasoning amongst students. This research also indicates that instructors need to clarify the meaning behind "more data is better" so that students can understand when, exactly, this is true. The researchers also noted in a similar vein that a clearer discussion of standard deviation and standard error might help students with differentiating these quantities. Further, they found that students often conflate quantum uncertainty (e.g., the Heisenberg uncertainty principle) with measurement uncertainty in quantum mechanical experiments and, therefore, care should be taken in advanced laboratory courses to help students distinguish between these concepts [206].

Other work related to this collaboration has shown that in classical physics, students often state the limitations of the experimental setup as the major cause of uncertainty, while in quantum mechanics, students often explained measurement uncertainty as related to the principles of the physics theory underlying the experiment, as well as statistics [221]. The researchers concluded that there is a split in student reasoning about classical and quantum experiments, and instruc-

tors should work to bridge this gap by providing additional instruction about the relationships between experiment, measurement uncertainty, and theory in courses at all levels, especially because statistical limitations and experimental setup limitations affect both quantum and classical experiments.

Other research has highlighted the benefits of using the term "uncertainty" instead of "error" when describing measurement variability. They posit that using "error" might be a cause of students' point-like reasoning, as it has a connotation of making mistakes rather than uncertainty as an inherent aspect of measurements [41]. Further, using uncertainty to describe inherent limitations and random variability, systematic effects to describe assumptions or approximations, and measurement mistakes to describe actual human errors might further aid student understanding of these concepts [114].

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

6.2.2   SPRUCE

6.2.2.1   General Overview and Development

SPRUCE is an RBAI centered around measurement uncertainty, and was designed to be administered pre- and post-instruction. Previous work has commented on the development, format [184, 235], and validation (see Chapter 4) of SPRUCE, though a brief summary is contained below.

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

---

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

SPRUCE was developed using an adaptation of evidence-centered design [169], beginning with the researchers conducting interviews with introductory laboratory instructors to determine which areas of measurement uncertainty they find important. Based on these interviews, we created assessment objectives for SPRUCE [184]. Assessment objectives, or AOs, are *"concise, specific articulations of measurable desired student performances regarding concepts and/or practices targeted by the assessment* [236]*:"* they are statements that are easy to directly assess in such a survey.

Table 6.1: SPRUCE assessment objectives, organized by assessment objective category.

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

These SPRUCE AOs were then refined during the process of writing and revising SPRUCE itself. Table 6.1 shows the final AOs for SPRUCE after iteration and refinement. They are divided into three categories – sources of uncertainty, handling of uncertainty, and distributions and repeated measurements – and cover a wide variety of measurement uncertainty concepts, while still maintaining a cohesive thematic structure to be able to target them in one assessment. All items (i.e., questions) on SPRUCE probe at least one of these AOs. More details about the validity of these AOs in relation to SPRUCE items are covered in Chapter 4.

6.2.2.2   Scoring

SPRUCE is scored using couplet scoring, a scoring scheme discussed at length in Chapter 4. Briefly, this scheme first identifies which AOs an item aims to measure. It then scores the responses to the item based on that AO only. The score for that one AO on one item is called a item-AO couplet. Items may address one or more AOs and thus have one or more scored item-AO couplets. Items that addresses multiple AOs will be scored multiple times, and the method of assigning points based on students' responses might differ for each couplet.

An example item and scoring scheme are shown in Figure 6.1 and Tab. 6.2, respectively. In this item, we address two different AOs on SPRUCE: *H1 - Propagate uncertainties using formulas* and *H2 - Report results with uncertainties and correct significant digits*. Students need to answer this multiple choice item only once, but we are able to draw conclusions about proficiencies along two different axes (i.e., AOs) from their answers. The scoring scheme itself is provided in Tab. 6.2. This example illustrates how, for one item, multiple answers might be scored as correct depending on the AO and what answer is considered correct depends on what AO is being scored.

> You and your lab mates decide to measure 20 oscillations at a time. Using a handheld digital stopwatch, you measure a time of 28.42 seconds for 20 oscillations. You estimate the uncertainty in your measurement of 20 oscillations to be 0.4 seconds, based on an online search for human reaction time. What value and uncertainty do you report for the period of **a single oscillation**?
>
> ○ 1.421 ± 0.02 s   ○ 1.42 ± 0.02 s   ○ 1.4 ± 0.02 s
> ○ 1.421 ± 0.4 s   ○ 1.42 ± 0.4 s   ○ 1.4 ± 0.4 s

Figure 6.1: SPRUCE item 3.3 (with alternate numbers to protect test security), in which students are attempting to determine the period of oscillation for a mass hanging vertically from a spring. This single item addresses two AOs, H1 and H2, which handle error propagation and significant figures, respectively.

For AO H1, the answers that are given credit are those where students have appropriately propagated error, in this case dividing by 20. Thus, options A, C, and E present choices where students have shown proficiency in error propagation and receive credit for couplet item 3.3 - AO

Table 6.2: Example scoring for couplets of item 3.3, showing how one multiple choice item results in information about two separate measurement uncertainty topics based on the different answers students might give.

Answer Option		Score	
		H1	H2
A	1.412 ± 0.02 s	1	0
B	1.412 ± 0.4 s	0	0
C	1.41 ± 0.02 s	1	1
D	1.41 ± 0.4 s	0	0
E	1.4 ± 0.02 s	1	0
F	1.4 ± 0.4 s	0	1

H1. On the other hand, for AO H2, the answers that are given credit are those where student have provided an answer with correct significant figures. In this case, the answer options with matching decimal places in the result and uncertainty are options C and F, so students selecting either of those would receive credit for couplet item 3.3 - AO H2.

Students only answer this item once. If they pick the "correct" overall answer, which is option C, they would receive credit on both couplets. However, they can receive credit on one couplet, but not the other by providing other answers, or they receive no credit if they select answer options B or D. In this way, we can separate student proficiencies in two different areas of measurement uncertainty by scoring along these axes to obtain information about them separately. All items in SPRUCE are scored according to these conventions, by first aligning the items with AOs and then scoring items as couplets. This leads to 31 item-AO couplets scored on SPRUCE from its 19 items. These couplets are then treated similarly to conventional item scores on a traditional assessment, in that they form the base unit of scoring.

After all of the couplets are scored, we combine them to create ten different AO scores: one score for each AO on SPRUCE. These AO scores are obtained by simply adding up all of the couplet scores pertaining to each AO for each student. We then round these scores to the nearest integer using typical rounding conventions (i.e., 0.5 rounds up), as some couplets allow for partial credit. (See Chapter 4 for justification). These integers scores are reported as the AO-level scores; in some cases, we normalize these to one by dividing by the number of couplets in each AO

for easier comparisons. Typically, in reporting raw scores, these are normalized to one, whereas when we perform other statistical analyses (such as ordinal logistic regression), we keep these as non-normalized integers. Table 6.3 shows the number of couplets and possible scores for each AO on SPRUCE; this table also shows that several couplets on SPRUCE allow for partial credit in increments of 0.25, rather than simply 0 or 1 as scores.

Table 6.3: AO Couplets and Score Options. Each AO is targeted by different numbers of couplets, and therefore has different total possible scores. Some AOs offer partial credit, which is then rounded to the nearest integer after summing all couplet scores for that AO, such that all final AO scores are integers.

	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]

In order to calculate one overall test score on SPRUCE, we add the normalized AO scores together to produce a score out of 10. This is then normalized to a score out of 1. Although the AO scores provide more fine-grained information than one single overall score, we still provide an overall score as a measure of student proficiency in measurement uncertainty as a whole, which is helpful for instructors and interesting from a research perspective. We also used this overall score in validating SPRUCE via classical test theory (see Chapter 4). This method of calculating the overall score (using the normalized AO scores) weights each AO equally, rather than weighting each couplet equally, in order to remove biases from some AOs that are sampled more than others. By weighting each AO equally, we produce a final score that accounts equally for all ten areas of measurement uncertainty and is therefore a good measure of overall student proficiency with measurement uncertainty and is also consistent with instructor expectations.

6.3   Methods

6.3.1   Data Collection and Cleaning

We collected data from 31 physics laboratory courses at 20 institutions in the United States during the Spring 2023 and Fall 2023 semesters (see Tab. 6.5 for details on these institutions). Of the courses, 23 were introductory (accounting for $1{,}379/1{,}576 = 87.5\%$ student responses) and eight were beyond introductory (accounting for $197/1{,}576 = 12.5\%$ student responses). Courses were solicited via the authors' contacts, as well as through posting advertisements on the Advanced Laboratory Physics Association (ALPhA) listserv and two American Physical Society (APS) discussion boards (Forum on Education and Topical Group on Physics Education Research). Student demographics, including gender, race, and major, are presented in Tab. 6.4. We note that these demographics, which represent the 1,576 matched pre-post responses, are representative of the full sample of completed post-test responses.

We collected 3,733 total pre-test responses and 2,710 total post-test responses for a total of 6,443 total responses. We then removed responses based on the following conditions. First, students who did not consent to having their data used for research were excluded, resulting in a loss of 691 responses (10.7%). Second, students who either did not answer the filter question (i.e., closed the survey before reaching that question) or answered the filter question incorrectly were excluded; this step removed a total of 1,153 of the 6,443 responses (17.9%). The filter question is placed after three of the four experiments on SPRUCE, ensuring students have answered at least 11 of the 19 items and therefore are scored on at least 21 of the 31 couplets. Finally, in order to examine the impact of instruction, we matched students using their student names and ID numbers to have matched pre-test and post-test responses for students. If students took only one of these (either only the pre-test or only the post-test), their results were excluded. Thus, we present an analysis of 1,576 matched pre-post responses from the two semesters of data collection, or about 48.9% of total responses to SPRUCE in that time frame.

We also conducted student interviews during the Fall 2022 semester while SPRUCE was

in beta testing, and some of these interview data is used in the work presented here. These 27 interviews each lasted approximately one hour. Students were solicited for interviews from courses in which SPRUCE was currently being piloted. During these think-aloud interviews, students took SPRUCE while sharing their screen with the interviewer and were asked to explain their reasoning for each item they responded to. These interviews provided evidence of student reasoning for each answer option, both correct and incorrect [235].

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

6.3.2   Analysis Methods

To answer our first research question regarding students' overall proficiency with measurement uncertainty, we analyze post-instruction data only. We use only matched post-instruction responses to maintain a single student population for the entire chapter, though results with all post-test responses are similar. Here, we examine the student scores on each AO and their overall scores on the assessment. AO scores are calculated by summing the couplet scores for each AO, rounding to the nearest integer, and normalizing to one (here, we normalize the AO-level scores to one to allow easier comparisons between AOs). Finally, the overall score is the sum of these normalized AO scores, also normalized to one. The scoring processes are detailed further in Sec. 6.2.2.2 and in

Table 6.4: Student Demographics: Race, Gender, Year, and Major [N = 1,576]. Because all demographic questions except year in school allow multiple responses and because these questions were optional, the numbers will not add up to 100%.

	Num. Students	Percent Students
Gender		
Man	905	57.4
Woman	614	39.0
Non-binary	52	3.3
Not Listed	9	0.57
Race		
White	1,206	76.5
Asian	248	15.7
Hispanic/Latino	138	8.8
Black	59	3.7
American Indian or Alaska Native	19	1.2
Native Hawaiian or other Pacific Islander	11	0.70
Not Listed	37	2.3
Year in School		
First year	477	30.3
Second year	551	35.0
Third year	320	20.3
Fourth year	170	10.8
Fifth year	31	2.0
Sixth year or beyond	14	0.89
Major		
Engineering	606	38.5
Physics	204	12.9
Biology	184	11.7
Computer Science	127	8.1
Math/Applied Math	99	6.3
Astrophysics	96	6.1
Biochemistry	92	5.8
Chemistry	71	4.5
Engineering Physics	45	2.9
Astronomy	33	2.1
Geology/Geophysics	23	1.5
Physiology	23	1.5
Other Science	172	10.9
Non-science major	41	2.6
Open Option/Undeclared	36	2.3

Table 6.5: Institution Information [N = 20] including highest degree offered and minority-serving status. HSI indicates a Hispanic serving institution and AANAPISI indicates an Asian American and Native American Pacific Islander serving institution.

	Num. Institutions
Highest Degree	
PhD	6
Master's	5
Bachelor's	8
Associate's	1
Minority Serving Status	
HSI	4
AANAPISI	1

Chapter 2.

Statistically, with post-test data, we report normality statistics in the form of the Anderson-Darling test, as well as skewness (the third moment) and kurtosis (the fourth moment). The Anderson-Darling test can detect whether data are normally distributed and, rather than a binary outcome, provides a significance level that gives information about the degree to which the data presented are normal [25, 172]. Skewness is a measure of the asymmetry of a distribution and kurtosis is a measure of the tails of the data compared to a normal distribution.

To answer one component of our second research question, regarding the impact of instruction, we look at the significance of the shifts from pre-test to post-test scores both at the level of the overall score and at the level of each individual AO, with scores calculated as described above. We perform a Wilcoxon signed-rank test [262] to compute this significance. This is a nonparametric test of the null hypothesis that for randomly selected scores from two populations (in this case, the pre-test and the post-test scores are the two different populations), the probability of one being greater than the other is the same as the reverse. It can be considered a nonparametric version of the dependent t-test. In this case, because we are comparing populations that are not expected to be equal (assuming instruction has an impact on student performance on SPRUCE), we anticipate that the null hypothesis will fail — that is, we would expect that the distributions of these two groups are not identical.

In order to determine how much of an impact instruction has, we also utilize Cohen's d as a measure of effect size [51]. Effect size is a measure of the magnitude of the shift, as opposed to the Wilcoxon signed-rank test, which simply indicates whether the shift is statistically significant (as a binary).

6.3.2.1   Analysis of Covariance

Another component of our second research question requires analysis of student performance on SPRUCE overall (using the post-test overall score) and how this correlates with students' pre-test score, major, and gender. To do so, we use a two-way analysis of covariance (ANCOVA), due to the relatively continuous nature of overall scores (as opposed to the small integer-only nature of the non-normalized AO scores). ANCOVA decomposes the dependent variable's variance into a part explained by the covariate, a part explained by the independent variables, and a residual variance [97, 121]. In our case, our dependent variable is post-test overall score, our categorical independent variables are student major and gender, and our covariate is the pre-test overall score. Using ANCOVA, we can explore whether student major or gender is correlated with post-test performance on SPRUCE, while controlling for pre-test performance. We conducted this analysis in both Python and R to validate the results are the same with two different statistics packages, and we find the results to be in agreement with one another. We note that interaction terms might be significant, and these are addressed in more detail in Sec. 6.4.

The general model we implement for ANCOVA is:

$$S_{post} = \beta_0 + \beta_1 S_{pre} + \beta_2 \left(\text{Gender}\right) + \beta_3 \left(\text{Major}\right), \tag{6.1}$$

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

ANCOVA has several assumptions that must be met in order for it to be an appropriate

statistic to use. The details of these assumptions and our data's adherence to them are discussed in App. A.

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

6.3.2.2   Ordinal Logistic Regression

As a final component of addressing our second research question, we examine the correlations between students' major and gender and their AO-level post-test scores, as well as determine the impact of an instructor's reported importance of that AO on these scores, we perform ordinal logistic regression [87]. We take SPRUCE AO scores as ordinal (the scores have a clear order attached to them, as a student who scores a one has a higher performance than a student who scores a zero), but the AO scores are not continuous (i.e., within an AO, only integer scores are possible before normalizing). Total possible scores for each AO are presented in Tab. 6.3. For ease of analysis, we use these rounded, non-normalized integer AO-level scores.

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

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

$$\log\left(\frac{\Pr(S_{post} \leq j)}{\Pr(S_{post} > j)}\right) = \alpha_j + \beta_1 S_{pre} + \beta_2(\text{Gender}) + \beta_3(\text{Major}) + \beta_4(\text{Importance}), \qquad (6.2)$$

where $\Pr(S_{post} \leq j)$ is the cumulative probability that the single AO post-test score is either j or lower (where j is an integer score on that AO), $\Pr(S_{post} > j)$ is the cumulative probability of the score being higher than j, α_j is the y-intercept for score integer j, β_1 is the coefficient for the pre-test score on that particular AO, β_2 is the coefficient for students' gender, β_3 is the coefficient for students' major, and β_4 is the coefficient for the importance variable (i.e., how important instructors ranked that particular AO on a Likert scale). We note that interaction terms were considered in our overall analysis, and these terms are addressed in more detail in Sec. 6.4.

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

One positive aspect of logistic regression, as compared with linear regression, is that logistic regression is a nonparametric technique, meaning that there are no assumptions necessary about the underlying distribution of the data. Not only does this mean that the data do not need to be normal, but it also means that we do not require homoscedasticity, or constant variance of the residuals in the data [175]. This is because logistic regression uses maximum likelihood estimation (MLE), an

iterative procedure to find the solution, instead of ordinary least squares (OLS) regression. MLE maximizes the likelihood that individual students have scores given by the dependent variable (in this case, post-test scores) based on their scores on the predictor variables (in this case, pre-test scores, major, gender, and the level of importance their instructor places on that AO). Logistic regression does, however, have several assumptions that must be met in order for it to be applied to our data. We discuss these assumptions as well as the adherence of our data to them in detail in App. B.

6.4 Results and Discussion

6.4.1 Overall Student Proficiency with Measurement Uncertainty

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

The mean overall post-test score on SPRUCE, as calculated based on methods described in Section 6.2.2.2 above, is 0.523 ± 0.005 with a standard deviation of 0.189. A histogram showing the distribution of overall post-test scores is shown in Figure 6.2, where the scores have been normalized to 100 in this case only for ease of understanding. This distribution is indicative that our data are visually normal. To quantify this, we perform an Anderson-Darling test for normality and find the post-test scores are normal to a significance level of 1.0% with a skewness of -0.49 ± 0.121 and a kurtosis of -0.55 ± 0.12, indicating normal data. For both skewness and kurtosis, values between -1 and 1 generally indicate normality [98] [2] . Similar figures showing the distributions for all 10 AO scores are shown in Fig.6.3. We also present the average score on each AO in Tab. 6.6; note that these are normalized to one, with uncertainty presented as the standard error.

These data help demonstrate student proficiency with measurement uncertainty. For exam-

---

[2] note that this reference uses a measure of kurtosis which 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
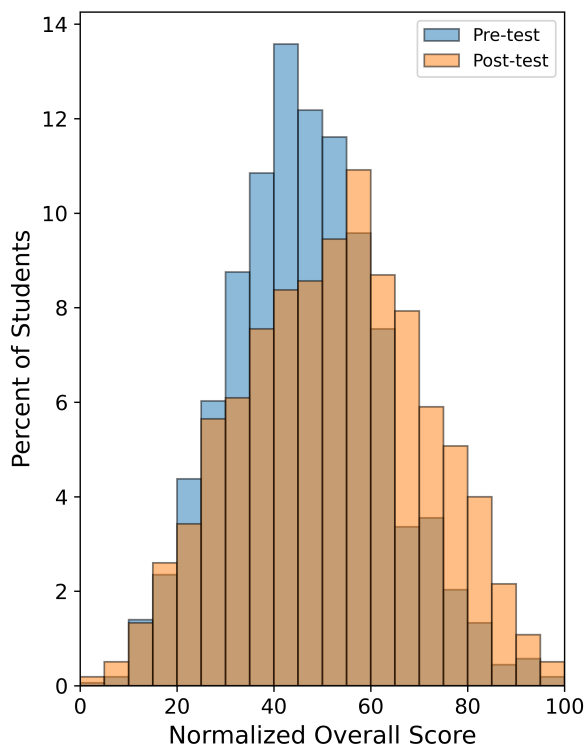
Figure 6.2: Pre-test (blue) and post-test (orange) overall scores on SPRUCE normalized to 100. In aggregate, students improve from pre-test to post-test, as can be seen by the clear shift in the histogram. The distributions themselves are considered normal, with skewness and kurtosis levels for both pre- and post-test distributions well within the limits of normality and and Anderson-Darling tests showing that both distributions are normal to a significance level of 1.0%. The ranges of scores show that SPRUCE does not suffer from ceiling or floor effects in overall score.

ple, students tend to do well at *D3: Calculate and report the mean of a distribution for the best estimate of the measurement*, which is the only AO with an average post-test score greater than 70%. On the other hand, students are less successful on AOs *D5: Determine if two measurements (with uncertainty) agree with each other*, *H1: Propagate uncertainties using formulas*, and *H2: Report results with uncertainties and correct significant digits*, which are perhaps areas instructors might focus on for improvement. All three of these AOs have post-instruction scores of less than 40%.
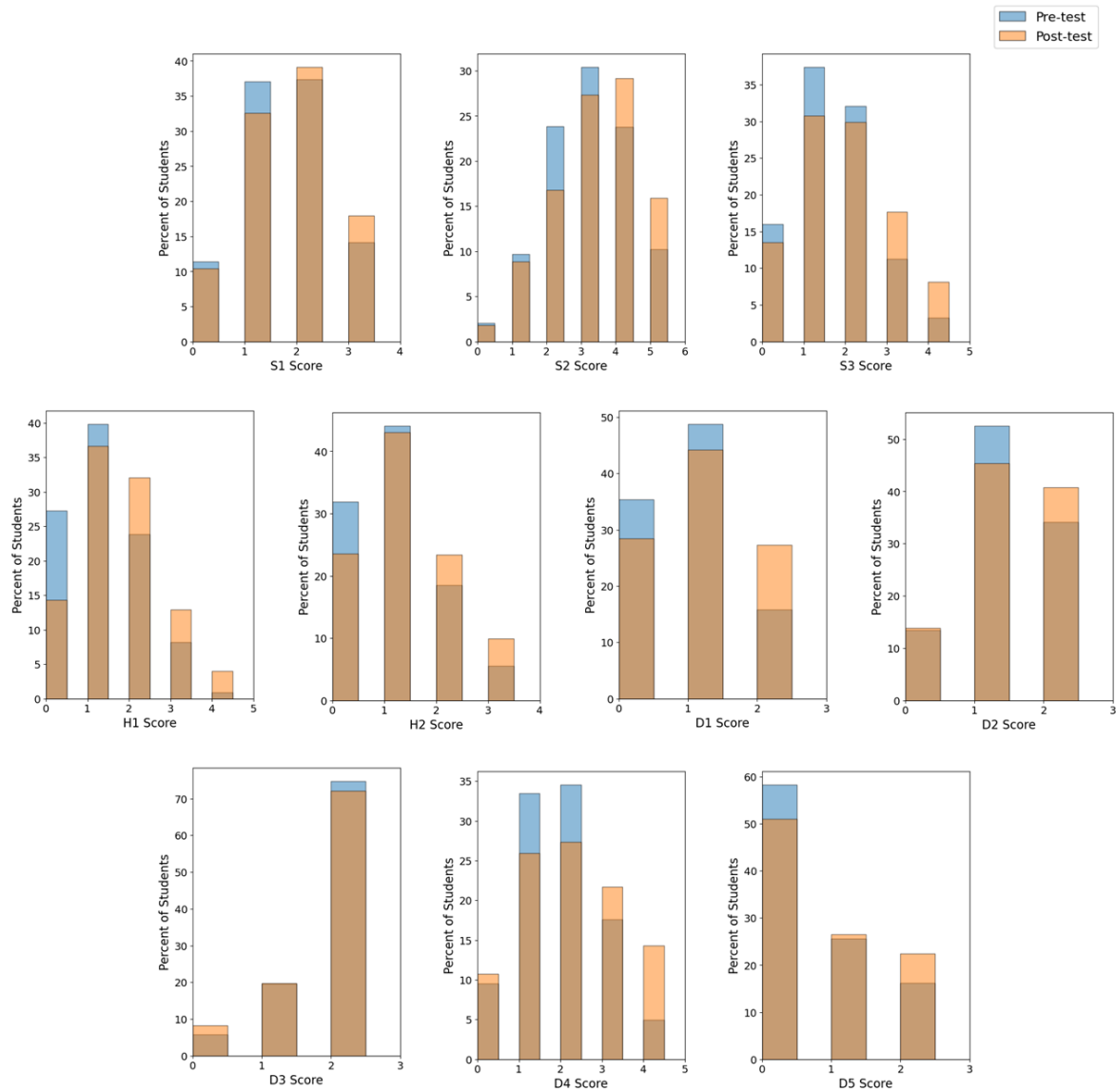
Figure 6.3: Distribution of pre-test (blue) and post-test (orange) scores for each of the 10 assessment objectives on SPRUCE. These distributions are presented after rounding but before normalizing scores to one. Due to the scoring scheme used, these AO scores can only be integers (see Table 6.3) and therefore, these distributions are not continuous.

Table 6.6: AO Average Scores, pre-test and post-test [N = 1,576], normalized to 100. Error presented is standard error, shown as uncertainty in the last digit (e.g., 51.4(7) = 54.9 ± 0.7). D3, which relates to students reporting the mean, has the highest score for both the pre-test and post-test, and likely exhibits ceiling effects. H1 (error propagation) and D5 (measurement comparison) have the lowest scores, and thus represent student proficiencies that have substantial room for improvement.

	Average Score, Pre-Test	Average Score, Post-Test
S1	51.4(7)	54.9(7)
S2	59.0(6)	62.5(5)
S3	37.1(6)	41.6(6)
H1	28.9(6)	38.9(6)
H2	32.6(7)	39.9(7)
D1	40.2(9)	45.8(8)
D2	60.3(8)	56.0(7)
D3	84.5(7)	81.9(7)
D4	43.8(6)	50.7(6)
D5	29.0(9)	35.7(8)
Overall	46.7(4)	52.3(5)

6.4.2   Impact of Instruction

In this section, we explore the answer to the second research question, relating to the impact of instruction on student proficiency with measurement uncertainty. We first examine the significance of the shifts from pre-test to post-test both at the overall score level and at the AO level. We then examine the correlations gender and major have with post-test score using ANCOVA (overall test) and ordinal logistic regression (AO-level). Additionally, we examine the impact of the importance an instructor places on a specific AO on student performance on that AO using ordinal logistic regression.

As shown in the pre-test and post-test distributions in Fig. 6.2, there is a clear shift of overall SPRUCE scores from pre- to post-instruction. We can quantify the significance of this shift using the Wilcoxon signed-rank test, as described in Sec. 6.3.2, and find that the pre-post shift is significant at p ≪ 0.0001 with an effect size of $d = 0.33 \pm 0.04$.

Pre- and post-test scores along with the effect sizes (Cohen's d) of the shifts for each AO are shown in Fig. 6.4. Again, the Wilcoxon signed-rank test shows that all of the pre-post shifts at the

AO level are significant with p ≪ 0.0001 aside from AOs D2 (which is significant at p = 0.0009) and D3 (which is significant at p = 0.004).

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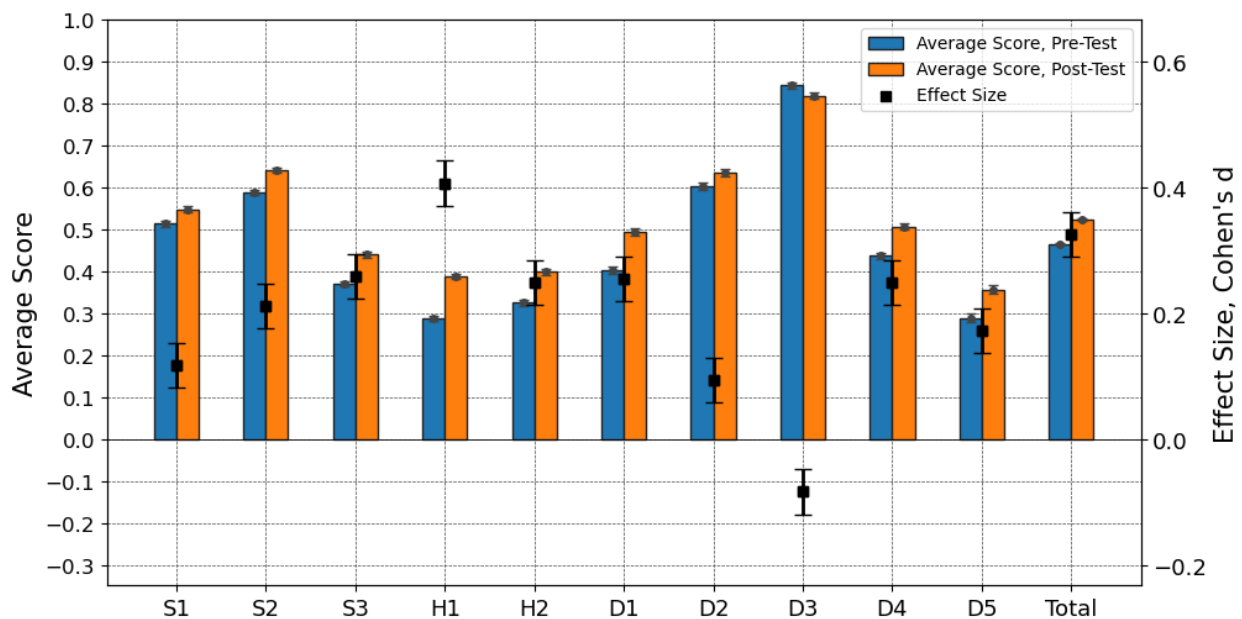

Figure 6.4: AO scores and overall score for both pre- and post-test [N = 1,576]. Error bars on the scores represent the standard error. Effect size (calculated via Cohen's d) is shown with black squares using the scale on the right, with error bars representing the standard error. The pre-post shifts were significant for all AO scores and for overall scores (as determined via the Wilcoxon signed-rank test), with varying effect sizes for these shifts. All shifts were positive aside from AO D3, which exhibits ceiling effects.

6.4.2.1 Impact of Instruction on Overall Post-Test Score: Correlation with Major and Gender

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

Only students who responded to both demographic questions about major and gender and selected only man or only woman for gender were included in this analysis, which resulted in N = 1,503 responses from the original N = 1576. A breakdown of number of students by gender and major is shown in Tab. 6.7.

Based on previous work on analysis of laboratory assessments, We split student majors into the following categories [239]:

(1) Physics, engineering physics, astrophysics

(2) Other engineering

(3) Other science and math (including astronomy and computer science)

(4) Non-science majors

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

The ordering of the groups must be chosen for this analysis (and in the ordinal logistic regression analysis discussed later). In both analyses, we order the majors as follows: non-science, other science/math, other engineering, and physics/engineering physics/astrophysics, as is done in other assessment analysis (e.g., PLIC [239]).

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

Table 6.7: Number of students by gender and major [N = 1,503]. The genders indicated are for students who selected only one single gender, whereas the majors indicated might be one of several majors selected by students, but students were placed into only one group based on their major such that no students are double-counted in this table. The numbers add up to 1,503 due to 73 of the 1,576 matched responses either did not include their demographic information or did not select majors compatible with this analysis. Physics includes students majoring in physics, engineering physics, and astrophysics. Other engineering includes all other types of engineering students. Other science includes students majoring in a science not listed above, including chemistry, astronomy, and computer science. Finally, non-science includes all possible majors outside of science.

	Men	Women	Total
Physics	227	113	340
Other Engineering	401	172	573
Other Science / Math	241	292	533
Non-Science	28	29	57
Total	897	606	1,503

Initially, we used the model Post $\sim$ Pre + Gender + Major + Gender $\times$ Major (which indicated a post-test score dependent variable, with independent variables of pre-test score, gender, and major and interaction term gender $\times$ major) , with an interaction term between major and gender included to test for its significance. We find this interaction term to be borderline in its significance (F test, p = 0.046). Because 0.05 is an arbitrary cutoff and this p-value is on the edge, we have chosen to treat it as not significant. If it were significant, we would have to split the data and do six separate ANCOVA analyses for each category (for example, we would need to examine men only in the model Post $\sim$ Pre + Major, and similarly for the other major and gender categories). This would obfuscate the conclusions one can draw from the data. Therefore, we choose to treat the borderline interaction term as not significant and use only the model presented in Equation 6.1.

Results of the ANCOVA analysis are shown in Tab. 6.8. Partial η^2 is an indicator of the effect size of each of these predictors (pre-test, major, and gender) on the post-test score by indicating the amount of variance each explains in the post-test score. A partial η^2 of at least 0.01 indicates a small effect, and anything above 0.06 is at least a medium-strength effect [97]; reported partial η^2 values should be considered a lower bound due to shared variance between the covariate (pre-test score) and independent variables (major and gender), as discussed further in App. A. All p-values in this table are calculated via the F-test, with gender having one degree of freedom and major having three degrees of freedom. Instruction is not accounted for in this model and is likely the cause much of the residual variance.

While pre-test score is a significant predictor of post-test score (both by p-value and by partial η^2), when we control for this, we find that both major and gender are predictors of post-test score. Major is more significant, and accounts for more variance than gender does. However, much of the variance in post-test score is not accounted for by any of the three predictors used in the model, leading us to presume that instruction (which is not included in the model) plays a significant role in post-test scores. Instruction helps students improve their SPRUCE overall scores, indicating some success in increasing students' proficiency in measurement uncertainty.

Table 6.8: ANCOVA results, including p-values and partial η^2, a measure of the amount of variance explained by each of the predictors. We find that pre-test is a significant predictor of post-test score and explains much of the variance in the post-test scores. Similarly, major and gender also are significant predictors of post-test score, but account for less of the variance. All variance not explained by these three predictors must be explained by some other variable not included in the model, such as the impact of instruction.

Predictor	p	partial η^2
Pre-Test	< 0.001	0.360
Gender	0.022	0.003
Major	< 0.001	0.012

6.4.2.2 Impact of Instruction on AO-Level Post-Test Scores: Correlation with Major, Gender, and AO Importance to Instructor

We performed an ordinal logistic regression analysis with pre-test AO score, major, gender, and importance of the AO to the instructor as explanatory variables for the post-test AO score. Again, only students who responded to both the demographic question about major and gender were included in this analysis, with the additional requirement that the instructor for the course responded to the course instructor survey, meaning that N = 1,486 for this analysis (73 of the 1,576 matched responses did not include the gender and major demographic information or did not respond with only man or only woman for their gender and were thus excluded from this analysis, and a further 17 students were enrolled in classes in which the instructor did not respond to the question in the instructor survey regarding importance for each AO).

The model we use is described by Equation 6.2 [3] . In order to examine the impact of interaction terms, we also model the following for each AO:

---

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

$$\log\left(\frac{\Pr(S_{post} \leq j)}{\Pr(S_{post} > j)}\right) = \alpha_j + \beta_1 S_{pre}+$$

$$\beta_2(\text{Gender}) + \beta_3(\text{Major}) + \beta_4(\text{Importance})+$$

$$\beta_5\left(\text{Major} \times \text{Gender}\right) + \beta_6\left(\text{Major} \times S_{pre}\right). \quad (6.3)$$

The variables in this equation are identical to those in Equation 6.2, but we have added extra terms to account for interactions between students' major and gender and students' major and pre-test score. We include only these interaction terms in our model because they are the ones that have reasonable theoretical explanations for correlation. For example, the importance of a specific AO to a course is conceptually distinct from a student's gender. This is true for all other possible interaction terms not included in the model. In terms of a gender and major interaction (which is similarly seen in the ANCOVA analysis), this interaction can be explained conceptually by noting that physics majors are more likely to be men, and physics majors are more likely to do well on SPRUCE. In terms of the major and pre-test interaction, this can conceptually be explained by the fact that physics majors are more likely to do well on SPRUCE even before taking a course in which SPRUCE is administered, likely due to high school preparation or prior coursework in physics.

We report the odds ratios for each of the AOs in Tab. 6.9. It is important to note that odds ratios can not be compared between AOs with different numbers of ordinal levels associated with them. We can, however, compare the odds ratios for AOs with the same number of ordinal levels (e.g., S3 and H1 can have their odds ratios compared, but H1 and H2 can not). An area of nacent research in educational statistics is determining how to compare statistical analyses from groups with different number of levels.

Three AOs have significant interaction terms ($p < 0.01$). AOs S3 and D1 show significant interaction between pre-test and major, and AO H2 shows a significant interaction between gender and major. The odds ratios reported are for the model in Equation 6.2, that is, the model without the interaction term, because odds ratios for the standalone terms in models with interaction terms

Table 6.9: Odds Ratios for Major, Gender, and Importance. Values shown are 95% confidence intervals. We denote significance with an asterisk (*) (i.e., the confidence interval does not cross 1) and † indicates a significant interaction term present in the model. This table is arranged in order of number of couplets (logistic levels) in each AO (two, three, four, five). Yellow cells indicate a positively correlated predictor for that AO and blue cells indicate a negatively correlated predictor for that AO. Pre-test is a significant predictor in all models, whereas gender, major, and importance only play a role in certain AOs. In particular, importance is only significant for on AO (S1) and it is an inverse predictor.

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

do not have meaning due to the collinearity between the standalone and interaction terms. However, significant interaction terms indicate that when interpreting odds ratios, one should take caution to remember that the full effect is not explained by these two variables alone, but rather that one impacts the other.

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

Finally, and most notably, instructor-rated importance of an AO is only a significant predictor of post-test score for AO S1. However, it is actually an *inverse* predictor in this case; that is, instructors that rated this AO as important were more likely to have students perform *worse* on this AO on the post-test compared to students in other courses. This, combined with the lack of

significance on the other AOs, indicates that students are not necessarily achieving the instructor-stated goals for the course with respect to measurement uncertainty. One potential cause of this is that instructors who don't report a particular AO as important still teach those concepts as well as those who do report it as important. Overall, instructors are not having a significant, positive impact on the areas of measurement uncertainty that they deem important as compared to instructors who do not report those areas as important. To further highlight this, the odds ratios for importance are presented in Fig. 6.5. We again note that the odds ratios for AOs with different numbers of couplets should not be directly compared.

6.4.3   AOs of Interest

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

6.4.3.1   AO S1: Estimate size of random/statistical uncertainty by considering instrument precision

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

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

This AO is probed three times by SPRUCE. Two of these items are coupled numerical open response and one is coupled multiple choice. In all cases, students are asked to provide a
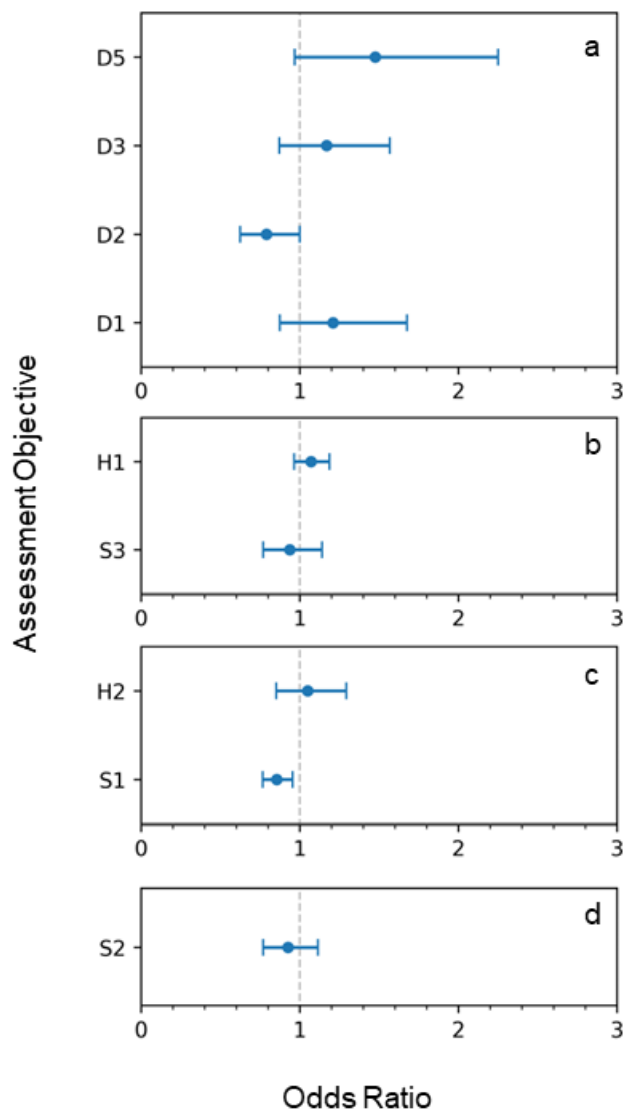
Figure 6.5: Odds ratios for importance, separated by number of couplets per AO - two (a), three (b), four (c), and five (d) couplets. These odds ratios show the impact of instructor-reported importance of an AO on students' final post-test score. Ideally, these odds ratios would be greater than one with 95% confidence. However, none of them statistically significant, indicating that the level of importance placed on an AO by instructors is not correlated with student performance on that AO, with a single exception: S1 scores are slightly negatively correlated with importance, due to its odds ratio lying below one with 95% confidence.

measurement and an uncertainty associated with that measurement based on a specific instrument shown.

One of these items presents students with two graduated cylinders filled with water showing

a "before" and an "after" measurement in order to determine the volume of an irregular object. Students report the volume shown in both of these circumstances, as well as the uncertainty in the measurement for both. Importantly, it is the same graduated cylinder in both measurements, with measurement markings every 100 milliliters.

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

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

This indicates that they believe all graduated cylinders have the same uncertainty, regardless of measurement markings on the instrument. This is directly opposed to AO S1, which requires students to understand that the precision of their measurements is directly related to the precision of the specific instrument they are using. In this case, 0.05 milliliters is much too precise for a graduated cylinder markings every 100 milliliters.

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

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

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

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

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

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

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

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

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

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

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

including digital instruments, can aid in determining the uncertainty of the measurement can potentially help students overcome these challenges.

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

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

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

During the interview phase, we found a common response to questions of collecting more data was simply a blanket statement about having more data being the best practice without providing reasoning as to why it is better to have more data, similarly to results from prior studies [206, 225]. For example, one student said:

I've always understood that it's best practice to take measurements multiple times in experi-ments.

Another common line of reasoning for students is the desire to take several measurements

because of human error. For example, one student stated:

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

Again, this student has a similar understanding that taking more data is better, and provides minimal evidence as to why, while seeming to employ point-like reasoning of taking multiple measurements to ensure reproduction of the "true value". This student specifically quotes three measurements, similar to findings in prior studies [139]. Further, this quote exemplifies the issues regarding using "error" as opposed to "uncertainty" when describing random and systematic effects, as previously explored [41, 114]. Both of these students left off key parts of a fully correct response to items probing this AO. For example, in this second case, the student did not select answer options about calculating the uncertainty and reducing the impact of outliers as reasons for collecting more data.

Another student discussed the "true value" of a measurement:

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

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

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

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

This AO was previously examined in Chapter 5 and is expanded upon here with a more complete dataset and using only matched pre-post responses. It is probed by two isomorphic items that are shown in Figure 6.6. The first asks students whether their measurement agrees with other groups' measurements using a numeric representation (NRI) while the second asks students the same question using a pictorial representation (PRI).

While students generally perform poorly on both of these items, they tend to perform better on the pictorial version despite the items being identical in content. The post-test mean for AO D5 is 0.357 ± 0.008, the lowest score of all AOs. Of the 1,576 post-test responses to these items, 354 students (23 ± 2 %) answered both items correctly, while 103 students ($7 \pm 1\%$) correctly answered only the NRI, 315 students ($20 \pm 2\%$) correctly answered only the PRI, and 804 students (51 ± 3 %) answered both items incorrectly. Further, this AO shows some improvement from pre-test to post-test, but this improvement was small (effect size $d = 0.17 \pm 0.04$, pre-test mean $= 0.290 \pm 0.009$).

Figure 6.7 shows a heat map of the 905 most common student responses on the post-test for both the NRI and the PRI. One might expect responses to occur only along the diagonal, indicating students who selected the same answer combination for both the NRI and PRI, but this is not the case. Instead, students frequently select different answers to these items, indicating a need for further instruction in measurement comparison. Of the 1,576 total responses, only 433, or about 27% of students, selected the same answer combination for these items (whether a correct

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

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

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

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

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

☐ (A) $3.71\frac{\text{N}}{\text{m}} \pm 0.06\frac{\text{N}}{\text{m}}$ ☐ (E) $3.91\frac{\text{N}}{\text{m}} \pm 0.06\frac{\text{N}}{\text{m}}$

☐ (B) $3.71\frac{\text{N}}{\text{m}} \pm 0.17\frac{\text{N}}{\text{m}}$ ☐ (F) $3.91\frac{\text{N}}{\text{m}} \pm 0.17\frac{\text{N}}{\text{m}}$

☐ (C) $3.76\frac{\text{N}}{\text{m}} \pm 0.06\frac{\text{N}}{\text{m}}$ ◯ (G) None of these agree with my data

☐ (D) $3.76\frac{\text{N}}{\text{m}} \pm 0.17\frac{\text{N}}{\text{m}}$

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

☐ (A)

☐ (B)

☐ (C)

☐ (D)

☐ (E)

☐ (F)

◯ (G) None of these agree with my data

Figure 6.6: Two Isomorphic Items on SPRUCE. These items probe student understanding of measurement comparisons with uncertainty by presenting the same data in two different representations - a numerically represented item (NRI) and a pictorially represented item (PRI). The students first encounter the NRI and then, after answering several unrelated items, they encounter the PRI in a different experimental context. Responses of 'ABCD' and 'ABCDF' receive full credit, while no other combinations receive any credit. Note that the answer options on the PRI are in a different order when presented to students (DAEBFCG) than shown here; we present them in the same order as the answer options for the NRI in this dissertation for ease of understanding.
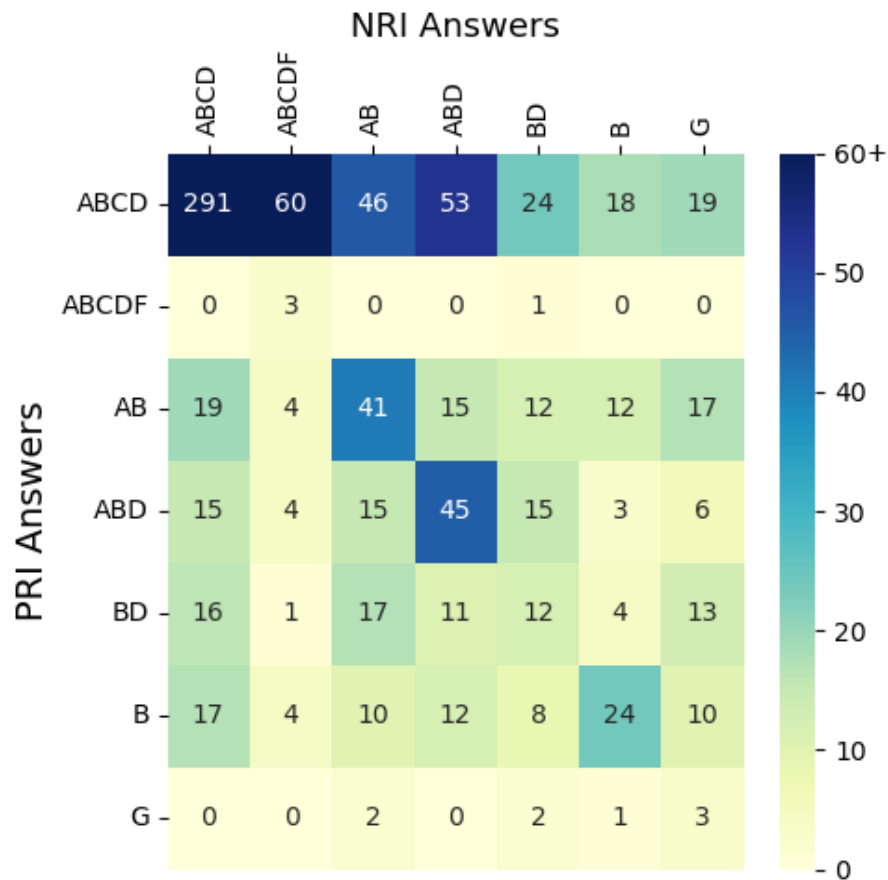
or incorrect combination).

Figure 6.7: Heat map showing the most common 905 of the total 1,576 post-test responses to the NRI and PRI. Both 'ABCD' and 'ABCDF' are accepted as correct responses. Responses along the diagonal indicate students who selected the same answer combination for both the NRI and the PRI, while off-diagonal elements indicate students who selected different answer combinations for these two items.

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

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

This ability in being able to switch between different representations aided this student in correctly

answering the numerically presented item; they also were able to correctly answer the pictorially presented item.

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

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

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

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

For the other four groups... their values did not put them in the same range as my value with its uncertainty so I don't believe they agree with my value.

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

Prior research into students' handling of different representations of the same problem [129–132, 243] shows that representation is very important in student proficiency at problem-solving in all areas of physics, not just in measurement uncertainty. Because many scientific papers present results numerically with uncertainties, being able to compare numeric results between papers is a vital skill for students to learn in experimental physics courses and we encourage instructors to help students learn to switch between different representations to bolster this skill. Further, instruction emphasizing that students should not prioritize their own measurements over others' could again help students become more proficient in this area.

6.5   Summary and Future Research

We have presented an overview of student proficiency in measurement uncertainty, including the impact of instruction and the correlation of students' gender and major with their performance on SPRUCE. We find that instruction does tend to lead to better scores on SPRUCE, both overall and at the individual AO level, though these effects vary between AOs. Further, we find that instructors rating specific areas as important do not correlate with student post-test scores (aside from one case in which this correlation is inverse). Overall, students excel at reporting the mean as a final answer and struggle with comparing measurements with uncertainties, propagating uncertainties using formulas, and correct use of significant figures.We also find that gender is a statistically significant but weak predictor of student performance on SPRUCE. Additionally, it is only correlated with performance on three of the 10 AOs on SPRUCE. Altogether, these results about gender show a promising step towards improving issues associated with gender bias in physics courses and assessments.

Further, from student interview data and the analysis of outcomes on SPRUCE, we present several suggestions for instructors. First, because students struggle with understanding why collecting more than one data point is important, we suggest that instructors emphasize this rather than providing minimum requirements without justification. Next, instructors should note that teaching students both numeric and pictorial representation methods of comparing measurements

with uncertainty, as well as teaching students how to switch between these representations provides students the best tools to properly analyze data. This has been shown in prior research and is apparent from our analysis of identical questions with different representations in SPRUCE. Additionally, instruction on comparing measurements with uncertainty could help bolster students skills in this area, because even after a semester of instruction, students struggle with this concept. Finally, because students sometimes struggle with identifying the precision of a measurement in relation to the instrument used to make that measurement, instructors should be deliberate in their treatment of this topic. This includes using various types of the same instrument with different measurement markings (e.g., rulers with different scales) to show that the specific instrument precision is important rather than treating each type of instrument as being the same. Finally, instruction about digital instrument uncertainty is important, as often students have more confidence in digital scales that the measurement uncertainty would suggest.

In the future, as more SPRUCE data are collected, we plan to perform further analyses about student proficiency in measurement uncertainty. With more data, we can perform more advanced statistical analyses such as a cluster analysis to identify groups of similarly-thinking students within the data [61]. Further, we hope to be able to perform ordinal logistic regression and ANCOVA to examine the correlation between race/ethnicity and student performance (an analysis not presented in this dissertation due to not having enough data from non-white students), as well as including gender minorities in future iterations of this work. We also aim to update the analysis of gender and major correlations with SPRUCE scores presented within this paper.

Additionally, future papers will investigate specific AOs further. For example, we are investigating student ideas surrounding accuracy and precision, as related to AOs S2 and S3 specifically. In addition to items related to accuracy and precision, students are presented with an initial question with four stereotypical bullseye targets and are asked to select which image depicts high precision and low accuracy. From this, we can correlate student performance on other items about accuracy and precision to see whether students understand the difference between these concepts at least in the bulls-eye representation.

In addition to helping to collect more data for these studies, instructors and researchers who are interested in using SPRUCE in their teaching and/or research can visit the SPRUCE website at [12] for more information about how to use it in their own classes and studies.

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

Chapter 7

Development of a global landscape of undergraduate physics laboratory courses

This chapter is adapted from an article submitted to Physical Review Physics Education Research [89]. The collaborators of this work are Micol Alemani, Michael F.J. Fox, P.S.W.M. Logman, Eugenio Tufino, and H.J. Lewandowski.

7.1   Introduction

Physics education is a global endeavour. As we strive to find the best methods of education for the next generation of physics students, undergraduate physics education can benefit from an international perspective, both for improving and comparing courses and to aid students studying worldwide.

In today's world, international collaboration is growing due to more accessible technology and the need to engage scientists across the world to answer important questions and solve critical issues that are global in nature [170]. Thus, physics education should cross the boundaries of countries to form a cohesive structure to enhance education of future scientists. In order to best conduct physics education research to improve education, we first need to understand the similarities and differences in how physics is taught worldwide, including degree requirements, classroom environments, and experiences of students. This will also help future collaborations amongst physicists, as they can better understand their collaborators' previous educational experiences, which could lead to a better appreciation of the variety of backgrounds of participants in a collaboration.

Here, we focus on the context of undergraduate physics laboratory courses due to the signif-

icant current work in this space [16, 43, 110, 159], as well as the importance of physics laboratory courses in general and the unique skills they can provide for students [138]. One important step in the process to improve physics laboratory education is to understand what these courses currently look like globally.

Towards this end, our ultimate goal is to create a taxonomy, or classification scheme, of undergraduate physics laboratory courses that can be applied worldwide. This taxonomy would have numerous applications, including to gather information about courses so that lab instructors and course developers may be inspired by others, as well as to facilitate comparisons that may be made through Physics Education Research (PER) studies. From a research perspective, it is difficult to know whether studies done in certain courses can be compared to others, as lab courses are a rich and complex space with a large variety of implementations. A taxonomy could help classify these courses, so research results can be used appropriately without over-generalization.

A taxonomy could also be used to standardize comparison data that is currently presented in reports to instructors about their courses from research-based assessment instruments (RBAIs) [233, 253], assessments used to determine how well students collectively are meeting course learning goals (as opposed to assessments used to individually evaluate students) [153]. Typically, results from RBAIs for laboratory courses present instructors with both their own students' performance, as well as data from other courses for comparison. Unfortunately, these comparison data usually include data from all students who have taken the assessment, regardless of the type of course it was used in. This makes it difficult to know whether the comparison data is appropriate. A taxonomy could be used to select comparisons data from only similarly characterized classes.

On the practical side, a taxonomy could allow instructors to learn from one another about what they do in order to improve or transform their courses. One could also use a taxonomy of lab courses to learn about student experiences in different systems; this may be helpful in graduate admissions, as students apply to study across the globe coming from undergraduate institutions around the world. Further, a taxonomy scheme could help facilitate international collaborations amongst physicists with different educational backgrounds.

Here, we present the development of a survey designed to collect data that will allow us to create a taxonomy of lab courses with additional data collection. The project goals include gathering input through interviews with lab instructors across the world to both understand the scope of lab instruction, and to gather input to the development of a research-based survey, which aims to capture the structure, goals, and activities in lab classes throughout the world. We present the development of the survey, as well a first look at the similarities and differences of 217 lab courses in 41 countries represented in the initial data collections. Future efforts will work to collect significantly more survey responses to be able to apply clustering methods [79] to create a taxonomy of labs courses that can be used by both instructors and education researchers.

7.2   Background

7.2.1   Prior PER on Laboratory Courses with a Global Perspective

PER in undergraduate physics laboratory courses is less common than research investigating lecture courses, though research focused on laboratory courses is growing in popularity [16]. Though undergraduate laboratory courses are often the only time students might get experience with experimental physics, these courses are frequently overlooked as less important than their lecture counterparts in the curriculum overall [74], as often students complete fewer lab courses than lecture courses. Further, laboratory instruction can be challenging due to personnel limitations, financial constraints, aging equipment, outdated experiments, and lack of advanced courses [81]. Despite these challenges, students have the opportunity to learn valuable skills in laboratory courses — such as hands-on technical skills, experimental design, modeling, troubleshooting, and data analysis techniques — that are often not otherwise covered in the physics curriculum [138]. Because these courses can be critical to the development of students' experimental skills, knowledge, and habits of mind and are often resource-intensive, it is vital to ensure that the courses are meeting their goals with the support of research-based practices and assessments being developed by researchers in PER [81].

Currently, the field of PER in laboratory courses is focused largely in the United States, but is quickly becoming more international. In this section, we highlight some contributions to the PER literature from the community outside of the USA. For example, one paper from Taiwan discusses methods of integrating technology into physics lab courses, including potential options for virtual and remote laboratory instruction, including the creation of a framework for others hoping to use technology to support inquiry-based activities [48]. A recent paper from Finland discusses modernization of a physics lab course at the University of Helsinki, with details about shifting to more open-inquiry activities [137]. Other work from Finland details methods of assessing students work in lab courses, including different types of examination and feedback [122]. Recent work from India discusses shifting an undergraduate electronics lab towards open-ended activities in order to help improve students' research skills, including technical skills, problem-solving abilities, and collaboration [171]. One paper from Germany presents a similar shift from prescriptive laboratory activities towards a skills-based course with more authentic experiments, finding that students are more engaged and are better able to master important laboratory skills, such as keeping a lab notebook [22]. Further, in Italy, a group of researchers examined the transformation of a lab course to include activities with Arduinos and smartphones, including an open-ended aspect [174]. Other work, from the Netherlands, shows a tendency towards open-inquiry lab courses [149, 150, 192].

In addition to these efforts in in-person labs, the COVID-19 pandemic necessitated research into remote laboratory activities. One such paper from the Netherlands described this abrupt transition where they investigated the use of Arduinos with open-inquiry activities [35]. An additional paper from France also details the use of Arduinos in a project-based lab course for third year students, in which they are given complete decision-making control over their experimental setup and what to investigate with the Arduinos in order to help them learn more about the nature of experimental physics [34]. Further work from the Netherlands reports the creation of a Mach-Zehnder interferometer from children's toys with an Arduino detector. This work highlights the ability to achieve experimental physics learning goals without the need for expensive resources [82].

In addition to research from individual countries, collaborations between researchers in dif-

ferent countries have become more common as modern technology allows us to connect with people around the world. For example, a collaboration between universities in Finland and the United States investigated students' abilities in critical thinking over the course of a semester [180]. Another collaboration involving researchers in Germany, Finland, Switzerland, and Croatia examined the use of digital experiments in physics lab courses, including development of a questionnaire in four languages to investigate student use of these online experiments aimed at remote learning [140]. Additionally, researchers in Germany and the United States investigated student views of experimental physics in German lab courses, finding distinct differences between these students and their American counterparts. [227]. Another collaboration between researchers in Denmark, Czechia, and Slovenia explored teacher education regarding lab courses with global input from a discussion at a conference. This work specifically focused on learning goals and the role of labs in teaching physics [27].

These collaborations and international studies can reveal similarities and differences in the ways that lab courses are taught around the world [191], allowing us to challenge our perspectives on how lab courses are taught. However, it is difficult to directly compare research from all of these studies without an understanding of the basic laboratory course structure, goals, and activities at these different institutions. Even within the United States, laboratory courses differ vastly between different colleges and universities [111]; adding an international component to the mix further complicates this. Some prior research has investigated physics lab instruction in North America [111] using course instructor surveys from RBAIs, but this does not include a broader international aspect. Holmes et al. found that physics lab instruction across North America varies considerably in many aspects, including course goals, activities, and pedagogical methods. Another recent paper compared instructional strategies during the pandemic from one university from each of the three countries (United States, Sweden, and Australia) and found that, despite the widely varying locations and cultural constructs, all universities struggled with successfully implementing emergency remote instruction [204].

One class of studies that aims to be broader than just including a few countries is the use

of RBAIs. These assessments are intended for gathering information about students' performance collectively, rather than for assigning individual students grades. They are frequently used to determine whether a course is meeting its learning goals [153]. One such RBAI is the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) [254, 257, 268, 270]. Despite the name, this assessment instrument has been used broadly with greater than 100,000 total responses from many countries (mostly in North America, Europe, and Asia). It has also been translated to several different languages (including Swedish [106], Italian [174], German [227], Chinese, Spanish, Hebrew [143], Norwegian, and Amharic) to allow for easier administration in other countries. Another globally used RBAI, the Physics Lab Inventory of Critical Thinking (PLIC) [114, 194, 240, 241], which focuses on critical thinking in physics labs, including data analysis and measurement uncertainty. This survey is available in Chinese, Finnish [180], German, and Spanish, in addition to English. However, even if the data collection using these RBAIs includes many countries, they did not include input from more than a few at most in the development of the instruments.

7.2.2 Prior work on characterizing laboratory courses

There has been some previous work on characterizing undergraduate physics lab courses that we can build on for the current study.

One effort to characterize laboratory courses is related to the administration of E-CLASS and PLIC. To implement E-CLASS using the centrally administered version in English, instructors complete a course information survey, which gathers information about the course itself, including information about the level (introductory or beyond introductory), whether it is algebra- or calculus-based, the number of students and staff, and how frequently students participate in various activities. A nearly identical set of questions is used in the administration of the English version of PLIC [111]. While this information is useful in many research applications, it is limited in scope and does not provide enough for course classification for worldwide applications [111, 252]. For example, it is missing information about activities, pedagogies, and course design. Some questions

from the course information surveys for E-CLASS and the PLIC were used as input for the survey discussed in this chapter, including questions about the level of the course (introductory versus beyond introductory), type of institution, and information about course goals [111].

Another effort we draw from was initiated during the COVID-19 pandemic, where research was conducted regarding the switch to emergency remote laboratory instruction. Researchers created a survey to collect information from course instructors about the structure, goals, and components, as well as other features of their remote or hybrid lab courses [33, 83, 245]. The pandemic instructor survey included some similar questions to those in the E-CLASS course information survey, but also included questions about changes to course goals and activities (whether they were incorporated in the course both before the pandemic and while the course was remote). While the data were helpful in analyzing differences between instruction pre-pandemic versus during the pandemic, the questions were not extensive enough to fully characterize laboratory courses and were created locally on an extremely short timescale to capture the rapidly evolving situation.

7.2.3 Collaborator Professional Positionality Statements

As the work presented here relies, in part, on the professional experience and networks of the collaborators of this project, I present backgrounds relevant to the work.

Gayle Geschwind is a Ph.D. student currently working in PER on the development, validation, and analysis of a measurement uncertainty focused RBAI for use in undergraduate physics laboratory courses. She has prior experience both as an undergraduate and as a graduate student working in experimental physics; her undergraduate career focused mainly on experimental atomic, molecular, and optical physics while her graduate experimental physics experience is in biophysics. Finally, she has been a teaching assistant in a sophomore-level undergraduate physics lab course at the University of Colorado Boulder. She was not present at the workshop where the initial idea for the taxonomy work originated, but was brought in some time later to work on the project.

Micol Alemani is the laboratory course coordinator at the University of Potsdam, which she completely redesigned using research based results. Her physics background is in solid state physics,

but she is now focusing on physics education and its research. During her PhD she built a low temperature scanning tunnelling microscope and use it to investigate the manipulation of isolated molecules on surfaces. During her postdoc, she studied electrical transport properties of graphene. She worked as an experimental physicists both in US and Germany and thought lab courses in both these countries. As a student she did her lab courses in Italy. She is an active member of the working group about undergraduate laboratory courses called Arbeitsgruppe Physikalische Praktika (AGPP) [4] of the German Physical Society (DPG) and since 2021 serves as board member for the physics education section of the DPG.

Michael F.J. Fox has been the second-year laboratory course coordinator at Imperial College London since 2023. His research is focused on student learning in undergraduate teaching laboratories and equity in physics. Previously, he completed a PhD in theoretical plasma physics, taught high-school physics in London for 3 years, and worked as a postdoctoral researcher with H. J. L. in PER related to laboratory courses completing work on the quantum industry [85], as well as analysis of E-CLASS data [83] and development of the MAPLE survey [84].

Paul Logman is the laboratory course coordinator at the Leiden Institute of Physics in The Netherlands. His physics background is in applied physics, but since 2009 he is focusing on physics education. From 1993-2014 he worked as a high school physics teacher. In 2014 he finished his PhD in physics education in which he used educational design research to develop a teaching-learning sequence on the general law of energy conservation in which students reinvent that law by performing various experiments. After finalizing his PhD he started work in Leiden where he has been redesgining the undergraduate lab courses since 2017 using educational design research. He is a board member of the Groupe International de Recherche sur l'Enseignement de la Physique (GIREP - International Research Group on Physics Teaching) [6] and co-leader of the GIREP Thematic Group on Laboratory Based Teaching in Physics (LabTiP) [5].

Eugenio Tufino is a Ph.D. student actively involved in PER. His research focuses on the integration of active learning methods in both introductory university physics courses and high school settings (in particular, using the ISLE approach). With many years of experience as a high

school physics teacher, he has developed a keen interest in the use of digital technologies to enhance physics learning. He has been very involved in the introduction of the E-CLASS assessment tool in Italy. E.T. attended the workshop in April 2022, where he presented the first results of the implementation of E-CLASS in Italy.

H.J. Lewandowski is a physics professor who runs two research groups, one in physics education research and one in experimental chemical physics. Her PhD was in the field of experimental atomic physics (Bose-Einstein Condensation), and her postdoc was in experimental molecular physics (Cold molecule spectroscopy). She has over 25 years experience designing, constructing, and using table-top experimental apparatus. Her work in PER began in 2010 and has focused mostly on laboratory courses at the undergraduate level. Besides the current work, she has had many international collaborations in PER, including with researchers from China, Oman, South Africa, Denmark, Norway, United Kingdom, Italy, and Germany. Additionally, she has taught all of the undergraduate physics laboratory courses at the University of Colorado numerous times. She also led efforts to transform three of these courses through research-based practices. She has served on the Board of Directors of the Advanced Laboratory Physics Association (ALPhA) [1] for 11 years, including two years as President of the organization. She was present at the workshop in April 2022 that initiated this project.

7.3   Methods

7.3.1   Survey Creation

7.3.1.1   Initial Creation of Ideas and Organization

The project idea emerged from a workshop at Imperial College London where the collaborators met to discuss international comparisons of teaching labs using the E-CLASS survey. The first step towards development of the survey was a collective brainstorming session over Zoom, where we discussed the most important aspects of undergraduate physics laboratory courses and created a virtual whiteboard to collect and partially organize the ideas. Many of the ideas organized on the

whiteboard were based on the Spinnenweb (spiderweb) representation of curriculum and learning proposed by van den Akker [232]. This model examines why students learn in different facets. A large fraction of the survey was modeled off this structure, including sections such as grouping, assessment, goals, activities, content, instructional staff roles, and materials and resources.

After this initial process, the results of the whiteboard activity were loosely organized into a document by category. This document included many of the ideas that eventually went into the survey, but was missing several important concepts and included many items that we decided to remove. For example, a list of equipment that students might have access to in the course was removed for being too unwieldy for both survey participants and researchers to handle. We also added questions from previous E-CLASS/PLIC and pandemic instructor surveys.

This list of lab components, along with feedback provided by the collaborators and external physics education researchers, was transformed into a survey format. This involved organizing the information we wished to probe into meaningful categories, ordering those categories in an intuitive way, and turning ideas we wanted to probe into questions. All collaborators iterated on the survey several times until a cohesive final product emerged. Many of the iterations involved language changes, with interviews (as detailed below) later validating the wording choices made for the survey. At this stage, the survey document was coded into a Qualtrics survey in order to prepare for the interview validation phase of survey development.

7.3.1.2 Interviews with Lab Instructors

We conducted interviews with lab instructors from around the world. The purpose of the interviews was to (1) make sure the questions were interpreted as intended, especially for people from countries where English is not the primary language, (2) to make sure the answer choices well-represented lab courses around the world, and (3) to see if the questions represented the types of information the lab instructors felt was important to capture.

Interviews were solicited from contacts known by the collaborators. These contacts were compiled into a list and solicitations were chosen such that only one person per country would be

solicited with a reasonable worldwide spread of country participants. We chose to exclude Germany, Italy, the Netherlands, and the United Kingdom from interviews due to already having researchers on our team from these countries. We did include the United States in interviews due to the diversity in universities in this country. We conducted interviews only with those who currently teach or have previously taught undergraduate physics laboratory courses.

In total, we conducted 23 interviews with participants from 22 countries; the United States was sampled twice. A map of the countries participating in interviews is shown in Fig. 7.1. Interviews took place with instructors from the following countries: Australia, Brazil, Canada, Chile, China, Colombia, Finland, France, Georgia, Greece, India, Indonesia, Ireland, Israel, Kenya, Norway, Oman, Pakistan, Poland, South Africa, South Korea, and the United States. Each country had one instructor interview except for the United States, which had two (the first was to check the interview protocol as well as collect data; the USA is varied enough to warrant two interviews). Because our authors are from Germany, Italy, England, and the Netherlands, we specifically did not contact instructors in these countries for interviews because our authors could provide the necessary information.

We solicited 32 participants from 26 countries in total, leading to approximately a 72% success rate. Interviews were solicited in several rounds, and we determined after 23 interviews that changes to the survey had become minimal and therefore, further interviews would be unnecessary. (We made changes to the survey after nearly every interview and would present the newest version to the next interviewee).

I conducted all of the interviews over Zoom. Each interview lasted between 33 – 76 minutes. Both video and audio of the interviews were recorded for later analysis. Additionally, the interviewer took notes during the interview and, in most cases, these notes provided the basis for further changes to the survey. Interviewees were provided with a link to the survey during the interview and were asked to talk through the questions while sharing their screen. They were instructed to consider all undergraduate physics laboratory courses they had knowledge of when considering whether the questions made sense, as well as whether the questions spanned the space of all knowledge they felt

was important to collect.

The interview protocol was created to validate the survey questions by answering three validation questions:

(1) Are the survey questions understandable to those who are not from a country represented by a collaborator? Are the survey questions interpreted in the way they were intended?

(2) Do the survey questions make sense in the context of courses the instructor has taught?

(3) Do the survey questions fully span the space of information that the instructor feels is important to capture about undergraduate physics laboratories?

For most survey questions, interviewees were asked whether they understand the question, as well as whether anything was missing; in some cases, interviewees were also asked to explain in their own words the concepts and ideas presented in certain questions, especially in the goals and activities sections of the survey.

In terms of the first validation question, we probe whether the (mostly) American English used throughout the survey is understandable both to those who have learned English as a second language, as well as to those who speak English as their primary language, but know a different dialect (British, Canadian, or Australian English, for example). Differences arise due to variations in terms used by British English (e.g.,'revise' means 'study' in British English, but means 'alter' in American English), as well as terms not commonly used for those who speak another language as their primary language (e.g., the term 'rubric' was determined to be unfamiliar to many non-native English speakers, but when the concept was described, interviewees were familiar with it).

To answer the second of the validation questions, interviewees were asked about whether each survey question was applicable in courses familiar to them. In some cases, questions had to be added or logic had to be introduced in order to ensure that people taking the survey would be able to appropriately answer all questions without confusion due to certain questions not being applicable to their courses. For example, in one case, we added a question about whether the lab

course meets weekly. In the case that it does, the instructor is sent to a page with the original questions about number of hours per week the course meets and the number of weeks the course runs. In the case that the course does not meet weekly, instructors are sent to a different set of questions in which they are asked to describe their course meeting (how often, how many hours per term, etc.).

Answering the third validation question helped us to add, remove, and revise questions to ensure we gathered all of the information we need to create a taxonomy and describe the state of undergraduate physics laboratory courses around the world. Interviewees were asked whether questions regarding goals of the course and activities students participate in were complete or missing important relevant information. Many times, interviewees suggested additions that helped to make the survey more accurately capture the full breadth of components of the courses. In addition to ensuring each individual question fully spanned the space of information we hoped to gather about that specific topic being probed, interviewees were asked at the end of the survey whether they felt anything was missing that they thought should have been included.

Many changes to the survey occurred concurrent with the interview process. As interviewees made comments about the survey, changes were discussed with members of the team and implemented so that updated questions could be validated during the interview phase. More details of the results of this process are discussed in Sec. 7.4.1.

Separately, we found that instructors might benefit from taking this survey, as well as from published works that will come from it. Some of the questions, especially those about branded approaches to instruction and the use of RBAIs, were very interesting to interviewees. These questions include links to outside resources about various instructional methods and assessments. One interviewee, while examining the question about branded approaches to instruction (such as ISLE and SCALE-UP), said:

It looks very interesting, actually... I'm going to open all of them... that looks nice. Here, we are very far away from that... I'm going to come back to this. Thank you so much for that
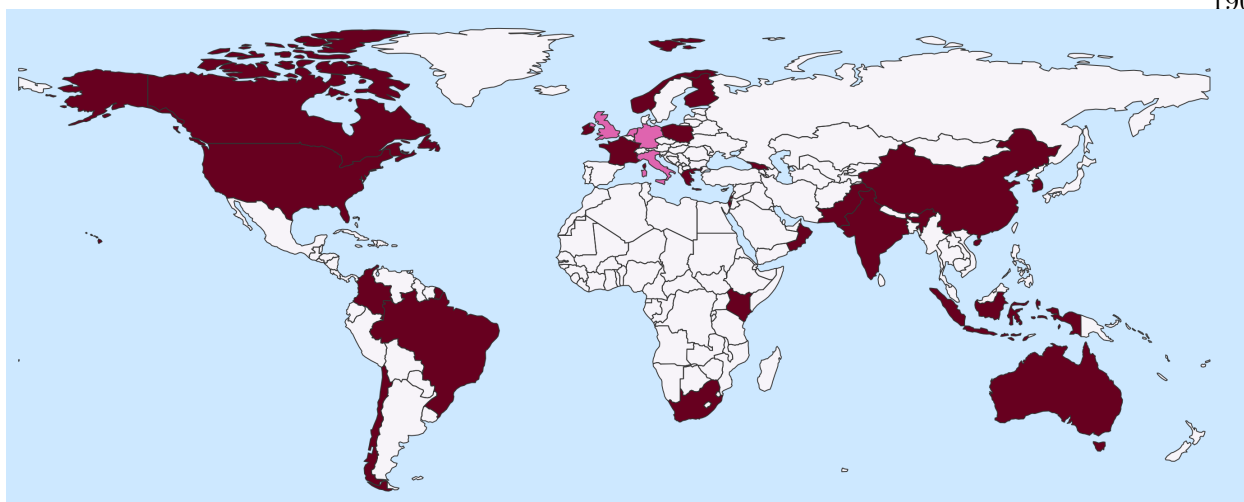
Figure 7.1: World map indicating the countries where we interviewed a lab instructor. Each maroon country had one instructor interviewed, aside from the United States, which had two. The pink countries represent the home countries of the collaborators of this chapter, aside from the United States, and therefore no interviews were solicited from instructors in these countries because the collaborators could provide the necessary feedback. We interviewed 23 instructors from 22 countries.

[sic] links.

This interviewee was excited to learn about methods they had not previously been familiar with. Questions such as these, and those about course goals and activities, can allow instructors to reflect on their courses and consider possible new teaching methods, assessments, and course goals. We hope this unintended impact of the survey can also be useful for improving laboratory instruction.

7.3.2   Survey Dissemination & Data Collection

We first disseminated the final version of the survey using all of the international contacts of the collaborators. We compiled a list of our contacts ($\sim$130), then emailed a solicitation to take the survey to people we knew through our professional networks. This solicitation invited people to take the survey, as well as to pass it along to others they know (whether in their own department or at other institutions), which is often referred to as snowball sampling [95].

We also posted the survey on a variety of listservs relevant to lab instructors. These included the ALPhA listserv [1], two American Physical Society (APS) discussion boards (Forum on Educa-

tion [2] and Topical Group on Physics Education Research [3]), a newsletter distributed to members of GIREP, and a JISCmail forum for physics education researchers and instructors in the United Kingdom [8]. Further, the survey was advertised during the ALPhA Beyond First Year (BFY) IV Conference, American Association of Physics Teachers Summer Conference, Physics Education Research Conference, and at the GIREP Conference (all in July 2023) during various collaborators' talks and poster presentations.

Next, we compiled a list of 171 countries worldwide and focused our efforts on those not represented in our sample thus far, especially those within world regions not well-represented. Each collaborator then searched for publicly available information about institutions with physics departments (e.g., from institution websites) within $10 - 15$ unique countries and contributed to a database with contact information for department heads and lab instructors. This approach has limitations, as many institutions did not have contact information easily accessible. However, this led to a large increase in responses: we received about 53 additional responses from 24 countries to add to our data set. The solicitation e-mail we sent to these institutions requested that they send to their contacts as well. In selecting countries, we chose places where we did not have a well-represented sample, such as much of Asia, Eastern Europe, South America, and Africa. This helped spread the survey to a more global audience than just where our contacts were located.

Data presented here were collected from 20 June 2023 until 10 January 2024, though the survey is still open and collecting responses [9]. We include data only from people who responded to a minimum number of questions (about 80% of the survey). This removes 121 responses with no questions answered and 18 responses that answered only some questions, leaving 217 out of the initial 356 responses. Because we do not force responses to most questions on the survey, aside from a couple that are necessary for future logic within the survey, each question has a varying number of responses. The number of respondents is reported for each part of the results section as appropriate. The median time to take the survey was 18 minutes.[1]

---

[1] Median is reported to exclude the outliers of those who leave the survey open for multiple days without actively filling it out before submitting it, thereby skewing the mean and making it an inappropriate statistic to report.

7.3.2.1   Limitations of our Data

First, aside from the United States, the countries where the collaborators are located (Germany, Italy, the United Kingdom, and the Netherlands) are oversampled based on the number of institutions present in each country, while most other countries are undersampled by this same metric. For example, we have only three responses from China and two responses from India, two largely populated countries with strong physics programs. We also have only four responses from Africa and nine responses from South America, therefore undersampling large areas of the world. We are also missing many countries entirely. The United States is underrepresented compared to responses received from other countries based on the number of higher education institutions — we have only 63 responses for the USA. Additionally, within the United States, there are a large variety of types of institutions, so getting a truly representative sample of institutions would be difficult and is something we do not currently have at this point in our data collection.

Second, some of our responses are clustered at specific institutions. One example of this is Canada — while we have six responses from Canada, five of them are from one university within Canada and all six responses are from a single province.

Third, since the survey was disseminated primarily using our contacts and listservs we are members of, those who responded are more likely to be interested or involved in physics education research to some extent. This also biased the countries from which we received responses, as those with relationships with the collaborators were more likely to fill out the survey and pass it on to their colleagues, hence why the country bias is skewed towards our home countries. There is also bias in who chose to fill out the survey: those with interest in improving their programs and are invested in the quality of lab teaching are more likely to fill out the survey.

Fourth, the survey is available only in English. While many of our colleagues do speak English, we miss many people who do not know English well enough to complete the survey. We chose not to translate the survey at this time due to constraints on resources and expertise.

Finally, because of these limitations, we do not present uncertainties in our tables in most

cases. This is because, due to the sampling bias in our data, our unknown systematic error is likely larger than the uncertainty determined by statistical means. It could be therefore misleading to present the statistical uncertainty.

7.4   Results and Discussion

7.4.1   Lab Taxonomy Survey

The final version of the survey itself is a major result of this work, as it was developed through a systematic process and extensive interviews, and will hopefully serve as a foundational tool for years to come. An adaptation of the Qualtrics version of the survey is presented in App. D. The survey is delivered online via Qualtrics and remains open for data collection [9]. We note that the survey does not collect information about the instructors (e.g., name, e-mail address, demographics), but rather focuses only on the course itself.

The survey is structured with eight separate sections to capture a variety of information including: overall course and institution characteristics, students, grouping of students, instructional staff, goals, activities, evaluation, and an optional open text box for any additional items, including a suggestion for lab activity titles. Some sections initially included in the survey were removed as we iterated through development. Sometimes, this occurred for practical considerations. For example, one section detailing equipment available in the lab course was removed due to being too laborious for both those taking the survey and the researchers analyzing the data, as well as likely not being helpful in classifying these courses. Other areas were removed because we are surveying instructors, who might not know the information. For example, instructors likely do not know how much time students are spending outside of the laboratory studying, preparing, or writing lab reports.

Interviews helped shape the final form of the survey. The first general category of survey edits were simple improvements, including the addition of a progress bar, bolding 'select all that apply' wherever it appeared in order to draw attention to it (based on several interviewees missing this text

in the questions), inclusion of a back button, and other general readability improvements. These were minimal edits that did not significantly change the contents of the survey, but rather improved the user experience. In addition to these changes, minor wording changes and clarifications were made throughout the survey to improve understandability to a wider audience.

In the following, we present each section of the survey together with the description of the changes we made during their development.

7.4.1.1 Overall Characteristics

This section of the survey asks for basic information both about the institution where the course is taught and about the course. Institution questions include the location and name of the institution and highest degree it grants. Course information collected in this section includes general information, such as the name of the course, the intended level of the course (introductory vs. beyond introductory), a checklist of physics topics covered by the course, number of students enrolled in a typical term, and the basic setup of the course — for example, whether students participate in project work, and the types of experiments students do (weekly, many experiments per week, or experiments that last longer than one week). Other questions include whether the lab course is integrated with a lecture course and whether the course includes lectures on statistics, data analysis, or experimental techniques. This section also includes questions about project work which are displayed if participants indicate that projects are part of the course.

Many edits were made to this section during the interview phase. Several items were added to the list of topics that might be covered in a physics laboratory course, including quantum information, geophysics, and modern physics; this change occurred due to interviewees suggesting topics that they felt were important, but were not included in the original list. A question was also added to this category to probe the level of the lab, introductory or beyond introductory, after it became clear that the year of the students taking the course is not sufficient to provide this information (e.g., a course for life science majors might be mostly second or third year students, but it might be the first physics laboratory course these students take and is therefore considered

to be introductory). Another question was added to probe whether the course meets weekly. If the respondents choose 'no,' we added an open text box for them to provide details of their course meeting schedule. We added this text box as several interviewees mentioned that their courses followed unique course meeting schedules. Because we can't account for every such case, we determined that an open text box was the best method to collect this information and may help refine future iterations of the survey.

7.4.1.2   Students

This section of the survey first asks about students' majors and the percentage of students in the course earning a degree in physics or astrophysics. Finally, this section asks participants to estimate how many years the students have been at the university.

The most significant change that occurred in this section due to interviews was to include a question asking about the percentage of students in the course that are physics majors. We decided that simply asking about the majors of students was not enough information to determine whether the course is intended for physics majors, non-majors, or both; the inclusion of this question helps make that more clear. In addition to this change, more majors were added to the list of potential degrees students might be earning as a result of interviewee requests.

7.4.1.3   Students' group work

This section of the survey inquires about how students work in the lab: alone or with others. If students work with others, participants are asked several followup questions, including the typical size of a group, how groups are chosen, and whether students stay in the same groups for the entire course.

No significant changes were made to this section section during the interview phase aside from slight wording changes to make things more easily understood.

7.4.1.4   Instructional Staff

This part of the survey asks about the types of instructional staff present in the lab with students. These might include faculty, lab technicians, graduate teaching assistants (TAs), and undergraduate learning assistants (LAs). In this section, participants are also asked about training provided to TAs and LAs — both the frequency of this training and the topics covered (e.g., familiarization with equipment, pedagogy instruction, and grading training).

In this section, we added lab technicians during the interview phase at the request of interviewees. Slight wording changes were also made to the questions inquiring about the types of training.

7.4.1.5   Goals

In this section, many potential goals for a physics laboratory course are listed, and participants rank these on a Likert scale consisting of Major Goal, Minor Goal, Not a Goal, and Future Goal (not currently a goal). The development of this unique Likert scale is discussed in more detail below.

The list of goals are as follows:

- Reinforcing physics concepts previously seen in lecture (confirming known results / seeing theory in an experiment)

- Learning/discovering physics concepts not previously seen in lecture

- Developing technical knowledge and skills (e.g., making measurements and hands-on manipulation of equipment)

- Designing experiments

- Developing mathematical model(s) of experimental results

- Learning how to analyze and interpret data (e.g., linear regressions, uncertainty)

- Learning how to visualize data (e.g., plotting)

- Developing lab notebook keeping skills

- Developing scientific writing skills (e.g., lab reports)

- Developing other communication skills (e.g., oral presentations, poster presentations)

- Making quick and simple approximations to predict experimental outcomes (e.g., back of the envelope calculations)

- Developing expert-like views about the nature of the process of doing experimental physics (e.g., experimentation is iterative, not linear)

- Developing collaboration and teamwork skills

- Reflecting on and evaluating one's own learning and knowledge (metacognition)

- Enjoying experimental physics and/or the course

The goals section went through major revisions during the interview process. First, the Likert scale was changed; initially, it included only Major Goal, Minor Goal, and Not a Goal. However, we observed that many interviewees would say that one of the goals is not a goal of their course, but they would be interested in implementing it. They would then often select 'Minor Goal', despite stating it is not a goal of their course. In order to address this issue, we introduced a fourth Likert option: Future Goal (not currently a goal). While analyzing the survey data, we currently collapse this category with Not a Goal, but it helps to provide more accurate results in our data collection.

Additionally, the list of goals presented underwent revisions. Some goals, such as developing communication skills, were split into several goals [in this case, the split was into three goals: developing lab notebook keeping skills, developing scientific writing skills (e.g., lab reports), and developing other communication skills (e.g., oral presentations, poster presentations)]. This was due to interviewee input about the concepts they thought were covered by the goal. In this particular

example, when asked to define "communication skills," interviewees had many different ideas about what this might include. Therefore, we split this into three distinct goals in order to collect the most accurate data. Further, we added goals at the request of interviewees, such as 'enjoying experimental physics and/or the course' and 'making quick and simple approximations to predict experimental outcomes (e.g., back of the envelope calculations)'.

Other changes to the goals section as a result of interviews included wording changes to clarify meaning, as well as adding examples to the goals to make them more easily understood.

7.4.1.6 Activities

The first part of this section of the survey asks participants whether they use any officially branded approaches to lab instruction in their course [e.g., Investigative Science Learning Environment (ISLE) Physics [76], Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) [30], and Modeling Instruction [36]], as well as whether they use any RBAIs to evaluate the course (e.g., Survey for Physics Reasoning on Uncertainty Concepts in Experiments, or SPRUCE [184,235]; Modeling Assessment for Physics Laboratory Experiments, or MAPLE [84]; and E-CLASS [254]).

The next part of this section lists several activities students might participate in during an undergraduate physics laboratory course and asks how often students engage with them along the following Likert scale: Very frequently, Somewhat frequently, 1-2 times per semester/term, Would like to use in the future, and Never. Details of the development of this Likert scale are given below. The activities probed are divided into the following categories:

- Data Analysis and Visualization

- Communication

- Student Decision-Making

- Materials and Resources

- Modeling and other activities

Within the above categories, examples of activities include: quantify uncertainty in a measurement, write lab reports, develop their own research questions, and calibrate measurement tools; the full list of activities can be found in App. D.

The activities section also underwent significant changes during the interview process. A question probing whether research-based assessments are used to evaluate the course was added as a result of interviewees mentioning during interviews that they use some of these assessments.

The Likert scale used in the list of activities was changed to make things clearer. Originally, the scale was: Always, Often, Sometimes, Rarely, Never. However, this scale was not appropriate for some activities. For example, students might design and present a poster once during a course, and it is unclear which of the scale points this should fall into, because it is not typical for students to make posters "frequently" when compared with other activities, such as quantifying uncertainty in a measurement, which might occur with every lab experiment. Similarly, it is unclear what it means to 'always' complete a safety training — some courses might require a single training, whereas others might have a few that students have to complete.

The new Likert scale we implemented is: Very frequently, Somewhat frequently, 1-2 times per semester/term, Would like to use in the future, and Never. This scale has several advantages over the previous one. First, it includes an aspirational scale point (i.e., Would like to use in the future), which can be collapsed with Never when analyzing the survey data, but again helps discourage those who select a different option despite not using the activity (similar to the aspirational Likert scale point in the Goals section). Second, the new scale helps clarify events that might happen only one or two times in a semester, such as a poster presentation or a safety training. Finally, for activities that might happen more commonly in courses — such as keeping a lab notebook or writing their own code — it provides several scale points that are more easily understood. Because 1-2 times per semester/term is an option, it is clear that 'somewhat frequently' means that students participate in the activity more than this, while 'very frequently' indicates a higher degree. Therefore, based

on interviewee responses to this scale, we feel that we have appropriate knowledge of what it means each time someone selects a particular scale point.

In addition to a new Likert scale, many of the activities were also changed. In some cases, wording was altered or examples were added to clarify meaning. Some activities were combined into one option after it was determined that interviewees could not always tell the difference between them. One example of this is combining 'refine experimental apparatus or procedure to reduce random uncertainty' and 'refine system to reduce systematic uncertainty' into 'refine experimental apparatus or procedure to reduce uncertainty (statistical and/or systematic)'. This was due to many interviewees not being able to give appropriate examples differentiating random and systematic uncertainty, and therefore the information obtained from probing these separately was inaccurate.

Finally, some activities were added to this section, such as 'engage with PhET simulations' and 'complete safety training' due to interviewee requests.

7.4.1.7   Evaluation of students' work

This part of the survey probes how students are graded (evaluated) in the course. The first question probes whether students are assigned individual or group grades, while the second lists potential parts of the course that might factor into student grades and asks participants to select all that are used for their course. Examples include taking a quiz at home before the lab, lab notebooks, lab reports, written exams, poster presentations, practical exams (i.e., hands-on exams), and peer feedback on other students' work. Finally, participants are asked whether they use a rubric (a set of guidelines about how something is graded) to grade student work.

The evaluation section of the survey also underwent significant changes due to interviews. Initially, participants were provided with a list of items that might potentially be included in student grades and were asked to rate them on a Likert scale: Not used, Marked/graded for inclusion in final course grade, Marked/graded but not used to determine final course grade. This led to confusion, especially about certain items that are not directly used in the final grade, but might be used in some indirect way. For example, attendance at each individual course meeting might make up an

overall attendance grade that is then used to determine the final course grade. Interviewees were then uncertain about where on the Likert scale to include attendance. We changed this question to be multiple response, and ask participants to select all of the items on the list that are used in grading the course. Because it is a binary option, interviewees understood how to handle indirect affects on grades (they did select these items). Removing the Likert scale helped make the survey more clear.

Additionally, some items were removed (such as reflection questions) due to interviewee lack of understanding around these points, added (such as worksheets) due to interviewee requests, and reworded to help with clarity and understanding.

7.4.1.8   Optional Long Entry

In this section, two long-form text boxes are provided for participants. Both are listed explicitly as optional; while most of the rest of the survey's questions are optional, these are the only two questions which state this. The first asks participants to enter the titles of lab experiments in any language they wish. This is useful in looking at trends of common laboratory themes that might be present, especially in introductory labs (e.g., during the interviews, many instructors discussed using a pendulum activity in introductory mechanics). While this qualitative data might not provide the most accurate evaluation of themes — for example, some might choose not to include this information, and others might have laboratory titles that don't fully reflect the activities — this will still provide a wealth of information. We encourage any language entry to allow participants to simply copy and paste lab manual titles in order to make this step easier. Online tools such as Google Translate provide an accurate enough translation to qualitatively code themes in later steps of the analysis.

The second text box asks for any additional comments that participants might have. This is useful in cases where the survey may not have fully captured the experience of the participant in teaching their course or for any clarifying comments they would like to make about their prior responses.

The optional long entry section was not significantly altered during the interview process.

7.4.2   Survey Results

We present here an overview of physics laboratory courses around the world based on the data we have collected thus far. We received responses from 217 unique courses in 41 countries. A figure showing a map of the countries where instructors responded to the survey is shown in Fig. 7.2, with a full list of the number of respondents per country located in Tab.7.1.
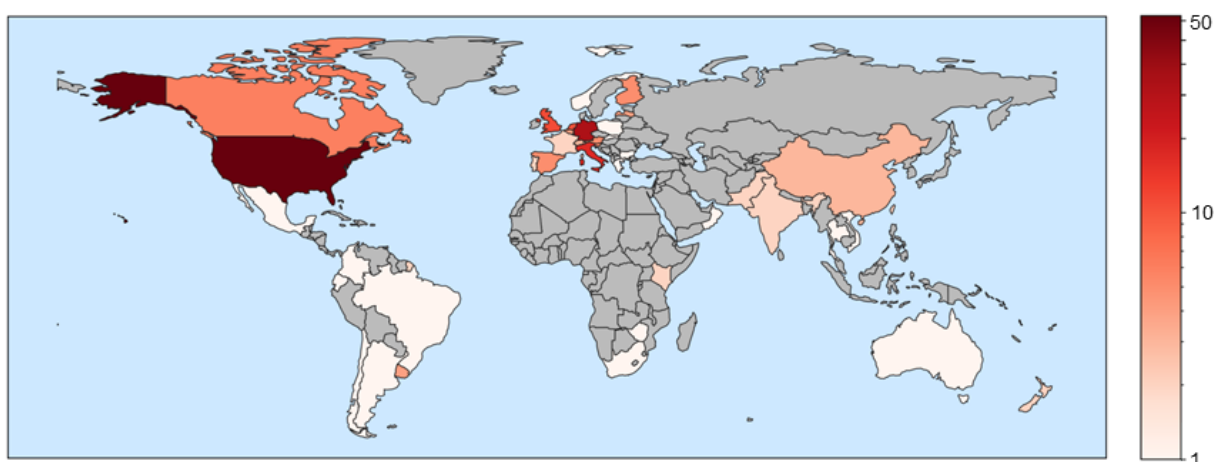

Figure 7.2: World Map of Survey Responses. Shown on a log scale, each colored country has at least one response; countries in gray have no responses. We received the most responses from the United States (63 courses).

As discussed previously, our data is limited — it is skewed towards the countries the collaborators of this chapter are from and is lacking representation in many areas. In future iterations of this work, we hope to present a more representative sample.

7.4.2.1   Overall Characteristics

Of our respondents, most courses (166/217) were offered at PhD-granting institutions, with fewer being offered at Master's-granting institutions (20/217), Bachelor's-granting institutions (25/217) and Associate's-granting institutions (6/217). Most of this variation comes from within

Table 7.1: Respondents by Country. The number of unique courses per country that are included in the final data set (N = 217), as well as the percent of responses from each country are listed. The countries where the authors are from have a higher than average percentage of responses.

Country	Num. Responses	% Responses
United States	63	29
Germany	33	15
Italy	17	7.8
United Kingdom	11	5.1
Netherlands	10	4.6
Canada	6	2.8
Hong Kong	6	2.8
Austria	5	2.3
Finland	5	2.3
Slovenia	5	2.3
Spain	5	2.3
Belgium	4	1.8
Latvia	4	1.8
Uruguay	4	1.8
China	3	1.4
Switzerland	3	1.4
Czech Republic	2	0.92
France	2	0.92
Kenya	2	0.92
India	2	0.92
New Zealand	2	0.92
Pakistan	2	0.92
Portugal	2	0.92
Taiwan	2	0.92
Argentina	1	0.46
Australia	1	0.46
Brazil	1	0.46
Bulgaria	1	0.46
Chile	1	0.46
Colombia	1	0.46
Ecuador	1	0.46
Greece	1	0.46
Mexico	1	0.46
Norway	1	0.46
Oman	1	0.46
Poland	1	0.46
Slovakia	1	0.46
South Africa	1	0.46
Thailand	1	0.46
Vietnam	1	0.46
Zimbabwe	1	0.46

the United States - 33 of the non-PhD granting institutions are in the USA (63 total responses) whereas only 18 are outside of it (217 total responses).

Of the courses surveyed, 137 are introductory, 79 are beyond introductory, and we have no data about one course. We discuss a split of the data by introductory and beyond introductory where appropriate in our analysis.

Next, we examine the number of students per course and the number of students per section in the course (i.e., the number of students present in the laboratory room at one time). Distributions for both of these are shown in Fig. 7.3. The median number of students per course is 50. Overall, there are only a few very large courses with more than 500 students. Most courses (188/216) have fewer than 200 students per course. Not surprisingly, only three beyond introductory courses have more than 200 students. The median number of students per section is 18. Most courses (152/217) have between 10 and 40 students per section. Very few courses (14/217) have more than 75 students present in the lab at any given time; of these courses, most (12/14) are introductory, with only one beyond introductory.

We probe the topics covered in the course by providing a multiple-response list of possible topics and the option to type in additional topics not listed (No clear patterns emerge from the not listed responses). We present a view of the topics covered in the courses in Fig. 7.4, with a split shown between introductory and beyond introductory courses. Classical mechanics is the topic selected most often for introductory courses, while optics and laser physics was the most common topic for beyond introductory courses. More specialized areas of physics, such as plasma physics and geophysics, are rarely covered in any lab courses.

In addition to the course topics, we also asked instructors to provide the titles of their lab experiments in a long-form text box. Of the 217 respondents to the survey, 111 chose to do so, and we received 1,078 lab titles from these courses. After translating all of the titles to English, we qualitatively coded them to determine the most common types of experiments occurring in undergraduate physics lab courses. These experiments are presented in Tab. 7.2. This includes codes with at least 15 courses using them. Of the 1,078 lab titles received, 96 (8.9%) were uncoded
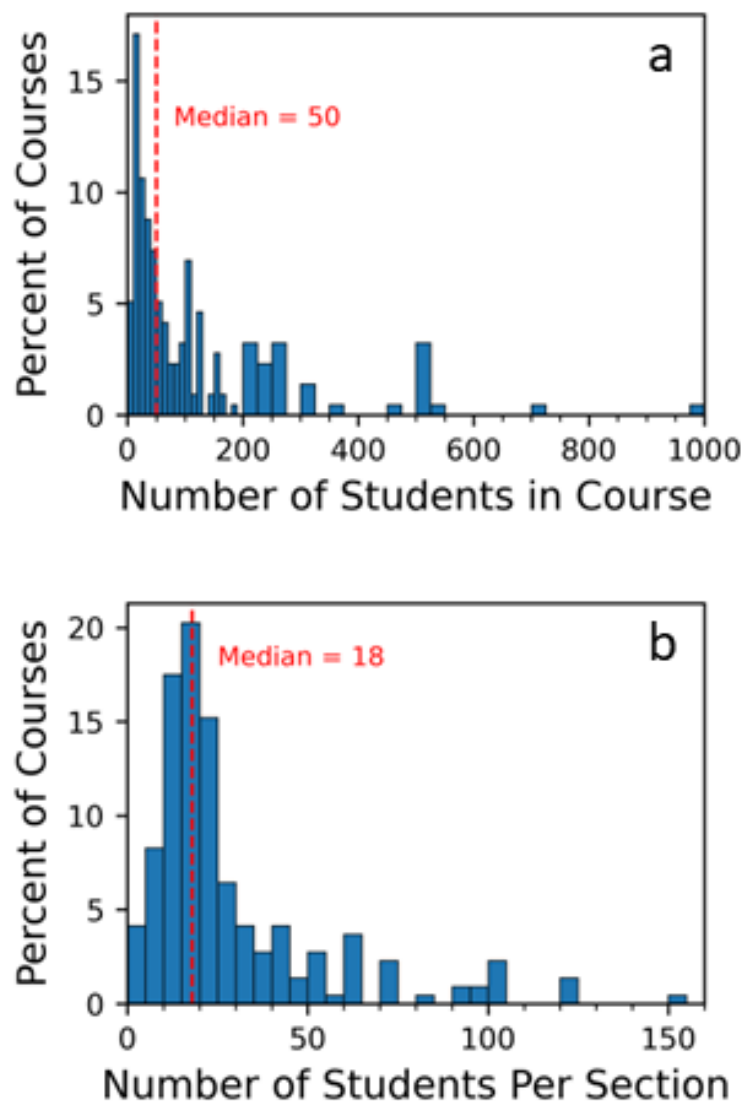
Figure 7.3: Number of students per course and per section. Upper histogram (a) shows the distribution of the total number of students in the course, with a median shown as a red dashed line (50) [N = 216]. Lower histogram (b) shows the distribution of the number of students per section of the course (i.e., number of students in the lab room at any time), with a median shown as a red dashed line (18) [N = 217].

due to being too vague (e.g., classic mechanics), activities beyond a lab experiment (e.g., poster preparation), or too specific (e.g., interaction and collaboration of kirigami). The other 982 lab titles were categorized with one to three codes. The code definitions for all codes and more information about the process of coding the lab titles are located in App. C. Additionally, we created a word
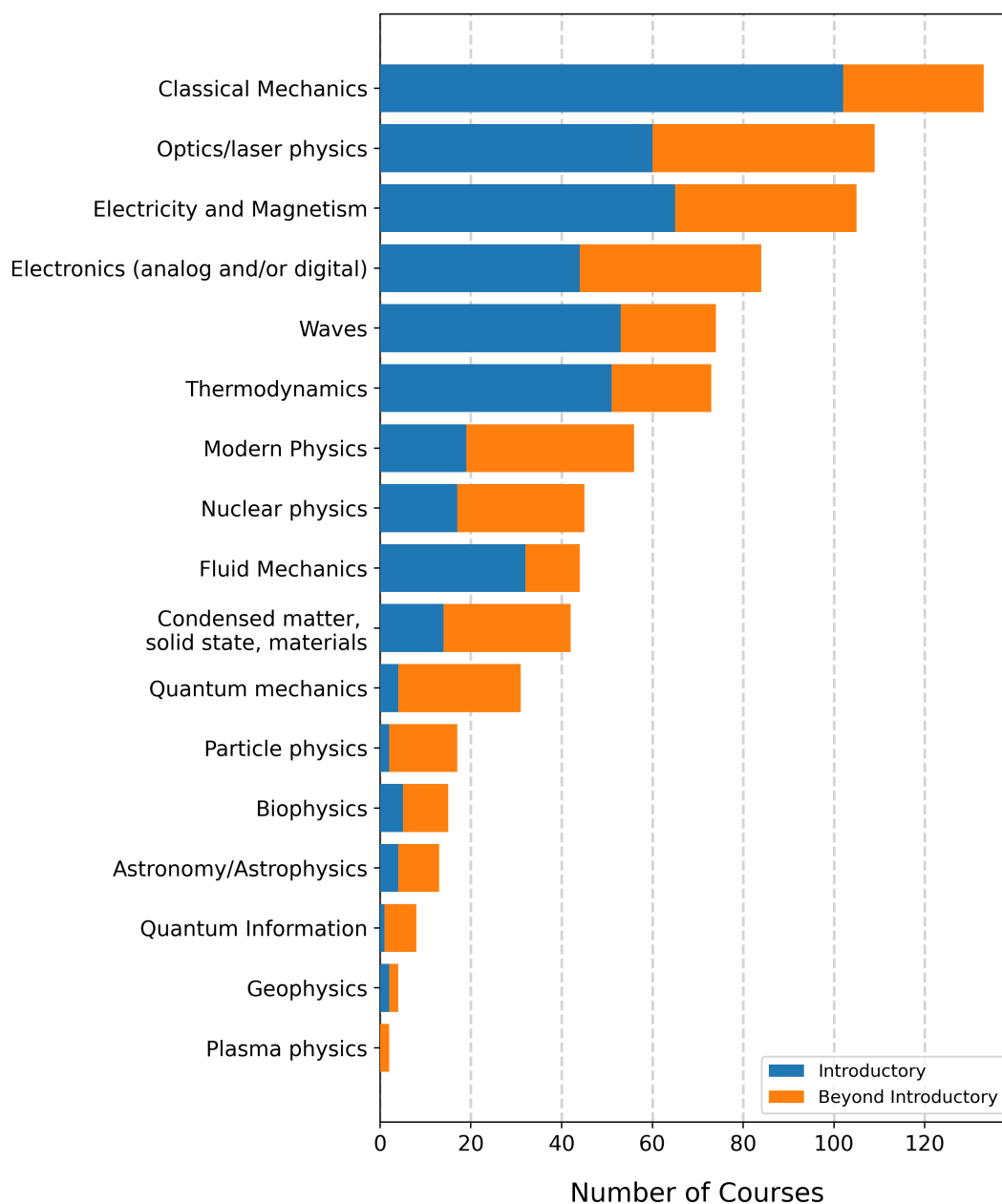
Figure 7.4: Topics covered in the course, split by introductory (blue, left) and beyond introductory (orange, right) courses [N = 216]. The bottom axis shows the absolute number of courses that included the topic. The most common topic for introductory courses is classical mechanics, and the most common topic for beyond introductory courses is optics and laser physics.

cloud of these lab titles after removing stop words [120] to give a visual representation of the data (See Fig. 7.5). We hope that, as we collect more data, qualitative coding of lab titles will help in creating a taxonomy: we can work on grouping courses that complete similar types of experiments.

Table 7.2: Most common experiments as given by titles of lab experiments. These are the codes for all categories with at least 15 courses reporting at least one lab in this category. A total of 111 courses providing 1,078 lab titles were qualitatively coded to determine the most common experiment types. The definitions of these codes and others not shown are provided in App. C

Topic	Num. Courses	Num. Lab Titles
Optics (intermediate)	68	108
Kinematics	36	55
Dynamics (mechanics)	33	56
Electronics (intermediate)	29	69
Electronics (simple)	29	56
Spectroscopy	28	37
Test and measurement equipment	27	30
Thermodynamics	26	56
Introduction to measurement and uncertainty	25	36
Pendulum	23	33
Optics (simple)	23	39
Particle physics	18	44
Optics (advanced)	17	49
Advanced materials and solid state	17	30
Waves	17	22
Electric fields and electrostatics	16	24
Fluids	15	21
Magnetic fields	15	18

We also asked whether the laboratory course is integrated with lecture (i.e., if both are one course combined) or if the laboratory course is a separate course. There are 129 courses that are separate, while 88 are combined with a theory course. Additionally, 136 courses include lectures about statistics and/or data analysis, whereas 81 do not.

Next, respondents reported the number of weeks the course runs for, as well as the number of hours per week students are scheduled to be in the lab. The distributions for these are shown in Fig. 7.6. The median number of weeks is 12, and the median number of scheduled hours per week is three. Further, data about the number of hours per week beyond the scheduled time that students spend in the lab is presented in Tab. 7.3; in most cases, students do not spend any time beyond what is scheduled in the lab.

We further investigate the number of experiments per lab meeting students complete and if students have choice over which experiments they complete. This question is multiple response,
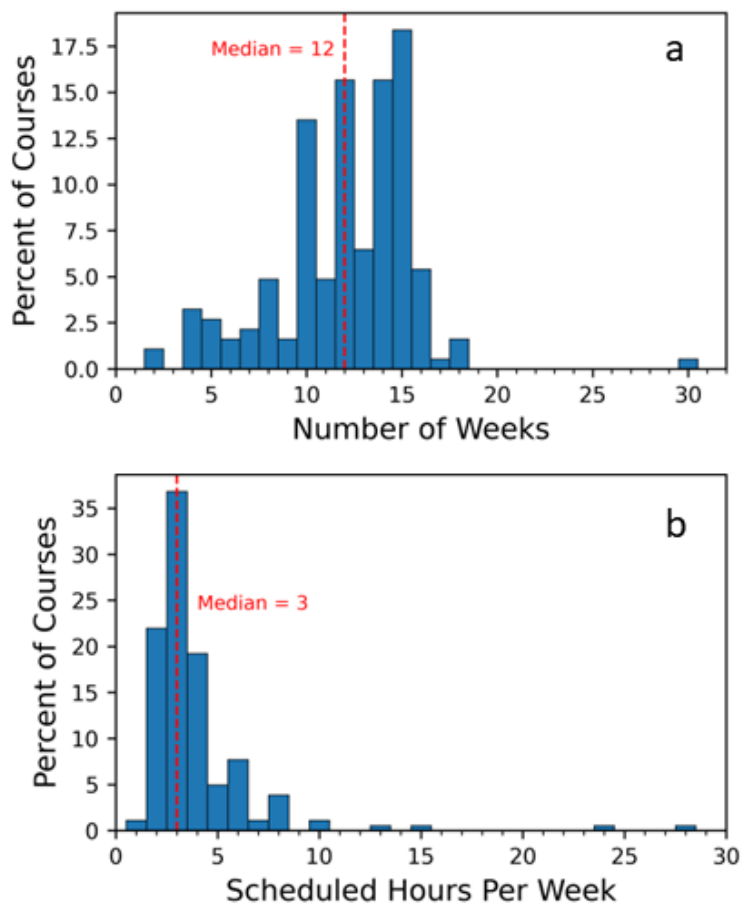
Figure 7.5: A word cloud showing the 200 most common words after removing stop words from the lab titles and using basic lemmatization [N = 111 courses with 1,078 lab titles]. This helps us form a visual representation of the types of experiments happening in undergraduate physics lab courses around the world. Electronics, mechanics, and optics experiments dominate the word cloud.

Table 7.3: Number of hours per week beyond the scheduled time students spend in the lab [N = 185]. In most courses, students do not spend any time in the lab other than what is scheduled.

	Num. Courses	% Responses
0 hours	103	56
1 - 3 hours	56	30
4 - 6 hours	7	3.8
More than 6 hours	6	3.2
Unknown	4	2.2
Not allowed	9	4.9

and respondents can choose whether students complete multiple experiments per meeting, one experiment per meeting, one experiment per multiple meetings, or a multi-session open-ended project. The distribution of responses, split by introductory and beyond introductory courses, is presented in Fig. 7.7. In most courses, students spend time doing one experiment per meeting of the course. Students are often not given a choice of which experiments they do, with 125 courses not allowing students any choice in which experiments they complete, 59 allowing students to choose their experiments for some portion of the course, and 33 allowing students to choose experiments

Figure 7.6: Number of weeks the course runs for (a) [N = 185] and number of hours per week students are scheduled to be in the lab (b) [N = 182]. Red dashed lines shown the median. The median number of weeks the course runs is 12, and the median number of hours per week is 3.

all of the time.

For those courses where there is a project component, we asked four additional questions about the project. There are 48 courses that contain some project component. Of these courses, students spend a median of 4.5 weeks engaged in project work. In 26 courses, students choose their project topic "all of the time", compared with 16 courses that allow students to choose "some of the time", and six courses that do not allow students to choose their own project topic. Thirty-two courses allow students to design their own project "all of the time", whereas 13 courses allow students to do this only "some of the time", and only three courses do not allow students to design
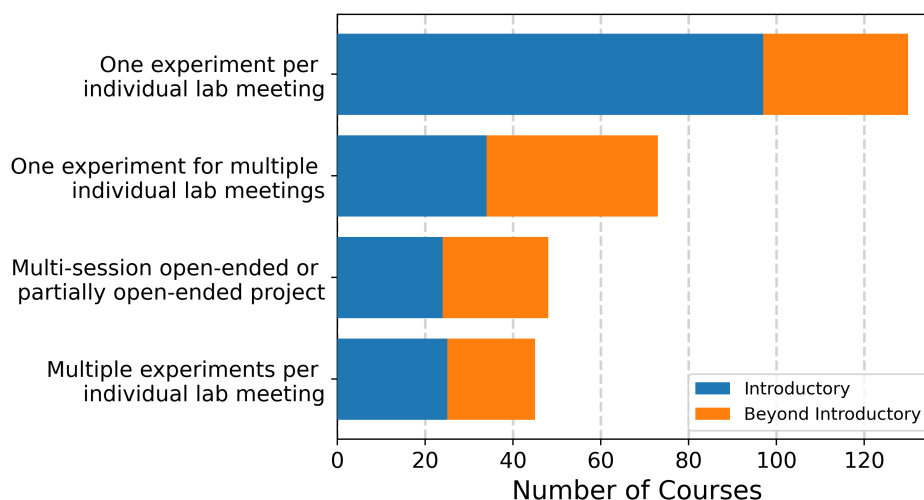
Figure 7.7: Types of experiments students complete in the course, split by introductory (blue, left) and beyond introductory (orange, right) [N = 216]. This question was multiple response, so respondents could select as many of these options as apply to their course. Most courses involve some component in which students complete one experiment per meeting of the course.

their own project. Finally, in 26 courses, students always build their own experimental apparatus for the project, while in 19 courses they do this sometimes and in three courses they never do this.

7.4.2.2   Students

Respondents reported which major(s) their students have in the course, as well as the fraction of physics and astrophysics majors in the course. For the first of these questions, we presented a list of possible majors as a multiple response question along with the option to write in majors not contained on the list. The second of these questions is multiple choice. We present the results of these items along with a split by introductory and beyond introductory in Tab. 7.4. Because the question about the major(s) of the students is multiple response, we received many different potential groupings. The table, therefore, represents only the most common groupings (i.e., a minimum of four courses chose this grouping). Nearly a third of all courses include only physics majors, with 27% of introductory courses and 41% of beyond introductory courses having only physics majors. The second-most common combination overall is physics and astrophysics/astronomy majors, which accounts for another 9% of responses. Overall, 166 out of the 217 courses surveyed (76%)

included physics majors, and therefore, 51 courses (24%) do not include any physics majors.

Another point of discussion about majors is that the United States typically treats physics courses differently than courses outside of the United States. Within the USA, it is very common to combine many different majors into one course at the introductory level, whereas outside of the USA, this is less common. Outside of the USA, it is instead more common to have an introductory physics course only for physics majors, one for those training to teach high-school physics, a separate course only for engineering majors, etc. When examining only introductory courses, 81 % (29 out of 36) introductory courses in the USA contain 0-25% physics majors while outside of the USA, this number drops to 39% (39 out of 101 courses).

Table 7.4: Most common grouping of student majors in a class and percentage of the course that is physics majors. Only the most common groupings are shown (i.e., a minimum of four courses in that grouping).

	% Responses (N = 217)	% Responses, Intro (N = 137)	% Responses, Beyond Intro (N = 79)
Physics	32.2	27.0	41.8
Physics, Astrophysics/Astronomy	9.2	5.1	15.2
Another science	5.5	8.8	0.0
Physics, Physics/Astronomy teaching	5.5	2.2	11.4
Engineering	4.6	5.8	2.5
Physics, Engineering	2.8	14.6	5.1
Chemistry, Another science	2.3	3.6	0.0
Physics/Astronomy teaching	1.8	2.2	1.3
Other	35.9	43.8	22.3
	% Responses (N = 216)	% Responses, Intro (N = 136)	% Responses, Beyond Intro (N = 79)
0-25% physics majors	34.3	50.0	7.6
25-50% physics majors	7.9	6.6	10.1
50-75% physics majors	2.8	8.1	6.3
75-100% physics majors	50.5	35.3	75.9

Next, we provide information about how many years the students have been at the University when they take the course. Again, this question is multiple response, so instructors can choose as many options as apply to their course. These data are presented in Tab. 7.5. The courses in our data set lean heavily towards first-year and second-year students.

Table 7.5: Year of students in the course [N = 216]. This question is multiple response, so instructors can select all options that apply to their course.

	Num. Responses	% Responses
1st year	106	49.1
2nd year	83	38.4
3rd year	66	30.6
4th year	39	18.1
5th year or higher	9	4.2

7.4.2.3   Students working in groups

We next inquire about the ways in which students work together in the course. Our survey results showed that 204 courses indicated students work with at least one partner, and 13 courses indicated that students work alone. Data about these 204 courses is shown in Tab. 7.6, including the number of lab partners, whether students stay in the same group for the entire course, and whether students choose their own groups. In most cases, students are working in pairs of their own choice and stay with this lab partner for the entire term.

Table 7.6: Grouping of students [N = 204]. Most students work with one lab partner of their choice for the entire term.

	Num. Responses	Percent Responses
Groups of 2	118	57.8
Groups of 3	63	30.9
Groups of 4	19	9.3
Groups of 5+	9	4.4
Stay in same groups	165	80.9
Switch groups	39	19.1
Choose their groups	139	68.1
Are assigned groups	31	15.2
Both options	34	16.7

7.4.2.4   Instructional Staff

This section of the survey asks about the number of different types of instructional staff present in the lab room with the students. Because we also know the number of students present

in the lab at one time, we can determine the average number of students per staff. The means of this are shown in Tab. 7.7, including information about faculty members, lab technician, graduate and postdoctoral TAs, and undergraduate LAs. The distribution of the number of students per staff member is shown in Fig. 7.8. The mean number of students per staff member (after summing all possible types of staff members) is 9.9. Few courses utilize LAs, whereas many courses have faculty and TAs present with students.

Table 7.7: Mean number of students per staff member in the lab. We pair the question probing number of staff in the room with the question about the number of students in each section to determine these averages. On average, there are a total of 9.9 students per staff member. Means take into account only courses with at least one of that type of instructional staff (e.g., courses with no undergraduate LAs are not counted in the mean number of students per LA).

	Mean	Num. Courses
Students Per Faculty	25	186
Students Per Lab Technician	33	95
Students Per Graduate TA	20	116
Students Per Undergraduate LA	21	51
Students Per Staff (total)	9.9	215

Further, respondents provided information about the frequency and type of training for both graduate TAs and undergraduate LAs. This question is multiple choice in which respondents can indicate whether training happens once per term, once per academic year, or weekly. The types of training is a multiple response question, which allows respondents to select pedagogy, grading, and familiarization with lab equipment in any combination that applies to their course. Both of these questions also have 'not listed' options with the opportunity to write in a response; no patterns emerged from an analysis of these responses. These data are shown in Tab. 7.8. Nearly all courses that train TAs and/or LAs provide instruction to familiarize them with the lab equipment, while more than half also offer pedagogy or grading training. There is no standardized frequency of this training, with about one-third providing training once per term and one-quarter providing training once per academic year or weekly.
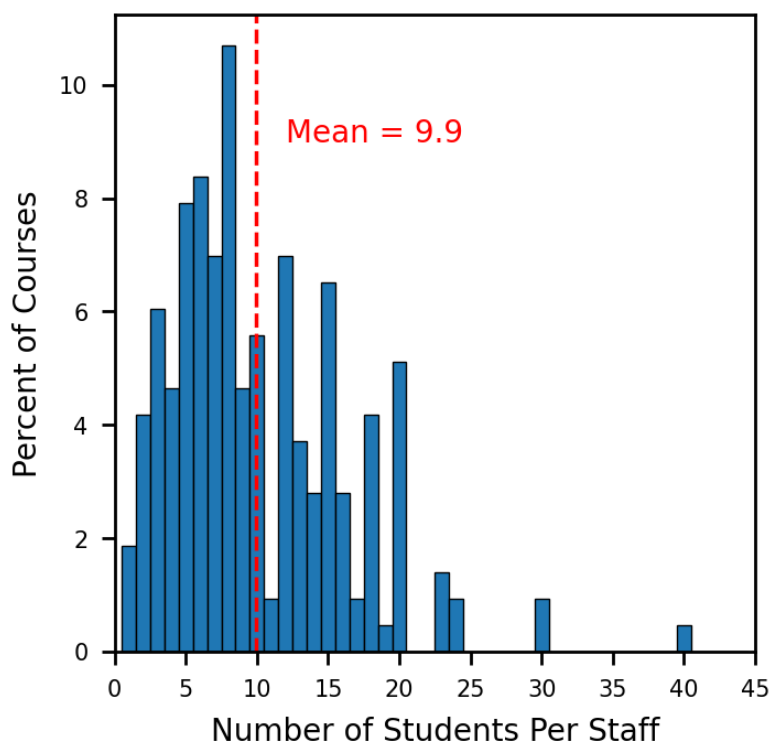
Figure 7.8: Distribution of the number of students per staff member in the lab room at any given time, with mean = 9.9 shown as a red dashed line.

Table 7.8: TA and LA training frequency [N = 137] and type [N = 136]. For the frequency question, respondents can type in an answer if none of the provided options capture their training schedule. Type of training is a multiple response question and also includes the ability to type in a response if a type of training is missing from the provided options.

	Num. Responses	Percent Responses
Once per term/semester	45	32.8
Once per academic year	35	25.5
Weekly	33	24.1
Other	22	16.1
Familiarization with equipment	130	95.6
Pedagogy	80	58.8
Grading	80	58.8
Other	7	5.1

7.4.2.5   Goals, Activities, and Evaluation

Potential course goals or learning objectives were presented as a list with options to select 'Major Goal', 'Minor Goal', 'Not a Goal', and 'Future Goal (not currently a goal)'. As previously discussed, the latter two of these categories are collapsed for all analysis. Each of the 15 goals presented had between 215 and 217 responses. A plot of the answers to this question is shown in Fig. 7.9. Other course goals are possible, but there is no "not listed" option available for this question. The current list of goals was refined through interviews, including the addition of extra goals as requested by interviewees.

We also examine the total number of goals (major plus minor) selected for each course. This distribution is shown in Fig. 7.10 and the mean is 11.8 goals out of the possible 15. On average, courses have 6.9 major goals and 4.9 minor goals. There are no significant differences in number of course goals (either total, major, or minor) for introductory and beyond introductory courses.

The first question in the activities section asks about whether a specific branded instruction technique is used (such as modeling instruction, SCALE-UP, or ISLE). Most courses (135/212) do not use any type of branded instruction method. Similarly, a question about RBAIs reveals that most courses (148/212) do not use any of these.

Next, we present respondents with a list of 41 possible activities broken up into five categories — data analysis, communication, student decision-making, materials, and modeling/other activities. Plots of the responses to the Likert-style questions for each of these categories are shown in Figs. 7.11, 7.12, 7.13, 7.14, and 7.15. These figures are broken down by Likert response (very frequently, somewhat frequently, 1-2 times per semester/term, and never, where we have again collapsed an aspirational scale point with never). We find that courses engage in a wide variety of activities. In some cases, such as in the collection of activities relating to both data analysis and student decision-making, at least half of the courses selected that they participated in all activities to some extent. The split between the frequencies students engage in activities also generally occurs as expected. For example, students typically write lab reports very frequently, but design
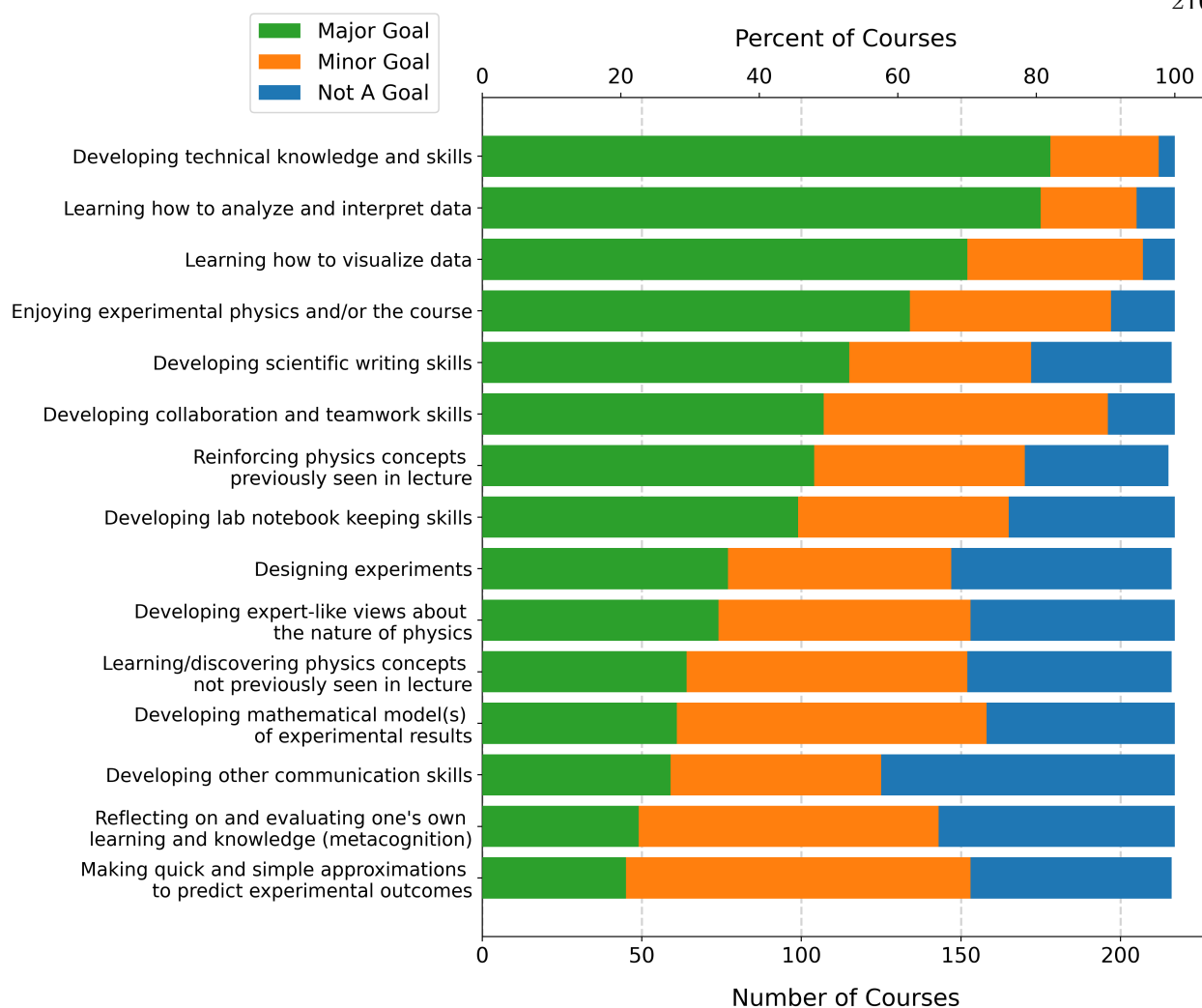
Figure 7.9: Course goals, split by major goal (green, left), minor goal (orange, middle), and not a goal (blue, right) for the course. Between 215 and 217 courses provided data for each goal, and the percentages are calculated using the full 217 courses for display purposes. The most commonly selected goal is developing technical knowledge and skills.

and present a poster 1-2 times per term. The student decision-making category, in particular, has a large number of activities with responses of 1-2 times per term.

A list of potential items that might be graded for inclusion in a student's final course grade is presented to respondents, and they are able to select all of the ones they use in their own course to assign student grades (multiple response). This list of 18 potential things might not fully span the space of items included in a grade, and so "not listed" with an option to write in other items
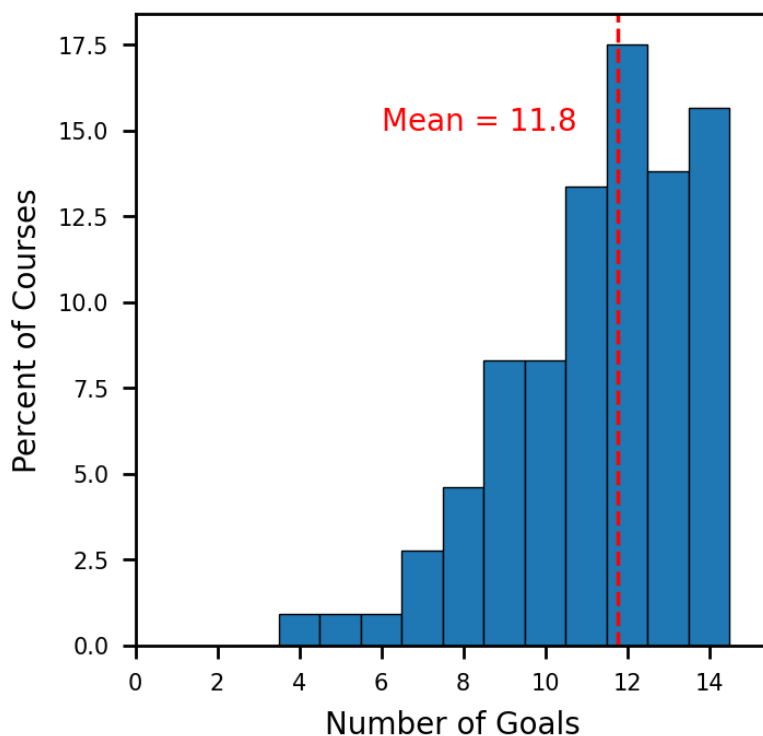
Figure 7.10: Distribution of the total number of goals (major plus minor) for each course. The maximum possible number of goals is 15 (all of the provided course goals). The mean number of goals per course (red dashed line) is 11.8.

is included in the survey. Tab. 7.9 shows the number of courses that include each option. The analysis of "not listed" write-in responses did not reveal any patterns. Most courses (about 75%) use lab reports to assign grades to students, with attendance and participation being the second most common item with more than half of the courses using this.

Finally, we can combine goals, activities, and items graded to determine how instructors are attempting to meet their course goals. I, along with H.J. Lewandowski and an outside PER postdoctoral researcher worked together to match the course goals with activities, as well as the course goals with items graded. We can then determine whether instructors are connecting their course goals with the activities students participate in and in the ways they are evaluated in the course. As a simple example, the course goal "Developing lab notebook keeping skills" can
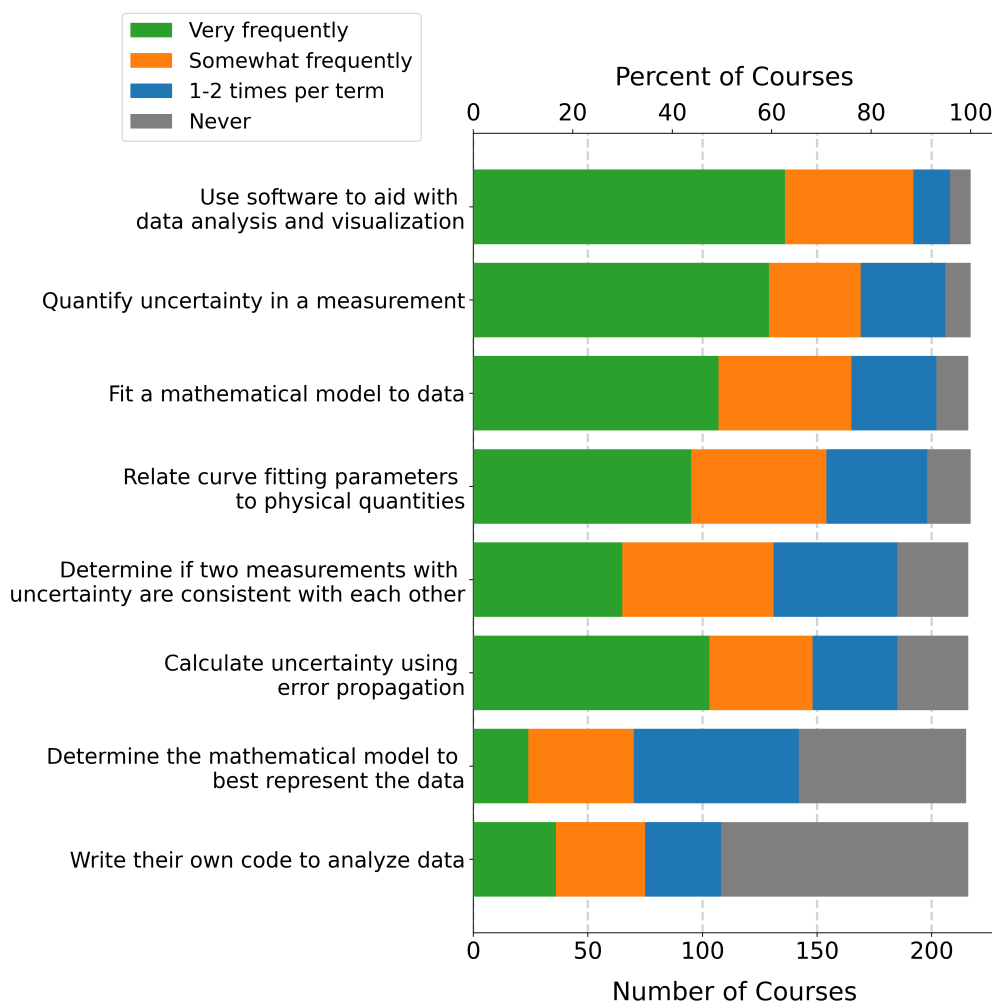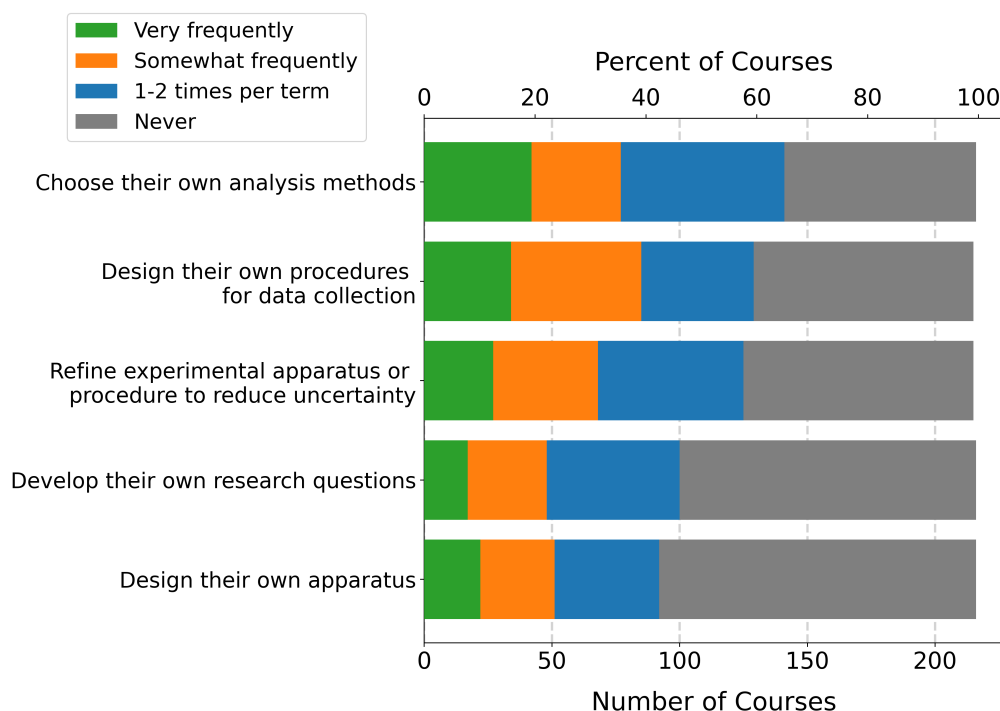
Figure 7.11: Courses with data analysis activities [N = 215 - 217]. Bars represent number of courses (bottom axis) and percent of courses (top axis) that include various activities related to analyzing data, such as error propagation and curve fitting. Bars are split based on frequency of the activity - very frequently (left, green), somewhat frequently (second from left, orange), 1-2 times per term (second from right, blue), and never (right, gray). In general, students participate in each of the data analysis activities to some extent in at least half of all courses. The least popular activity was for students to write their own code to analyze data.

be matched with activities "maintain an individual lab notebook" and "maintain a group lab notebook". This goal can be matched with "Lab notebooks" under items graded. After we came to agreement about matching, an analysis was done to determine how many activities instructors chose that matched the course goals, as well as how many items graded. We present the results of this analysis in Tab. 7.10. In this analysis, we collapse the categories of major goal and minor goal
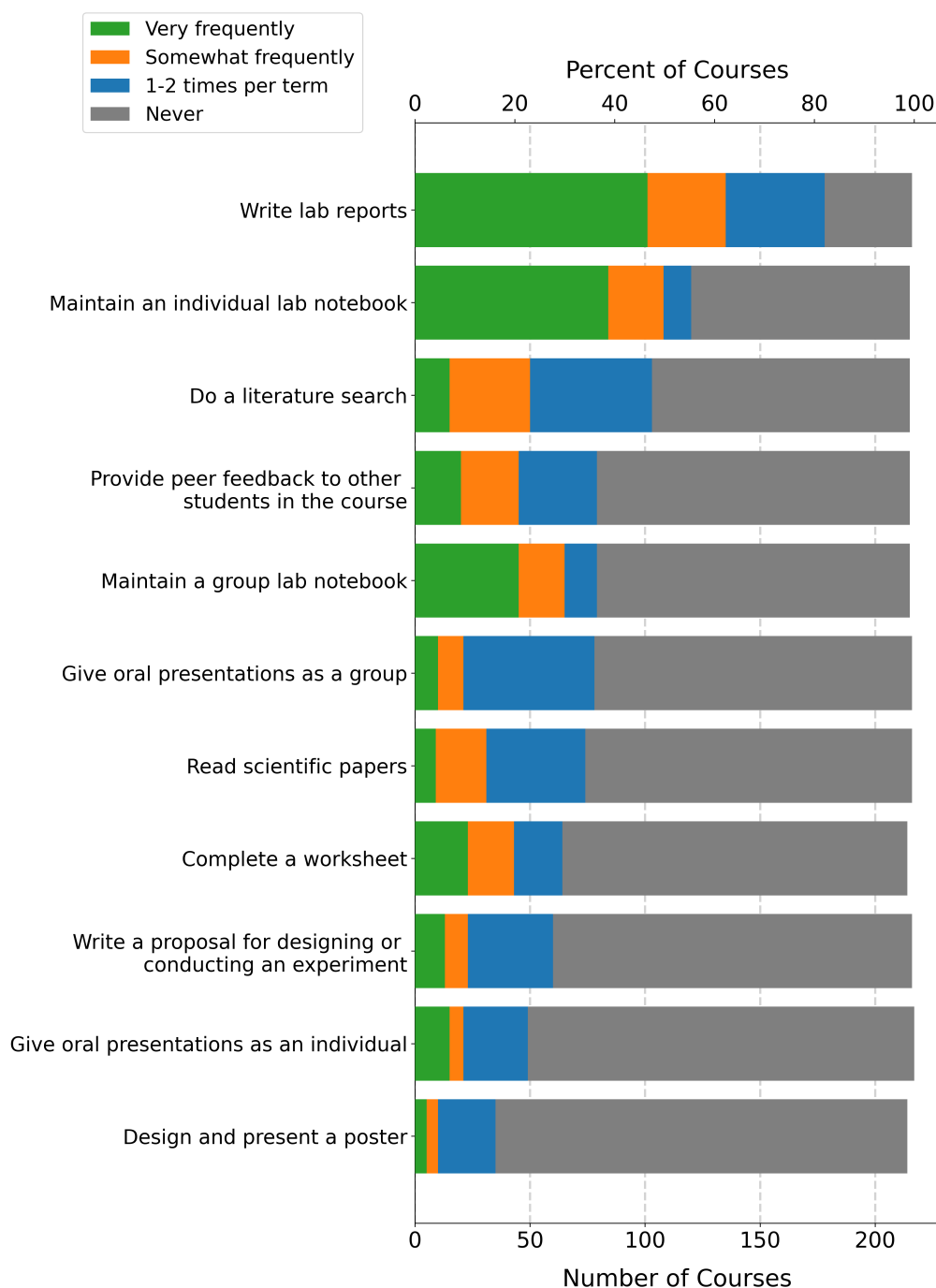
Figure 7.12: Courses with communication activities [N = 214 - 217]. Bars represent number of courses (bottom axis) and percent of courses (top axis) that include various activities related to communication skills, such as peer feedback and lab reports. Bars are split based on frequency of the activity - very frequently (left, green), somewhat frequently (second from left, orange), 1-2 times per term (second from right, blue), and never (right, gray). Writing lab reports is a common activity, with more than 80% of courses indicating that student engage in this to some extent. Very few courses have students present posters, give presentations, or write proposals for experiments.
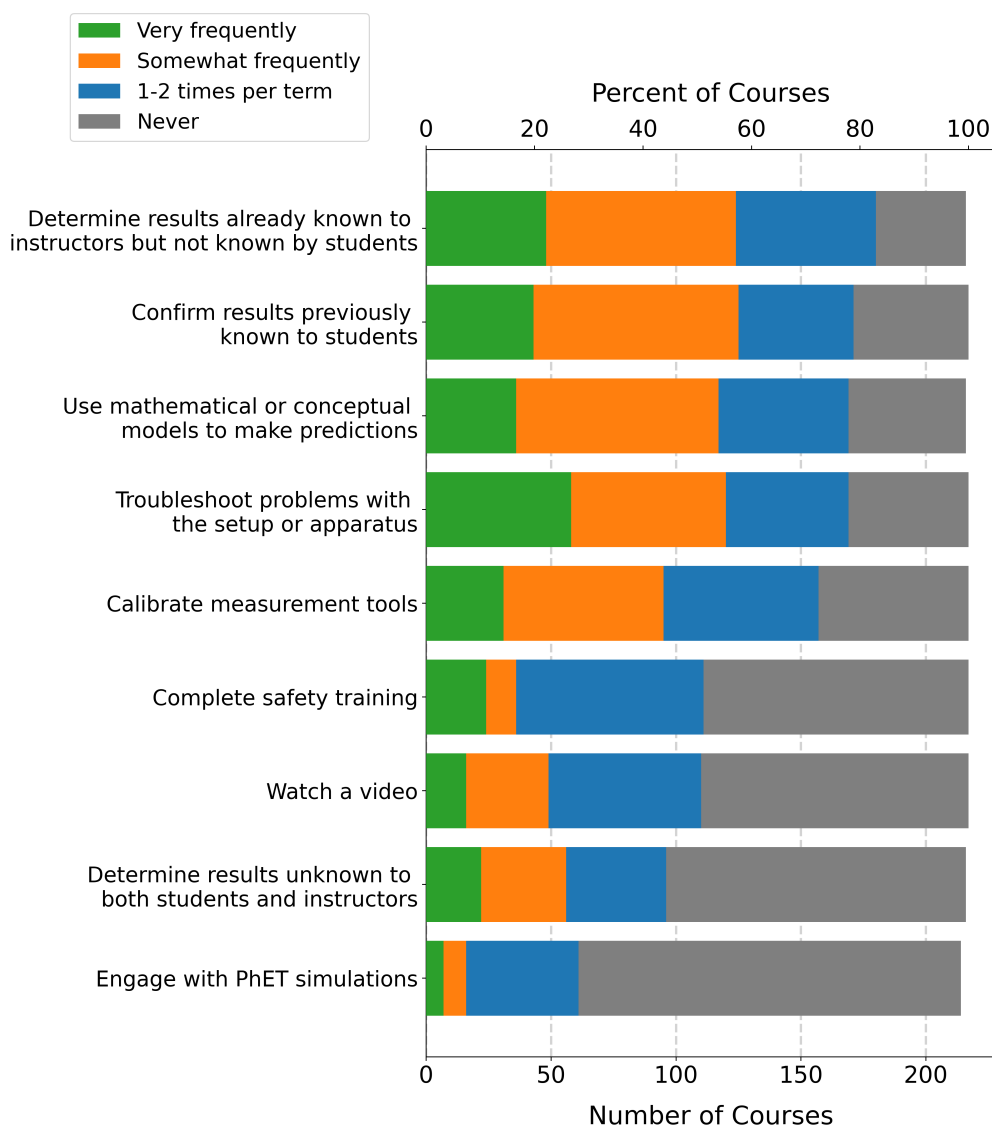
Figure 7.13: Courses with student decision-making activities [N = 215 - 217]. Bars represent number of courses (bottom axis) and percent of courses (top axis) that include various activities related to decisions made by students, including choosing their own procedures and analysis methods. Bars are split based on frequency of the activity - very frequently (left, green), somewhat frequently (second from left, orange), 1-2 times per term (second from right, blue), and never (right, gray). Students engage in these activities in many courses, although in most cases, they are only doing these 1-2 times per term as opposed to other activity categories that have more responses in the very frequently category.

together.

We find that courses typically engage in a higher percentage of activities related to a goal than items graded related to that goal; this percentage is often much larger (in some cases, more than four times). There is no correlation between having a certain goal for the course and the percentage of activities related to that goal or the percentage of items graded related to that goal, thus showing that instructors take many different paths in attempting to achieve their course goals.
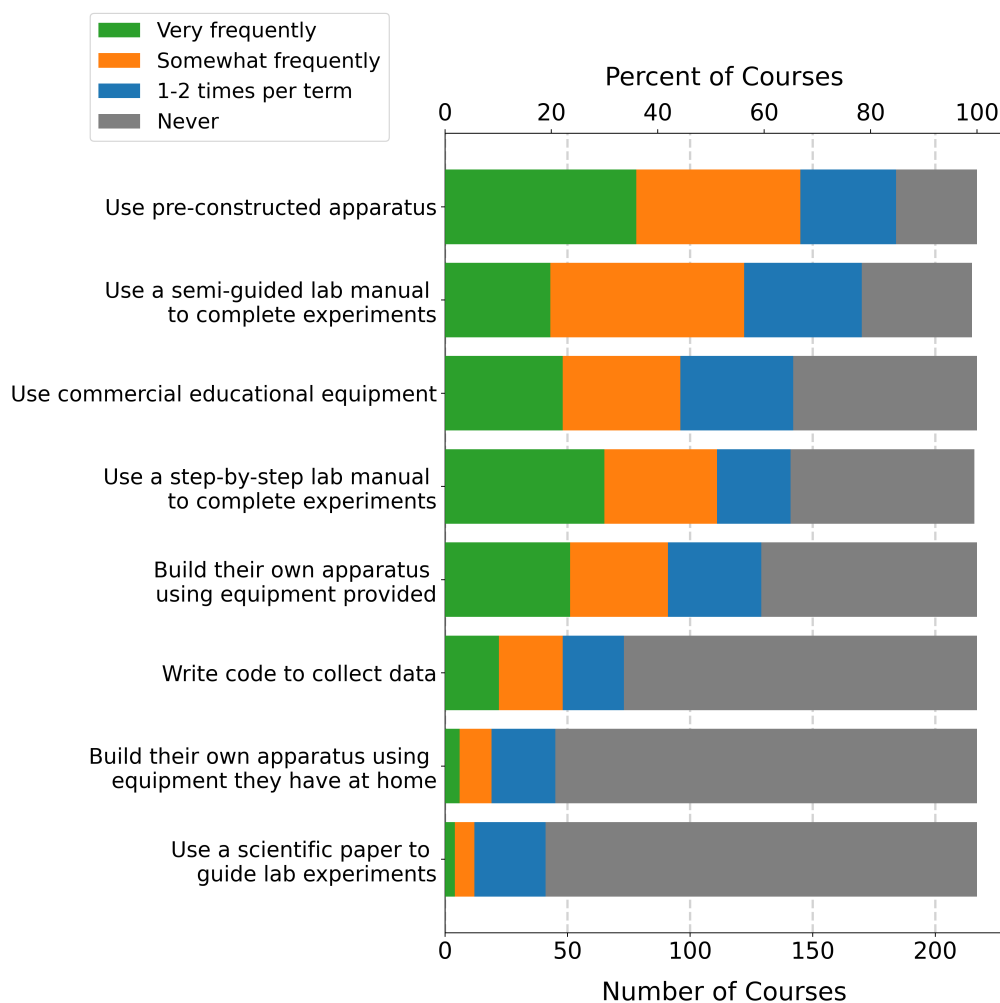
Figure 7.14: Students engagement with materials in lab courses [N = 215 - 217]. Bars represent number of courses (bottom axis) and percent of courses (top axis) that include various methods of student interaction with materials, including use of commercial equipment (such as PASCO or TeachSpin), as well as students building their own apparatus. Bars are split based on frequency of engagement with the activity - very frequently (left, green), somewhat frequently (second from left, orange), 1-2 times per term (second from right, blue), and never (right, gray).

7.5   Summary and Future Research

Here, we present the development of a survey designed to collect data that will allow us to create a taxonomy of lab courses with additional data collection. The project goals include gathering input through interviews with lab instructors across the world to both understand the scope of lab instruction, and to gather input to the development of a research-based survey, which

Figure 7.15: Courses with modeling and other activities [N = 214 - 217]. Bars represent number of courses (bottom axis) and percent of courses (top axis) that include various activities related to modeling, such as using models to make predictions, as well as other activities such as watching a video or completing a safety training. Bars are split based on frequency of engagement with the activity - very frequently (left, green), somewhat frequently (second from left, orange), 1-2 times per term (second from right, blue), and never (right, gray).

aims to capture the structure, goals, and activities in lab classes throughout the world. We analyzed survey results, including responses from 217 courses in 41 countries, to present an initial view of physics lab courses worldwide.

This global study found many similarities between these courses — for example, in almost

Table 7.9: Items included in final course grade [N = 216]. This multiple response item allows respondents to select multiple items, as well as an option to write in anything not listed.

	Num. Responses	Percent Responses
Lab report	160	74.1
Attendance/participation	117	54.2
Lab notebooks	99	45.8
Accuracy/precision of results	88	40.7
Oral presentation	69	31.9
Observation of students	62	28.7
Prelab calculations	59	27.3
Written exam	51	23.6
Interview/meeting after the lab	40	18.5
Worksheets	33	15.3
Partial lab report	31	14.4
Prelab measurement/analysis plan	29	13.4
Practical exam	27	12.5
Quiz/interview prior to working	24	11.1
Poster presentation	22	10.2
Prelab quiz	22	10.2
Peer feedback	19	8.8
Prelab video	18	8.3
Not Listed	31	14.4

all courses, students work with at least one partner. Additionally, a goal of nearly all courses is for students to develop technical knowledge and skills. We also find many differences between these courses, such as the number and types of goals of the course, the activities students participate in, and the student-to-staff ratio. Further, we find that there is no significant difference between introductory and beyond-introductory courses in terms of the number of course goals they have. We also find that in terms of data analysis and student decision-making, at least half of all courses participate in some extent in all activities in these categories.

We also find interesting results about the level of guidance provided to students. For example, students rarely use a scientific paper to guide their lab experiments - in most cases they either use a step-by-step lab manual or a semi-guided lab manual. In more than 80% of courses, students use a pre-constructed apparatus to some extent. Further, in most courses (more than 80%), students spend some time determining results already known to the instructor, but not yet known to the

students, though in nearly 80% of courses, students engage with activities where they are confirming results they have already learned in a lecture course. More research on open-ended lab courses and how they differ from traditional courses could be useful for understating the spectrum of guidance students receive in lab courses.

The results of this study have also raised several new questions in investigating undergraduate lab courses. For example, we find that students tend to work with other students in these courses, but we do not know why this is. It might be due to equipment limitations, logistical constraints due to two or more people being necessary to actually perform the experiment, pedagogical reasons, or perhaps simply tradition. We frequently assume that working collaboratively has pedagogical benefits, but this study does not investigate those, though group work in lab courses is addressed in other PER literature [60, 190, 222, 223, 246]. Further, we have no information about how well instructors are meeting their course goals. Because the average number of goals per course is so high (mean = 11.8), it seems unlikely that instructors can focus equal time to all of these goals. Even considering major goals only, courses have a mean of 6.9 major goals, which is a significant number of goals for a short course that might only meet a few hours per week for a term. We have no detailed information about actions instructors take in order to meet these goals aside from knowing some of the activities and graded items that might relate to these goals. Thus, the initial results of this survey suggest many ideas for future PER research in lab courses.

We hope to continue data collection in order to make more claims based on an expanded data set and eventually build a taxonomy of laboratory courses in order to help classify these courses and make comparisons easier. We need at least an order of magnitude more data in order to accomplish this goal, as we would eventually like to use a clustering analysis in order to analyze the data and find clusters that represent different types of courses [61, 79].

Further, if we collect a significant amount of data in individual countries, we would like to analyze the landscape of undergraduate physics laboratory courses in these specific countries; this would require many more responses from each individual country.

Finally, with more data, we can present a more complete view of the landscape of undergrad-

uate physics lab courses worldwide to give instructors and researchers a broad perspective as they work to improve physics laboratory instruction globally.

Table 7.10: Matching of goals with activities and items graded. The number of courses represents those who selected the goal as either a major or minor goal for their course. This table shows, if an instructor selects a goal, how many activities and items graded they have selected on average (mean) that match that goal. These are shown as fractions as well as percentages. The fractions allow visualization of the total number of matched activities and items graded for each goal, while the percentages allow for comparison between these matched items more easily. One goal (enjoyment of experimental physics and/or the course) is not shown in this table because no activities or items graded match that goal. Additionally, the goal related to approximations has no items graded matched with it (though it does have matched activities). In general, we find that instructors have a higher percentage of activities for a specific goal than items graded for that goal.

	Num. Courses	Activities		Items Graded	
		Fraction	Percent	Fraction	Percent
Developing mathematical model(s) of experimental results	158	5.7/7	81	1.8/5	36
Making quick and simple approximations to predict experimental outcomes	153	1.6/2	80	N/A	N/A
Learning how to analyze and interpret data	205	8.4/11	76	1.6/5	32
Reinforcing physics concepts previously seen in lecture	170	7.0/10	70	2.8/10	28
Developing scientific writing skills	172	1.3/2	65	1.2/3	40
Learning physics concepts not previously seen in lecture	152	9.1/14	65	2.2/6	37
Learning how to visualize data	207	2.5/4	63	1.4/4	35
Developing expert-like views about the nature of the process of doing experimental physics	153	18.2/30	61	2.4/8	30
Designing experiments	147	5.8/10	58	0.59/2	30
Developing lab notebook keeping skills	165	1.1/2	55	0.57/1	57
Developing technical knowledge and skills	212	3.2/6	53	1.4/4	35
Reflecting on and evaluating one's own learning (metacognition)	143	0.44/1	44	0.10/1	10
Developing collaboration and teamwork skills	196	1.2/3	40	0.66/2	33
Developing other communication skills	125	2.0/5	40	1.1/5	22

Chapter 8

Conclusions and Future Work

8.1 SPRUCE

8.1.1 Conclusions

Through this dissertation, I have shown the development of a new assessment, SPRUCE, designed to evaluate students' proficiency with measurement uncertainty in introductory physics laboratory courses. I further discussed a novel scoring scheme and its applications to SPRUCE, statistical validation of SPRUCE using classical test theory, and several important outcomes of analyzing data from SPRUCE.

Prior work determined the aspects of measurement uncertainty important to instructors [184]. These aspects were then turned into assessment objectives (AOs) for SPRUCE, followed by the creation of SPRUCE items centered around these AOs. An iterative process of student interviews and beta testing, refinement of items, and refinement of AOs occurred until the final form of SPRUCE emerged (see Chapter 3 for more details about the development of SPRUCE).

Next, when scoring SPRUCE, it became clear that a traditional scoring scheme would result in a significant loss of potential information contained within student responses. Thus, couplet scoring was developed. In this scheme, items are matched with all AOs they address and scored along each AO, producing couplets (instead of traditional item scores). This eventually leads to both fine-grained AO scores and a course-grained overall score for SPRUCE. The AO scores allow instructors and researchers to determine individual areas of measurement uncertainty in which

students do well or poorly (along with how students improve post-instruction in these individual areas), while the overall score, which weights each AO equally, provides information about students' overall proficiency with measurement uncertainty. Chapter 2 provides many more details about this scoring scheme, including the affordances and limitations it provides compared with more traditional scoring scheme.

After the development of both SPRUCE and the novel scoring scheme, we provided evidence for the validity and reliability of SPRUCE using classical test theory. This work examines the validity and reliability of SPRUCE at the couplet level, AO level, and overall for SPRUCE. It is essential to establish the validity and reliability of an assessment before conclusions from student data are drawn from it. These concepts ensure that the assessment measures what it is intended to (in this case, student proficiency with measurement uncertainty) and that it does so in a way that is reproducible across student populations. The details of this validation are provided in Chapter 4.

Finally, once SPRUCE was developed into its final form and validated, student data can be examined to explore student understanding of measurement uncertainty, both at a course-grained level and at the level of each individual AO. Through this work, I first examined post-instruction results only to provide a general overview of student proficiency after at least one semester of an undergraduate physics laboratory course. I found general trends in student understanding of concepts; for example, most students excel at reporting the mean of a distribution as their final result, whereas most struggle with comparing measurements with uncertainty.

After examining post-instruction data only, I analyzed the impact of instruction by looking at changes from pre-test to post-test results. Students do tend to improve after a semester of instruction, both at the overall score level and at the fine-grained AO level. I then determined the effects of major and gender on overall student performance, as well as the effects of these and the importance of the AO to an instructor on the AO-level performance. I found that major and gender tend to be significant predictors of student post-test scores, while the importance of an AO to an instructor typically is not a predictor of student performance.

Finally, I chose several AOs to examine deeply for student reasoning elements (provided in

student interviews) as well as student performance. Details of this and other student results are provided in Chapters 5 and 6.

In short, we find that students excel at certain areas of measurement uncertainty, such as reporting the mean of a distribution as the best approximation of a measurement, but struggle in other areas, such as propagating error and comparing measurements. Additionally, the importance an instructor places on specific areas of measurement uncertainty are uncorrelated with student performance in those areas. Finally, students' majors and genders do correlate with their performance on SPRUCE, with men generally outperforming women and physics majors outperforming other majors.

8.1.2   Future Work

SPRUCE is still actively collecting data. Future statistical validation can be done using item response theory (IRT) [61]. This method can estimate the characteristic parameters of items, or in our case, couplets, as well as determine students' latent abilities. The characteristic parameters of couplets in IRT are difficulty and discrimination. IRT assumes one unidimensional ability that affects students' responses to all couplets on SPRUCE and provides more information on the couplet level than CTT validation alone. While CTT focuses on a full-test validation, IRT operates at the item (or couplet) level. IRT models the response of each student of a given "ability" for each couplet on the assessment. IRT is based on the idea that the probability of a correct response to a couplet is a function of both the person and the couplet parameters. Thus, the core idea is to fit the following model for each couplet:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}, \tag{8.1}$$

where a and b are the couplet's difficulty and discrimination, respectively, and θ is the student's latent abilities.

Because of the format of SPRUCE and its scoring scheme, this validation of SPRUCE will likely have to be in the form of multidimensional IRT (MIRT). IRT assumes unidimensionality,

but SPRUCE, with its 10 AOs, is a multidimensional assessment. While it does measure one construct (measurement uncertainty), this construct is so broad that it is not entirely accurate to call SPRUCE unidimensional. Thus, MIRT can be useful in this further validation of SPRUCE. In other words, while IRT assumes that student ability is unidimensional (θ), MIRT does not require this assumption and instead allows for multidimensional student abilities. It combines elements of both factor analysis (discussed in Chapter 3) and IRT in order to extend IRT beyond unidimensionality. MIRT is a method of modeling the interactions between students and assessment couplets and attempts to determine the significant contributors to these interactions [197]. This method has been applied to the Force Concept Inventory (FCI) in order to help determine how many factors (or concepts) this assessment entails [219] using different methodology than previous studies, which mainly focused on factor analysis [209], cluster analysis [218], and IRT [242]. A MIRT analysis of SPRUCE can help validate its dimensionality, whether that be ten (determined by the AOs themselves), three (determined by the grouping of the AOs into sources of uncertainty, handling of uncertainty, and distributions and repeated measurements), or some other number or grouping that would have to be conceptually explained.

Further work in SPRUCE also entails a deeper look at the AOs and student proficiency in each. For example, two of the AOs deal with accuracy and precision, concepts students frequently confuse. SPRUCE also includes an item that is not scored in order to help calibrate student ideas around accuracy and precision. It presents a set of bullseye targets and asks students to select which one represents high precision and low accuracy. From this calibration, future work can examine how student ideas on the AOs related to accuracy and precision follow. This analysis can include data from previous student interviews in order to gain insight into student reasoning elements, as well as data from SPRUCE administrations. Further, other AOs of interest might be examined in more detail, especially those that were not previously examined in Chapters 5 and 6.

Finally, we can tie in the lab taxonomy work detailed in Chapter 7 to an analysis of SPRUCE data: once we are able to create a full taxonomy of lab courses, we can examine how students in different types of courses perform on SPRUCE.

8.2 Lab Taxonomy

8.2.1 Conclusions

Through this dissertation, I have described work related to building a taxonomy of under-graduate physics laboratory courses. Towards this end, I have created a survey designed to capture the significant aspects of these courses in order to aid in developing a classification scheme. While there is still not enough data from the survey to be able to make this scheme yet, I was able to present a global view of these courses in Chapter 7. This is the most global PER study in laboratory courses to date, consisting of data from 217 courses in 41 countries.

The creation of the survey itself is a result of this work, requiring interviews with instructors in 22 countries to validate the survey – both to ensure it is understandable to those who don't speak English as a first language, as well as to ensure it is applicable to most courses and that we can understand what a response to each question means.

From the data collected, we learn that there are several similarities and differences between the courses. Students typically work in groups (not alone) around the world, and most courses aim to help students develop their technical and hands-on lab skills. However, many of the goals and activities students participate in differ from course to course. Further, there are many difference in course sizes and number of staff present to help students.

Finally, both the survey itself and papers detailing results of the survey allow instructors interested in improving their courses to gather ideas and resources. The survey includes links to a variety of branded instruction styles and assessments. Further, the lists of potential course goals and activities can spark new ideas and methods for instructors.

Thus, while this work does not present information about the overall goal of creating a taxonomy of courses, it details information about courses around the world and provides a starting point towards the creation of this scheme.

8.2.2   Future Work

Further data collection can aid in achieving the overall goals of this project. First, if we can collect at least 1000 responses, a clustering analysis would provide insights into how certain groups correlate with one another; this would allow us to develop different groups of responses and examine how they relate to each other. Cluster analyses organize multivariate data into different subgroups, which can then be examined for conceptual similarities within a group and differences between groups in order to assign names to these clusters [79]. These clusters would represent different types of courses, and future courses could then be assigned to these clusters.

The clusters obtained from this analysis can then be used for the purposes of providing more accurate comparison data on instructor reports from RBAIs. If an instructor also fills out the lab taxonomy survey, their course could be classified into a cluster and then the instructor report they receive at the end of the semester with their students' results could include comparison data only from other courses within their cluster. This would aid in helping instructors fully understand the results of these assessments. Further, researchers could use these data to determine whether certain clusters outperform others and attempt to apply these findings to other courses in order to improve instruction.

Next, more data would allow us to perform analysis based on geography. There are many different analyses that can be done if we could split the data by country or region. If we examine data from only the United States, for example, we could analyze the ways in which undergraduate physics labs differ across the wide variations in institutions in the country. One example of this could be looking at whether courses at community colleges are similar to courses at four year colleges in order to determine the level of preparation provided at two-year colleges for students aiming to transfer to four-year institutions. We could also compare courses at similar universities – for example, between R1 institutions – in order to compare the types of courses happening at these places. In smaller countries, enough data provided would allow researchers to provide a landscape of the courses in that particular country.

8.3   Summary

Thus, through this dissertation, I have presented both work towards understanding student proficiency with measurement uncertainty and the development of a survey with the goal of creating a classification scheme for undergraduate physics laboratory courses. Both of these works aims to improve instruction in the undergraduate physics laboratory setting. While future work is necessary for fully realizing these goals, the work presented here is a significant step forward for both of these. My hope is that future researchers can expand on both analysis of SPRUCE results, as well as results from the lab taxonomy survey to aid instructors in providing the best possible experiences for students in lab courses.

Bibliography

[1] Advanced laboratory physics association (ALPhA). `https://advlab.org` [Accessed: March 25, 2024].

[2] American physical society forum on education. `https://engage.aps.org/fed/home` [Accessed: March 25, 2024].

[3] American physical society topical group on physics education research. `https://engage.aps.org/gper/home` [Accessed: March 25, 2024].

[4] Arbeitsgruppe physikalische praktika (agpp). `https://www.dpg-physik.de/vereinigungen/fachuebergreifend/ag/agpp` [Accessed: April 18, 2024].

[5] Girep thematic group on laboratory based teaching in physics (LabTiP). `https://www.girep.org/thematic-groups/laboratory-based-teaching-in-physics/` [Accessed: April 19, 2024].

[6] Groupe international de recherche sur l'enseignement de la physique - girep. `https://www.girep.org` [Accessed: March 25, 2024].

[7] Guide to instructional laboratories and experimental skills. Accessed on January 18, 2023, https://ep3guide.org/guide-overview/instructional-laboratories-and-experimental-skills.

[8] Jiscmail - physics education. `https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=PHYSICS-EDUCATION` [Accessed: March 25, 2024].

[9] Lab taxonomy survey, qualtrics. `https://cuboulder.qualtrics.com/jfe/form/SV_ehWi7ucbFNk1kFM` [Accessed: January 10, 2024].

[10] LASSO: Learning About STEM Student Outcomes. https://learningassistantalliance.org/.

[11] PhysPort Assessment Resources. https://www.physport.org/.

[12] SPRUCE for Instructors | JILA - Exploring the Frontiers of Physics. https://jila.colorado.edu/lewandowski/research/spruce-instructors.

[13] Thinking Critically in Physics Labs. Accessed on April 11, 2024, https://www.physport.org/curricula/thinkingcritically/.

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

[15] Analyzing and Interpreting Data | Next Generation Science Standards, 2022.

[16] Focused collection of Physical Review PER: Instructional labs-improving traditions and new directions, March 30, 2023 2022. `https://journals.aps.org/prper/collections/PER-LAB`.

[17] Abdi, H. Factor rotations in factor analyses. Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA (2003), 792–795.

[18] Adams, W. K. The Design and Validation of the Colorado Learning Attitudes about Science Survey. In AIP Conference Proceedings (2005), vol. 790, AIP, pp. 45–48.

[19] Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., and Wieman, C. E. New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. Physical Review Special Topics - Physics Education Research 2, 1 (Jan. 2006), 010101.

[20] Adams, W. K., and Wieman, C. E. Development and validation of instruments to measure learning of expert-like thinking. International journal of science education 33, 9 (2011), 1289–1312. Publisher: Taylor & Francis.

[21] Adlakha, V., and Kuo, E. Critical issues in statistical causal inference for observational physics education research. Physical Review Physics Education Research 19, 2 (Nov. 2023), 020160.

[22] Alemani, M. The redesign of an introductory physics laboratory course. Il Nuovo Cimento C 46, 5 (Aug. 2023), 1–4.

[23] American Educational Research Association, Ed. Report and recommendations for the reauthorization of the institute of education sciences. American Educational Research Association, Washington, D.C, 2011.

[24] Anderson, L. W., and Krathwohl, D. R. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Longman, 2001.

[25] Anderson, T. W., and Darling, D. A. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. The Annals of Mathematical Statistics 23, 2 (June 1952), 193–212.

[26] Aubrecht, G. J., and Aubrecht, J. D. Constructing objective tests. American Journal of Physics 51, 7 (July 1983), 613–620.

[27] Bearden, I., Dvořák, L., and Planinšič, G. Work group 2 position paper: Experiments and laboratory work in teacher education. Journal of Physics: Conference Series 2297, 1 (jun 2022), 012008.

[28] Beichner, R., Bernold, L., Burniston, E., Dail, P., Felder, R., Gastineau, J., Gjertsen, M., and Risley, J. Case study of the physics component of an integrated curriculum. American Journal of Physics 67, S1 (July 1999), S16–S24.

[29] Beichner, R. J. History and Evolution of Active Learning Spaces. New Directions for Teaching and Learning 2014, 137 (2014), 9–16.

[30] Beichner, R. J., Saul, J. M., Abbot, D. S., Morse, J. J., Deardorff, D. L., Allain, R. J., Bonham, S. W., Dancy, M. H., and Risley, J. S. The student-centered activities for large enrollment undergraduate programs (scale-up) project. In Reviews in PER Volume 1: Research-Based Reform of University Physics, E. Redish and P. Cooney, Eds. American Association of Physics Teachers, College Park, MD, 2007.

[31] Beichner, R. J., Saul, J. M., Allain, R. J., Deardorff, D. L., and Abbott, D. S. Introduction To Scale Up: Student Centered Activities For Large Enrollment University Physics. pp. 5.411.1–5.411.12.

[32] Ben-Akiva, M. E., and Lerman, S. R. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, 1985.

[33] Borish, V., Werth, A., Sulaiman, N., Fox, M. F., Hoehn, J. R., and Lewandowski, H. J. Undergraduate student experiences in remote lab courses during the COVID-19 pandemic. Physical Review Physics Education Research 18, 2 (July 2022), 020105.

[34] Bouquet, F., Bobroff, J., Fuchs-Gallezot, M., and Maurines, L. Project-based physics labs using low-cost open-source hardware. American Journal of Physics 85, 3 (Mar. 2017), 216–222.

[35] Bradbury, F. R., and Pols, F. A Pandemic-Resilient Open-Inquiry Physical Science Lab Course Which Leverages the Maker Movement. Electronic Journal for Research in Science & Mathematics Education 24, 3 (2020), 60–67. ERIC Number: EJ1285251.

[36] Brewe, E. Modeling theory applied: Modeling Instruction in introductory physics. American Journal of Physics 76, 12 (12 2008), 1155–1160.

[37] Brewe, E., Bruun, J., and Bearden, I. G. Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data. Physical Review Physics Education Research 12, 2 (Sept. 2016), 020131. Publisher: American Physical Society.

[38] Brewe, E., Kramer, L., and O'Brien, G. Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS. Physical Review Special Topics - Physics Education Research 5, 1 (June 2009), 013102.

[39] Brewe, E., Kramer, L., O'Brien, G., Henderson, C., Sabella, M., and Hsu, L. CLASS Shifts in Modeling Instruction. In AIP Conference Proceedings (Edmonton, Alberta (Canada), 2008), AIP, pp. 79–82.

[40] Brookes, D. T., Ektina, E., and Planinsic, G. Implementing an epistemologically authentic approach to student-centered inquiry learning. Physical Review Physics Education Research 16 (Dec. 2020), 020148. ADS Bibcode: 2020PRPER..16b0148B.

[41] Buffler, A., Allie, S., and Lubben, F. The development of first year physics students' ideas about measurement in terms of point and set paradigms. International Journal of Science Education 23, 11 (Nov. 2001), 1137–1156.

[42] Buggé, D., and Etkina, E. The long-term effects of learning in an ISLE approach classroom. In Physics Education Research Conference 2020 (Virtual Conference, July 2020), PER Conference, pp. 63–68.

[43] Caballero, M. D., Dounas-Frazer, D. R., Lewandowski, H. J., and Stetzer, M. R. Labs are Necessary, and We Need to Invest in Them. APS News 27, 5 (2018).

[44] Caballero, M. D., Greco, E. F., Murray, E. R., Bujak, K. R., Jackson Marr, M., Catrambone, R., Kohlmyer, M. A., and Schatz, M. F. Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study. American Journal of Physics 80, 7 (July 2012), 638–644. Publisher: American Association of Physics Teachers.

[45] Campbell, B., Lubben, F., Buffler, A., and Sallih, A. Teaching Scientific Measurement at University: Understanding Student's Ideas and Laboratory Curriculum Reform. Southern African Association for Research in Mathematics, Science and Technology Education, 2005.

[46] Chabay, R., and Sherwood, B. Qualitative understanding and retention. AAPT Announcer 27, 2 (1997), 96.

[47] Chabay, R., and Sherwood, B. Restructuring the introductory electricity and magnetism course. American Journal of Physics 74, 4 (Apr. 2006), 329–336.

[48] Chen, S., Lo, H.-C., Lin, J.-W., Liang, J.-C., Chang, H.-Y., Hwang, F.-K., Chiou, G.-L., Wu, Y.-T., Lee, S. W.-Y., Wu, H.-K., Wang, C.-Y., and Tsai, C.-C. Development and implications of technology in reform-based physics laboratories. Physical Review Special Topics - Physics Education Research 8, 2 (Oct. 2012), 020113.

[49] Christman, E., Miller, P., and Stewart, J. Beyond normalized gain: Improved comparison of physics educational outcomes. Physical Review Physics Education Research 20, 1 (Apr. 2024), 010123.

[50] Coelho, S. M., and Séré, M. Pupils' Reasoning and Practice during Hands-on Activities in the Measurement Phase. Research in Science & Technological Education 16, 1 (May 1998), 79–96.

[51] Cohen, J. Statistical Power Analysis for the Behavioral Sciences. L. Erlbaum Associates, 1988, pp. 20–27.

[52] Costello, A., and Osborne, J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Practical Assessment, Research, and Evaluation 10, 1 (Nov. 2019).

[53] Cox, R., and Brna, P. Supporting the use of external representations in problem solving: The need for flexible learning environments. Journal of Artificial Intelligence in Education 6, 2-3 (1995), 239–302. Place: US Publisher: Assn for the Advancement of Computing in Education.

[54] Crocker, L., and Algina, J. Introduction to Classical and Modern Test Theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887, 1986, p. 34. ERIC Number: ED312281.

[55] Crocker, L. M., and Algina, J. Introduction to classical and modern test theory. Cengage Learning, Mason, Ohio, 2008. OCLC: 268675245.

[56] CUMMINGS, K., MARX, J., THORNTON, R., AND KUHL, D. Evaluating innovation in studio physics. American Journal of Physics 67, S1 (July 1999), S38–S44.

[57] CURETON, E. E., AND MULAIK, S. A. The weighted varimax rotation and the promax rotation. Psychometrika 40, 2 (June 1975), 183–195.

[58] DAY, J., AND BONN, D. Development of the Concise Data Processing Assessment. Physical Review Special Topics - Physics Education Research 7, 1 (June 2011), 010114.

[59] DAY, J., STANG, J. B., N. G. HOLMES, KUMAR, D., AND D.A. BONN. Gender gaps and gendered action in a first-year physics laboratory. Physical Review Physics Education Research 12, 2 (Aug. 2016), 020104.

[60] DEW, M., HUNT, E., PERERA, V., PERRY, J., PONTI, G., AND LOVERIDGE, A. Group dynamics in inquiry-based labs: Gender inequities and the efficacy of partner agreements. Physical Review Physics Education Research 20, 1 (Apr. 2024), 010121.

[61] DING, L., AND BEICHNER, R. Approaches to data analysis of multiple-choice questions. Physical Review Special Topics - Physics Education Research 5, 2 (Sept. 2009), 020103.

[62] DORAN, R. L. Basic Measurement and Evaluation of Science Instruction. National Science Teachers Association, 1742 Connecticut Ave, 1980, pp. 97–104. ERIC Number: ED196733.

[63] DOUCETTE, D., CLARK, R., AND SINGH, C. Students' attitudes toward experimental physics in a conceptual inquiry-based introductory physics lab. Canadian Journal of Physics 100, 6 (2022), 292–302.

[64] DOUNAS-FRAZER, D. R., AND H. J. LEWANDOWSKI. The Modelling Framework for Experimental Physics: description, development, and applications. European Journal of Physics 39, 6 (Oct. 2018), 064005. Publisher: IOP Publishing.

[65] DOUNAS-FRAZER, D. R., RÍOS, L., POLLARD, B., STANLEY, J. T., AND H. J. LEWANDOWSKI. Characterizing lab instructors' self-reported learning goals to inform development of an experimental modeling skills assessment. Physical Review Physics Education Research 14, 2 (Nov. 2018), 020118. Publisher: American Physical Society.

[66] DUFRESNE, R. J., GERACE, W. J., AND LEONARD, W. J. Solving physics problems with multiple representations. The Physics Teacher 35, 5 (May 1997), 270–275.

[67] DURBIN, J., AND WATSON, G. S. Testing for Serial Correlation in Least Squares Regression: I. Biometrika 37, 3/4 (1950), 409–428.

[68] DURBIN, J., AND WATSON, G. S. Testing for Serial Correlation in Least Squares Regression. II. Biometrika 38, 1/2 (1951), 159–177.

[69] DURBIN, J., AND WATSON, G. S. Testing for Serial Correlation in Least Squares Regression. III. Biometrika 58, 1 (1971), 1–19.

[70] EATON, P., JOHNSON, K., AND WILLOUGHBY, S. Generating a growth-oriented partial credit grading model for the Force Concept Inventory. Physical Review Physics Education Research 15, 2 (Dec. 2019), 020151. Publisher: American Physical Society.

[71] EBLEN-ZAYAS, M. The impact of metacognitive activities on student attitudes towards experimental physics. pp. 104–107. ISSN: 2377-2379.

[72] ENGELHARDT, P. V. An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests. In Getting Started in PER, C. Henderson and K. Harper, Eds., vol. 2. American Association of Physics Teachers, College Park, April 2009.

[73] ESHACH, H., AND KUKLIANSKY, I. Developing of an instrument for assessing students' data analysis skills in the undergraduate physics laboratory. Canadian Journal of Physics 94, 11 (Nov. 2016), 1205–1215.

[74] ETKINA, E. When learning physics mirrors doing physics. Physics Today 76, 10 (Oct. 2023), 26–32.

[75] ETKINA, E., KARELINA, A., RUIBAL-VILLASENOR, M., ROSENGRANT, D., JORDAN, R., AND HMELO-SILVER, C. E. Design and Reflection Help Students Develop Scientific Abilities: Learning in Introductory Physics Laboratories. Journal of the Learning Sciences 19, 1 (Jan. 2010), 54–98.

[76] ETKINA, E., KARELINA, A., AND VILLASENOR, M. R. Studying Transfer Of Scientific Reasoning Abilities. In AIP Conference Proceedings (Syracuse, New York (USA), 2007), vol. 883, AIP, pp. 81–84.

[77] ETKINA, E., AND VAN HEUVELEN, A. Investigative Science Learning Environment: Using the processes of science and cognitive strategies to learn physics. In Physics Education Research Conference 2001 (Rochester, New York, July 2001), PER Conference.

[78] ETKINA, E., AND VAN HEUVELEN, A. Investigative science learning environment – a science process approach to learning physics. In Reviews in PER Volume 1: Research-Based Reform of University Physics, E. Redish and P. Cooney, Eds., vol. 1. American Association of Physics Teachers, College Park, MD, 2007, pp. 1–48.

[79] EVERITT, B., Ed. Cluster analysis, 5th ed ed. Wiley series in probability and statistics. Wiley, Chichester, West Sussex, U.K, 2011.

[80] FAGEN, A. P., CROUCH, C. H., AND MAZUR, E. Peer Instruction: Results from a Range of Classrooms. The Physics Teacher 40, 4 (Apr. 2002), 206–209.

[81] FEDER, T. Undergraduate labs lag in science and technology. Physics Today 70, 4 (Apr. 2017), 26–29.

[82] FEENSTRA, L., JULIA, C., AND LOGMAN, P. A Lego® Mach–Zehnder interferometer with an Arduino detector. Physics Education 56, 2 (Jan. 2021), 023004.

[83] FOX, M. F., HOEHN, J. R., WERTH, A., AND LEWANDOWSKI, H. J. Lab instruction during the COVID-19 pandemic: Effects on student views about experimental physics in comparison with previous years. Physical Review Physics Education Research 17, 1 (June 2021), 010148.

[84] FOX, M. F. J., POLLARD, B., RÍOS, L., AND LEWANDOWSKI, H. J. Capturing modeling pathways using the Modeling Assessment for Physics Laboratory Experiments. In 2020 Physics Education Research Conference Proceedings (Sept. 2020), American Association of Physics Teachers, pp. 155–160.

[85] Fox, M. F. J., Zwickl, B. M., and Lewandowski, H. J. Preparing for the quantum revolution: What is the role of higher education? Phys. Rev. Phys. Educ. Res. 16 (Oct 2020), 020131.

[86] Gardner, M. 10 trends in science education. The Science Teacher 46, 1 (1979), 30–32.

[87] Gelman, A., and Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Dec. 2006.

[88] Gentile, J., Brenner, K., and Stephens, A., Eds. Undergraduate Research Experiences for STEM Students: Successes, Challenges, and Opportunities. National Academies Press, Washington, D.C., May 2017.

[89] Geschwind, G., Alemani, M., Fox, M. F., Logman, P., Tufino, E., and H. J. Lewandowski. Development of a global landscape of undergraduate physics laboratory courses. Physical Review Physics Education Research (2024). In review.

[90] Geschwind, G., Vignal, M., Caballero, M. D., and H. J. Lewandowski. Evidence for validity and reliability of a research-based assessment instrument on measurement uncertainty. Physical Review Physics Education Research (2024). In review.

[91] Geschwind, G., Vignal, M., Caballero, M. D., and H. J. Lewandowski. Using a research-based assessment instrument to explore undergraduate students' proficiencies around measurement uncertainty in physics lab contexts. Physical Review Physics Education Research (2024). In review.

[92] Geschwind, G., Vignal, M., and Lewandowski, H. J. Representational differences in how students compare measurements. In Physics Education Research Conference 2023 (Sacramento, CA, July 2023), PER Conference, pp. 114–119.

[93] Goertzen, R. M., Brewe, E., and Kramer, L. Expanded Markers of Success in Introductory University Physics. International Journal of Science Education 35, 2 (Jan. 2013), 262–288.

[94] Goldhaber, S., Pollock, S., Dubson, M., Beale, P., Perkins, K., Sabella, M., Henderson, C., and Singh, C. Transforming Upper-Division Quantum Mechanics: Learning Goals and Assessment. In Physics Education Research Conference 2009 (2009), pp. 145–148.

[95] Goodman, L. A. Snowball Sampling. The Annals of Mathematical Statistics 32, 1 (Mar. 1961), 148–170.

[96] Gorsuch, R. L. Factor Analysis, 2 ed. Psychology Press, New York, Oct. 1983, pp. 175–238.

[97] Hahs-Vaughn, D. L., and Lomax, R. G. An Introduction to Statistical Concepts, fourth ed. Routledge, Dec. 2019, pp. 615–635.

[98] Hahs-Vaughn, D. L., and Lomax, R. G. An Introduction to Statistical Concepts, fourth ed. Routledge, Dec. 2019, pp. 126–130.

[99] Hahs-Vaughn, D. L., and Lomax, R. G. An Introduction to Statistical Concepts, fourth ed. Routledge, Dec. 2019, pp. 1017–1018.

[100] HAHS-VAUGHN, D. L., AND LOMAX, R. G. An Introduction to Statistical Concepts, fourth ed. Routledge, Dec. 2019, pp. 997–1063.

[101] HAND, B., AND CHOI, A. Examining the Impact of Student Use of Multiple Modal Representations in Constructing Arguments in Organic Chemistry Laboratory Classes. Research in Science Education 40, 1 (Jan. 2010), 29–44.

[102] HANSEN, J., AND STEWART, J. Multidimensional item response theory and the Brief Electricity and Magnetism Assessment. Physical Review Physics Education Research 17, 2 (Nov. 2021), 020139. Publisher: American Physical Society.

[103] HEMPHILL, J. K., AND WESTIE, C. M. The Measurement of Group Dimensions. The Journal of Psychology 29, 2 (Apr. 1950), 325–342. Publisher: Routledge _eprint: https://doi.org/10.1080/00223980.1950.9916035.

[104] HENDERSON, R., FUNKHOUSER, K., AND CABALLERO, M. D. A longitudinal exploration of students' beliefs about experimental physics. pp. 214–219.

[105] HENDRICKSON, A. E., AND WHITE, P. O. Promax: A Quick Method for Rotation to Oblique Simple Structure. British Journal of Statistical Psychology 17, 1 (1964), 65–70.

[106] HENRIKSSON, J. En analys av hur en undervisning med Investigative Science Learning Environment (ISLE) bör påverka elevers syn på fysik, fysikinlärning och fysikexperiment. Samt en svensk översättning av två Research-Based Assessment Instruments (RBAIs) - CLASS och ECLASS. Bachelor's thesis, Uppsala University, 2020.

[107] HESTENES, D., WELLS, M., AND SWACKHAMER, G. Force concept inventory. The physics teacher 30, 3 (1992), 141–158. Publisher: American Association of Physics Teachers.

[108] HOELLWARTH, C., MOELTER, M. J., AND KNIGHT, R. D. A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms. American Journal of Physics 73, 5 (May 2005), 459–462.

[109] HOFSTEIN, A., AND LUNETTA, V. N. The Role of the Laboratory in Science Teaching: Neglected Aspects of Research. Review of Educational Research 52, 2 (June 1982), 201–217.

[110] HOFSTEIN, A., AND LUNETTA, V. N. The laboratory in science education: Foundations for the twenty-first century. Science Education 88, 1 (2004), 28–54.

[111] HOLMES, N., AND LEWANDOWSKI, H. Investigating the landscape of physics laboratory instruction across North America. Physical Review Physics Education Research 16, 2 (Dec. 2020), 020162.

[112] HOLMES, N. G., AND BONN, D. A. Quantitative Comparisons to Promote Inquiry in the Introductory Physics Lab. The Physics Teacher 53, 6 (09 2015), 352–355.

[113] HOLMES, N. G., OLSEN, J., THOMAS, J. L., AND WIEMAN, C. E. Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content. Physical Review Physics Education Research 13, 1 (May 2017), 010129.

[114] HOLMES, N. G., AND WIEMAN, C. E. Assessing modeling in the lab: Uncertainty and measurement. In 2015 Conference on Laboratory Instruction Beyond the First Year (College Park, MD, Nov. 2015), American Association of Physics Teachers, pp. 44–47.

[115] HOLMES, N. G., AND WIEMAN, C. E. Introductory physics labs: We can do better. Physics Today 71, 1 (01 2018), 38–45.

[116] HOLMES, N. G., WIEMAN, C. E., AND BONN, D. A. Teaching critical thinking. Proceedings of the National Academy of Sciences 112, 36 (Sept. 2015), 11199–11204.

[117] HURD, P. D. New Directions in Teaching Secondary School Science. Rand McNally, Chicago, 1969.

[118] JIRUNGNIMITSAKUL, S., AND WATTANAKASIWICH, P. Assessing student understanding of measurement and uncertainty. Journal of Physics: Conference Series 901, 1 (Sept. 2017), 012121.

[119] KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 3 (Sept. 1958), 187–200.

[120] KANT, J. Simple word cloud generator. Available at https://www.simplewordcloud.com.

[121] KEPPEL, G., AND WICKENS, T. D. Design and Analysis: A Researcher's Handbook, fourth ed. Pearson Prentice Hall, 2004, pp. 311–344.

[122] KETONEN, L., LEHTINEN, A., AND KOSKINEN, P. Assessment designs of instructional labs: A literature review and a design model. Physical Review Physics Education Research 19, 2 (July 2023), 020601.

[123] KIMBERLIN, C. L., AND WINTERSTEIN, A. G. Validity and reliability of measurement instruments used in research. American journal of health-system pharmacy: AJHP: official journal of the American Society of Health-System Pharmacists 65, 23 (Dec. 2008), 2276–2284.

[124] KIRK, R. E. Statistics: An Introduction. Cengage Learning, 2008, pp. 281–282.

[125] KLINE, P. A handbook of test construction: Introduction to psychometric design. A handbook of test construction: Introduction to psychometric design. Methuen, New York, NY, US, 1986, p. 143.

[126] KLINE, P. The New Psychometrics: Science, Psychology and Measurement - 1st Editi. Routledge, New York, NY, US, 1998, p. 29.

[127] KLINE, P. Handbook of Psychological Testing. New York, NY, US, 2000, p. 31.

[128] KLINE, R. Exploratory and Confirmatory Factor Analysis. In Applied Quantitative Analysis in Education and the Social Sciences. Routledge, 2013, ch. 6, pp. 171–207.

[129] KOHL, P. B., AND FINKELSTEIN, N. D. Student representational competence and self-assessment when solving physics problems. Physical Review Special Topics - Physics Education Research 1, 1 (Oct. 2005), 010104.

[130] KOHL, P. B., AND FINKELSTEIN, N. D. Effects of representation on students solving physics problems: A fine-grained characterization. Physical Review Special Topics - Physics Education Research 2, 1 (May 2006), 010106.

[131] KOHL, P. B., AND FINKELSTEIN, N. D. Patterns of multiple representation use by experts and novices during physics problem solving, 2008.

[132] Kohl, P. B., Rosengrant, D., and Finkelstein, N. D. Strongly and weakly directed approaches to teaching multiple representation use in physics. Physical Review Special Topics - Physics Education Research 3, 1 (June 2007), 010108.

[133] Kohlmyer, M. A., Caballero, M. D., Catrambone, R., Chabay, R. W., Ding, L., Haugan, M. P., Marr, M. J., Sherwood, B. A., and Schatz, M. F. Tale of two curricula: The performance of 2000 students in introductory electromagnetism. Physical Review Special Topics - Physics Education Research 5, 2 (Oct. 2009), 020105. Publisher: American Physical Society.

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

[135] Kontro, I. Development of Data Processing Skills of Physics Students in Intermediate Laboratory Courses. In Concepts, Strategies and Models to Enhance Physics Teaching and Learning, E. McLoughlin and P. van Kampen, Eds. Springer International Publishing, Cham, 2019, pp. 101–108.

[136] Kontro, I. Development of Data Processing Skills of Physics Students in Intermediate Laboratory Courses. In Concepts, Strategies and Models to Enhance Physics Teaching and Learning, E. McLoughlin and P. van Kampen, Eds. Springer International Publishing, Cham, 2019, pp. 101–108.

[137] Kontro, I., Heino, O., Hendolin, I., and Galambosi, S. Modernisation of the intermediate physics laboratory. European Journal of Physics 39, 2 (Jan. 2018), 025702.

[138] Kozminski, J., Lewandowski, H., Beverly, N., Lindaas, S., Deardorff, D., Reagan, A., Dietz, R., Tagg, R., EblenZayas, M., and Williams, J. AAPT recommendations for the undergraduate physics laboratory curriculum. American Association of Physics Teachers 29 (2014).

[139] Kung, R. L., and Linder, C. University students' ideas about data processing and data comparison in a physics laboratory course. Nordic Studies in Science Education 2, 2 (2006), 40–53.

[140] Lahme, S. Z., Klein, P., Lehtinen, A., Müller, A., Pirinen, P., Rončević, L., and Sušac, A. Evaluating digital experimental tasks for physics laboratory courses. PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung (Nov. 2023).

[141] Laverty, J. T., and Caballero, M. D. Analysis of the most common concept inventories in physics: What are we assessing? Physical Review Physics Education Research 14, 1 (Apr. 2018), 010123.

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

[143] Levy, S., Kapach, Z., Magen, E., and Yerushalmi, E. Re-defining lab norms via professional learning communities of physics teachers. In Physics Education Research Conference 2020 (Virtual Conference, July 2020), PER Conference, pp. 278–283.

[144] LEWANDOWSKI, H. J., BOLTON, D. R., AND POLLARD, B. Initial impacts of the transformation of a large introductory lab course focused on developing experimental skills and expert epistemology.

[145] LEWANDOWSKI, H. J., AND FINKELSTEIN, N. D. Redesigning a junior-level electronics course to support engagement in scientific practices. pp. 191–194.

[146] LEWANDOWSKI, H. J., HOBBS, R., STANLEY, J. T., DOUNAS-FRAZER, D. R., AND POLLARD, B. Student reasoning about measurement uncertainty in an introductory lab course. pp. 244–247. ISSN: 2377-2379.

[147] LI, J. C.-H. Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. Behavior Research Methods 48, 4 (Dec. 2016), 1560–1574.

[148] LINDELL, R., AND DING, L. Establishing reliability and validity: An ongoing process. AIP Conference Proceedings 1513, 1 (Jan. 2013), 27–29. Publisher: American Institute of Physics.

[149] LOGMAN, P. Engaging theoretically primed students in second year lab courses. Journal of Physics: Conference Series 2727, 1 (mar 2024), 012021.

[150] LOGMAN, P., AND KAUTZ, J. From dublin descriptors to implementation in bachelor labs. Journal of Physics: Conference Series 1929, 1 (may 2021), 012065.

[151] LOUVIERE, J., HENSHER, D., AND SWAIT, J. Stated choice methods: analysis and application, vol. 17. Jan. 2000.

[152] MADSEN, A., MCKAGAN, S. B., AND SAYRE, E. C. Best Practices for Administering Concept Inventories. The Physics Teacher 55, 9 (Dec. 2017), 530–536.

[153] MADSEN, A., MCKAGAN, S. B., AND SAYRE, E. C. Resource Letter RBAI-1: Research-Based Assessment Instruments in Physics and Astronomy. American Journal of Physics 85, 4 (Apr. 2017), 245–264. Publisher: American Association of Physics Teachers.

[154] MAJIET, N., AND ALLIE, S. Student understanding of measurement and uncertainty: probing the mean.

[155] MALONEY, D. P., O'KUMA, T. L., HIEGGELKE, C. J., AND VAN HEUVELEN, A. Surveying students' conceptual knowledge of electricity and magnetism. American Journal of Physics 69, S1 (2001), S12–S23. Publisher: American Association of Physics Teachers.

[156] MANN, H. B., AND WHITNEY, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics 18, 1 (Mar. 1947), 50–60.

[157] MARIC, D., FORE, G. A., NYARKO, S. C., AND VARMA-NELSON, P. Measurement in STEM education research: a systematic literature review of trends in the psychometric evidence of scales. International Journal of STEM Education 10, 1 (June 2023), 39.

[158] MARSHMAN, E., AND SINGH, C. Validation and administration of a conceptual survey on the formalism and postulates of quantum mechanics. Physical Review Physics Education Research 15, 2 (Sept. 2019), 020128.

[159] MAY, J. M. Historical analysis of innovation and research in physics instructional laboratories: Recurring themes and future directions. Physical Review Physics Education Research 19, 2 (Dec. 2023), 020168.

[160] MAZUR, E. Farewell, Lecture? Science 323, 5910 (Jan. 2009), 50–51. Publisher: American Association for the Advancement of Science.

[161] McCLUNG, M. S. Developing proficiency programs in California public schools: Some legal implications and a suggested implementation schedule. Sacramento: California State Department of Education (1978).

[162] McDERMOTT, L. C. Millikan Lecture 1990: What we teach and what is learned—Closing the gap. American Journal of Physics 59, 4 (Apr. 1991), 301–315.

[163] McFADDEN, D. Conditional logit analysis of qualitative choice behavior.

[164] MELTZER, D. E., AND OTERO, V. K. A brief history of physics education in the United States. American Journal of Physics 83, 5 (May 2015), 447–458.

[165] MILLAR, R., LUBBEN, F., GOT, R., AND DUGGAN, S. Investigating in the school science laboratory: conceptual and procedural knowledge and their influence on performance. Research Papers in Education 9, 2 (June 1994), 207–248. Publisher: Routledge _eprint: https://doi.org/10.1080/0267152940090205.

[166] MILLER, G. A., AND CHAPMAN, J. P. Misunderstanding analysis of covariance. Journal of Abnormal Psychology 110, 1 (2001), 40–48.

[167] MILLER, M. D., LINN, R. L., AND GRONLUND, N. E. Measurement and assessment in teaching. Merrill Prentice Hall, 2009.

[168] MISLEVY, R. J., HAERTEL, G., RICONSCENTE, M., RUTSTEIN, D. W., AND ZIKER, C. Evidence-Centered Assessment Design. In Assessing Model-Based Reasoning using Evidence-Centered Design. Springer International Publishing, Cham, 2017, pp. 19–24. Series Title: SpringerBriefs in Statistics.

[169] MISLEVY, R. J., AND RICONSCENTE, M. M. Evidence-Centered Assessment Design: Layers, Structures, and Terminology. Tech. rep., SRI International Center for Technology in Learning, 2005.

[170] MONASTERSKY, R., AND VAN NOORDEN, R. 150 years of Nature: a data graphic charts our evolution. Nature 575, 7781 (Nov. 2019), 22–23.

[171] NARAYANAN, S., SARIN, P., PAWAR, N., AND MURTHY, S. Teaching research skills for experimental physics in an undergraduate electronics lab. Physical Review Physics Education Research 19, 2 (July 2023), 020103.

[172] NELSON, L. S. The Anderson-Darling test for normality. Journal of Quality Technology 30, 3 (07 1998), 298–299.

[173] NUNNALLY, J. C., AND BERNSTEIN, I. H. Psychometric Theory, 3rd edition ed. McGraw-Hill, New York, Jan. 1994.

[174] ORGANTINI, G., AND TUFINO, E. Effectiveness of a Laboratory Course with Arduino and Smartphones. Education Sciences 12, 12 (Dec. 2022), 898.

[175] OSBORNE, J. W. Best Practices in Logistic Regression. SAGE Publications, Ltd, 2015.

[176] OTERO, V. K., AND MELTZER, D. E. 100 Years of Attempts to Transform Physics Education. The Physics Teacher 54, 9 (Dec. 2016), 523–527.

[177] PARAPPILLY, M., HASSAM, C., AND WOODMAN, R. J. Race to improve student understanding of uncertainty: Using LEGO race cars in the physics lab. American Journal of Physics 86, 1 (Jan. 2018), 68–76.

[178] PARNAFES, O., AND DISESSA, A. Relations between Types of Reasoning and Computational Representations. International Journal of Computers for Mathematical Learning 9, 3 (Sept. 2004), 251–280.

[179] PINTRICH, P. R., MARX, R. W., AND BOYLE, R. A. Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. Review of Educational Research 63, 2 (1993), 167–199.

[180] PIRINEN, P., LEHTINEN, A., AND HOLMES, N. G. Impact of traditional physics lab instruction on students' critical thinking skills in a Finnish context. European Journal of Physics 44, 3 (Mar. 2023), 035702.

[181] POLLARD, B., FOX, M. F. J., RÍOS, L., AND LEWANDOWSKI, H. J. Creating a coupled multiple response assessment for modeling in lab courses. pp. 400–405. ISSN: 2377-2379.

[182] POLLARD, B., HOBBS, R., DOUNAS-FRAZER, D., AND LEWANDOWSKI, H. J. Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses. In Physics Education Research Conference 2019 (Provo, UT, July 2019), PER Conference, pp. 458–463.

[183] POLLARD, B., HOBBS, R., DOUNAS-FRAZER, D. R., AND LEWANDOWSKI, H. J. Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses. pp. 458–463. ISSN: 2377-2379.

[184] POLLARD, B., HOBBS, R., HENDERSON, R., CABALLERO, M. D., AND LEWANDOWSKI, H. J. Introductory physics lab instructors' perspectives on measurement uncertainty. Physical Review Physics Education Research 17, 1 (May 2021), 010133. Publisher: American Physical Society.

[185] POLLARD, B., HOBBS, R., STANLEY, J. T., DOUNAS-FRAZER, D., AND LEWANDOWSKI, H. J. Impact of an introductory lab course on students' understanding of measurement uncertainty. In Physics Education Research Conference 2017 (Cincinnati, OH, July 2017), PER Conference, pp. 312–315.

[186] POLLARD, B., HOBBS, R., STANLEY, J. T., DOUNAS-FRAZER, D. R., AND LEWANDOWSKI, H. J. Impact of an introductory lab course on students' understanding of measurement uncertainty. pp. 312–315. ISSN: 2377-2379.

[187] POLLARD, B., AND LEWANDOWSKI, H. J. Transforming a large introductory lab course: impacts on views about experimental physics.

[188] POLLARD, B., WERTH, A., HOBBS, R., AND H. J. LEWANDOWSKI. Impact of a course transformation on students' reasoning about measurement uncertainty. Physical Review Physics Education Research 16, 2 (Dec. 2020), 020160. Publisher: American Physical Society.

[189] POLLOCK, S. J. Transferring Transformations: Learning Gains, Student Attitudes, and the Impacts of Multiple Instructors in Large Lecture Courses. In AIP Conference Proceedings (2006), vol. 818, AIP, pp. 141–144.

[190] POLS, C. F. J., DEKKERS, P. J. J. M., AND DE VRIES, M. J. Integrating argumentation in physics inquiry: A design and evaluation study. Phys. Rev. Phys. Educ. Res. 19 (Dec 2023), 020170.

[191] POLS, C. F. J., LEWANDOWSKI, H., LOGMAN, P., AND BRADBURY, F. Differences and similarities in approaches to physics LAB-courses. WCPE: World conference on Physics Education (2021).

[192] POLS, F. One setup for many experiments: enabling versatile student-led investigations. Physics Education 59, 1 (Oct. 2023), 015007.

[193] PRIEMER, B., AND HELLWIG, J. Learning About Measurement Uncertainties in Secondary Education: A Model of the Subject Matter. International Journal of Science and Mathematics Education 16, 1 (Jan. 2018), 45–68.

[194] QUINN, K. N., WIEMAN, C. E., AND HOLMES, N. G. Interview Validation of the Physics Lab Inventory of Critical thinking (PLIC). pp. 324–327.

[195] RAINEY, K. D., VIGNAL, M., AND WILCOX, B. R. Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage. Physical Review Physics Education Research 16, 2 (Aug. 2020), 020113.

[196] RAINEY, K. D., VIGNAL, M., AND WILCOX, B. R. Validation of a coupled, multiple response assessment for upper-division thermal physics. Physical Review Physics Education Research 18, 2 (Sept. 2022), 020116. Publisher: American Physical Society.

[197] RECKASE, M. Multidimensional Item Response Theory. Springer, New York, NY, 2009.

[198] RINDSKOPF, D. Reliability: Measurement. In International Encyclopedia of the Social & Behavioral Sciences, N. J. Smelser and P. B. Baltes, Eds. Pergamon, Oxford, Jan. 2001, pp. 13023–13028.

[199] ROSENGRANT, D., VAN HEUVELEN, A., AND ETKINA, E. Case Study: Students' Use of Multiple Representations in Problem Solving. AIP Conference Proceedings 818 (2006), 49–52.

[200] ROVINELLI, R. J., AND HAMBLETON, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschrift voor Onderwijsresearch 2 (1977), 49–60. Place: Netherlands Publisher: Tijdschrift voor Onderwijsresearch.

[201] RUMMEL, R. J. Applied Factor Analysis. Northwestern University Press, 1988.

[202] RÍOS, L., POLLARD, B., DOUNAS-FRAZER, D. R., AND H. J. LEWANDOWSKI. Using think-aloud interviews to characterize model-based reasoning in electronics for a laboratory course assessment. Physical Review Physics Education Research 15, 1 (June 2019), 010140. Publisher: American Physical Society.

[203] SADAGHIANI, H. R., AND POLLOCK, S. J. Quantum mechanics concept assessment: Development and validation study. Physical Review Special Topics - Physics Education Research 11, 1 (Mar. 2015), 010110. Publisher: American Physical Society.

[204] SALEHI, S., BALLEN, C. J., LAKSOV, K. B., ISMAYILOVA, K., PORONNIK, P., ROSS, P. M., TZIOUMIS, V., AND WIEMAN, C. Global perspectives of the impact of the COVID-19 pandemic on learning science in higher education. PLOS ONE 18, 12 (Dec. 2023), e0294821.

[205] SAWTELLE, V., BREWE, E., KRAMER, L. H., SINGH, C., SABELLA, M., AND REBELLO, S. Positive Impacts of Modeling Instruction on Self-Efficacy. pp. 289–292.

[206] SCHANG, A., DEW, M., STUMP, E. M., HOLMES, N. G., AND PASSANTE, G. New perspectives on student reasoning about measurement uncertainty: More or better data. Physical Review Physics Education Research 19, 2 (July 2023), 020105.

[207] SCHERR, R. E., AND HOLMES, N. G. Quantifying Uncertainty and Distinguishing Data Sets in Introductory Physics.

[208] SCHWAB, J. J. The Teaching of Science as Inquiry. Bulletin of the Atomic Scientists 14, 9 (Nov. 1958), 374–379.

[209] SCOTT, T. F., SCHUMAYER, D., AND GRAY, A. R. Exploratory factor analysis of a Force Concept Inventory data set. Physical Review Special Topics - Physics Education Research 8, 2 (July 2012), 020105.

[210] SEYMOUR, E., HUNTER, A.-B., LAURSEN, S. L., AND DEANTONI, T. Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. Science Education 88, 4 (2004), 493–534.

[211] SHULMAN, L., AND TAMIR, P. Research on teaching in the natural sciences. In Second handbook of research on teaching, R. Travers, Ed. Rand McNally, Chicago, 1973.

[212] SMITH, E. M., STEIN, M. M., AND HOLMES, N. G. How expectations of confirmation influence students' experimentation decisions in introductory labs. Phys. Rev. Phys. Educ. Res. 16 (Mar 2020), 010113.

[213] SMITH, E. M., STEIN, M. M., WALSH, C., AND HOLMES, N. G. Direct measurement of the impact of teaching experimentation in physics labs. Phys. Rev. X 10 (Feb 2020), 011029.

[214] SMITH, T. I., EATON, P., WHITE BRAHMIA, S., OLSHO, A., ZIMMERMAN, C., AND BOUDREAUX, A. Analyzing Multiple-Choice-Multiple-Response Items Using Item Response Theory. pp. 432–437. ISSN: 2377-2379.

[215] SPEARMAN, C. "General Intelligence," Objectively Determined and Measured. The American Journal of Psychology 15, 2 (1904), 201–292.

[216] STEIN, M. M., SMITH, E. M., AND HOLMES, N. G. Confirming what we know: Understanding questionable research practices in intro physics labs.

[217] STEIN, M. M., WHITE, C., PASSANTE, G., AND HOLMES, N. G. Student interpretations of uncertainty in classical and quantum mechanics experiments. pp. 573–578.

[218] STEWART, J., MILLER, M., AUDO, C., AND STEWART, G. Using cluster analysis to identify patterns in students' responses to contextually different conceptual problems. Physical Review Special Topics - Physics Education Research 8, 2 (Oct. 2012), 020112.

[219] STEWART, J., ZABRISKIE, C., DEVORE, S., AND STEWART, G. Multidimensional item response theory and the Force Concept Inventory. Physical Review Physics Education Research 14, 1 (June 2018), 010137. Publisher: American Physical Society.

[220] STRIKE, K., AND POSNER, G. In Cognitive structure and conceptual change, L. West and A. Pines, Eds. Academic Press, 1985, p. 211.

[221] STUMP, E. M., DEW, M., PASSANTE, G., AND HOLMES, N. Context affects student thinking about sources of uncertainty in classical and quantum mechanics. Physical Review Physics Education Research 19, 2 (Nov. 2023), 020157.

[222] STUMP, E. M., HUGHES, M., PASSANTE, G., AND HOLMES, N. Comparing introductory and beyond-introductory students' reasoning about uncertainty. Physical Review Physics Education Research 19, 2 (Oct. 2023), 020147.

[223] SUNDSTROM, M., WU, D. G., WALSH, C., HEIM, A. B., AND HOLMES, N. Examining the effects of lab instruction and gender composition on intergroup interaction networks in introductory physics labs. Physical Review Physics Education Research 18, 1 (Jan. 2022), 010102.

[224] SUSAC, A., BUBIC, A., MARTINJAK, P., PLANINIC, M., AND PALMOVIC, M. Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study. Phys. Rev. Phys. Educ. Res. 13 (Oct 2017), 020125.

[225] SÉRÉ, M., JOURNEAUX, R., AND LARCHER, C. Learning the statistical analysis of measurement errors. International Journal of Science Education 15, 4 (July 1993), 427–438.

[226] TABACHNICK, B. G., AND FIDELL, L. S. Using multivariate statistics, 5th ed ed. Pearson/Allyn & Bacon, Boston, 2007.

[227] TEICHMANN, E., LEWANDOWSKI, H., AND ALEMANI, M. Investigating students' views of experimental physics in German laboratory classes. Physical Review Physics Education Research 18, 1 (Apr. 2022), 010135.

[228] THIRY, H., LAURSEN, S. L., AND HUNTER, A.-B. What Experiences Help Students Become Scientists? A Comparative Study of Research and other Sources of Personal and Professional Gains for STEM Undergraduates. The Journal of Higher Education 82, 4 (July 2011), 357–388.

[229] THORNTON, R. K., KUHL, D., CUMMINGS, K., AND MARX, J. Comparing the force and motion conceptual evaluation and the force concept inventory. Physical Review Special Topics - Physics Education Research 5, 1 (Mar. 2009), 010105.

[230] THORNTON, R. K., AND SOKOLOFF, D. R. Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. American Journal of Physics 66, 4 (1998), 338–352. Publisher: American Association of Physics Teachers.

[231] Treagust, D. F. Development and use of diagnostic tests to evaluate students' misconceptions in science. International Journal of Science Education 10, 2 (Apr. 1988), 159–169. Publisher: Routledge _eprint: https://doi.org/10.1080/0950069880100204.

[232] Van Den Akker, J., Kuiper, W., and Hameyer, U. Curriculum Landscapes and Trends. Springer Netherlands, Dordrecht, 2003.

[233] Van Dusen, B., Shultz, M., Nissen, J. M., Wilcox, B. R., Holmes, N. G., Jariwala, M., Close, E. W., Lewandowski, H. J., and Pollock, S. Online administration of research-based assessments. American Journal of Physics 89, 1 (Jan. 2021), 7–8. Publisher: American Association of Physics Teachers.

[234] Vignal, M., Geschwind, G., Henderson, R., Caballero, M. D., and H. J. Lewandowski. Couplet scoring for research based assessment instruments. Discover Education (2024). In review.

[235] Vignal, M., Geschwind, G., Pollard, B., Henderson, R., Caballero, M. D., and Lewandowski, H. J. Survey of physics reasoning on uncertainty concepts in experiments: An assessment of measurement uncertainty for introductory physics labs. Physical Review Physics Education Research 19, 2 (Oct. 2023), 020139.

[236] Vignal, M., Rainey, K. D., Wilcox, B. R., Caballero, M. D., and H. J. Lewandowski. Affordances of Articulating Assessment Objectives in Research-based Assessment Development. In Physics Education Research Conference 2022 (Aug. 2022), pp. 475–480.

[237] Vignal, M., and Wilcox, B. R. Investigating unprompted and prompted diagrams generated by physics majors during problem solving. Physical Review Physics Education Research 18, 1 (Jan. 2022), 010104. Publisher: American Physical Society.

[238] Volkwyn, T. S., Allie, S., Buffler, A., and Lubben, F. Impact of a conventional introductory laboratory course on the understanding of measurement. Physical Review Special Topics - Physics Education Research 4, 1 (May 2008), 010108.

[239] Walsh, C., Lewandowski, H. J., and Holmes, N. G. Skills-focused lab instruction improves critical thinking skills and experimentation views for all students. Physical Review Physics Education Research 18, 1 (Apr. 2022), 010128.

[240] Walsh, C., Quinn, K. N., and Holmes, N. G. Assessment of critical thinking in physics labs: concurrent validity.

[241] Walsh, C., Quinn, K. N., Wieman, C., and Holmes, N. G. Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. Physical Review Physics Education Research 15, 1 (May 2019), 010135.

[242] Wang, J., and Bao, L. Analyzing force concept inventory with item response theory. American Journal of Physics 78, 10 (Oct. 2010), 1064–1070.

[243] Weliweriya, N., Huynh, T., and Sayre, E. Standing fast: Translation among durable representations using evanescent representations in upper-division problem solving. In Physics Education Research Conference 2017 (Cincinnati, OH, July 2017), PER Conference, pp. 432–435.

[244] WELLS, M., HESTENES, D., AND SWACKHAMER, G. A modeling method for high school physics instruction. American Journal of Physics 63, 7 (July 1995), 606–619.

[245] WERTH, A., HOEHN, J. R., OLIVER, K., FOX, M. F., AND LEWANDOWSKI, H. Instructor perspectives on the emergency transition to remote instruction of physics labs. Physical Review Physics Education Research 18, 2 (Nov. 2022), 020129.

[246] WERTH, A., OLIVER, K., WEST, C. G., AND LEWANDOWSKI, H. Assessing student engagement with teamwork in an online, large-enrollment course-based undergraduate research experience in physics. Physical Review Physics Education Research 18, 2 (Oct. 2022), 020128.

[247] WERTH, A., WEST, C. G., SULAIMAN, N., AND LEWANDOWSKI, H. Enhancing students' views of experimental physics through a course-based undergraduate research experience. Physical Review Physics Education Research 19, 2 (Oct. 2023), 020151.

[248] WHITE BRAHMIA, S., OLSHO, A., SMITH, T. I., BOUDREAUX, A., EATON, P., AND ZIMMERMAN, C. Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics. Physical Review Physics Education Research 17, 2 (Oct. 2021), 020129. Publisher: American Physical Society.

[249] WIEMAN, C., AND HOLMES, N. G. Measuring the impact of an instructional laboratory on the learning of introductory physics. American Journal of Physics 83, 11 (Nov. 2015), 972–978.

[250] WILCOX, B. R., AND LEWANDOWSKI, H. J. Correlating students' beliefs about experimental physics with lab course success. pp. 367–370.

[251] WILCOX, B. R., AND LEWANDOWSKI, H. J. Impact of instructional approach on students' epistemologies about experimental physics. pp. 388–391.

[252] WILCOX, B. R., AND LEWANDOWSKI, H. J. Open-ended versus guided laboratory activities:Impact on students' beliefs about experimental physics. Physical Review Physics Education Research 12, 2 (Oct. 2016), 020132.

[253] WILCOX, B. R., AND LEWANDOWSKI, H. J. Research-based assessment of students' beliefs about experimental physics: When is gender a factor? Physical Review Physics Education Research 12, 2 (Sept. 2016), 020130.

[254] WILCOX, B. R., AND LEWANDOWSKI, H. J. Students' epistemologies about experimental physics: Validating the colorado learning attitudes about science survey for experimental physics. Phys. Rev. Phys. Educ. Res. 12 (Mar 2016), 010123.

[255] WILCOX, B. R., AND LEWANDOWSKI, H. J. Developing skills versus reinforcing concepts in physics labs: Insight from a survey of students' beliefs about experimental physics. Physical Review Physics Education Research 13, 1 (Feb. 2017), 010108.

[256] WILCOX, B. R., AND LEWANDOWSKI, H. J. Improvement or selection? A longitudinal analysis of students' views about experimental physics in their lab courses. Physical Review Physics Education Research 13, 2 (Sept. 2017), 023101.

[257] WILCOX, B. R., AND LEWANDOWSKI, H. J. Students' views about the nature of experimental physics. Physical Review Physics Education Research 13, 2 (2017), 020110. Publisher: APS.

[258] WILCOX, B. R., AND LEWANDOWSKI, H. J. A summary of research-based assessment of students' beliefs about the nature of experimental physics. American Journal of Physics 86, 3 (Mar. 2018), 212–219.

[259] WILCOX, B. R., AND POLLOCK, S. J. Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. Physical Review Special Topics-Physics Education Research 10, 2 (2014), 020124. Publisher: APS.

[260] WILCOX, B. R., AND POLLOCK, S. J. Validation and analysis of the coupled multiple response Colorado upper-division electrostatics diagnostic. Physical Review Special Topics - Physics Education Research 11, 2 (Nov. 2015), 020130.

[261] WILCOX, B. R., ZWICKL, B. M., HOBBS, R. D., AIKEN, J. M., WELCH, N. M., AND LEWANDOWSKI, H. J. Alternative model for administration and analysis of research-based assessments. Physical Review Physics Education Research 12, 1 (June 2016), 010139. Publisher: American Physical Society.

[262] WILCOXON, F. Individual Comparisons by Ranking Methods. Biometrics Bulletin 1, 6 (Dec. 1945), 80.

[263] WILSON, J., POLLARD, B., AIKEN, J. M., CABALLERO, M. D., AND H. J. LEWANDOWSKI. Classification of open-ended responses to a research-based assessment using natural language processing. Physical Review Physics Education Research 18, 1 (June 2022), 010141. Publisher: American Physical Society.

[264] WILSON, J. M. The CUPLE physics studio. The Physics Teacher 32, 9 (Dec. 1994), 518–523.

[265] YANG, F. M. Item response theory for measurement validity. Shanghai archives of Psychiatry 26, 3 (2014), 171. Publisher: Shanghai Mental Health Center.

[266] ZHU, G., AND SINGH, C. Surveying students' understanding of quantum mechanics in one spatial dimension. American Journal of Physics 80, 3 (Mar. 2012), 252–259.

[267] ZWICKL, B. M., FINKELSTEIN, N., AND H. J. LEWANDOWSKI. Incorporating learning goals about modeling into an upper-division physics laboratory experiment. American Journal of Physics 82, 9 (Sept. 2014), 876–882. Publisher: American Association of Physics Teachers.

[268] ZWICKL, B. M., FINKELSTEIN, N., AND LEWANDOWSKI, H. J. Development and validation of the Colorado learning attitudes about science survey for experimental physics. AIP Conference Proceedings 1513, 1 (Jan. 2013), 442–445.

[269] ZWICKL, B. M., FINKELSTEIN, N., AND LEWANDOWSKI, H. J. The process of transforming an advanced lab course: Goals, curriculum, and assessments. American Journal of Physics 81, 1 (Jan. 2013), 63–70.

[270] ZWICKL, B. M., HIROKAWA, T., FINKELSTEIN, N., AND H. J. LEWANDOWSKI. Epistemology and expectations survey about experimental physics: Development and initial results. Physical Review Special Topics - Physics Education Research 10, 1 (June 2014), 010120. Publisher: American Physical Society.

Appendix A

ANCOVA Assumptions and Adherence

Hahs-Vaughn and Lomax discuss several assumptions [97] for ANCOVA. We provide evidence of our data meeting the following assumptions:

(1) Independence of observations

(2) Homogeneity of variance

(3) Normality of the residuals

(4) Linear relationship between dependent variable and covariate

(5) Independent variable(s) fixed by researcher

(6) Independence of covariate and the independent variable(s)

(7) Measure of covariate without error

(8) Homogeneity of regression slopes

First, ANCOVA requires independence of observations – that is, that observations are independent of one another both within and across samples. However, no data set collected from students will ever be truly adherent to this requirement in the strictest sense [175]. We are, of course, introducing a sampling bias by testing only students in physics courses. Further, since our data tends to be dominated by large R1 institutions, students are more homogeneous than the general population would predict. However, this will introduce only slight effects in our results,

and is a general issue with any assessment analysis in physics education research. The effects of this are small. We can check this assumption more rigorously both by examining plots of the residuals by group, as well as by using the Durbin-Watson statistic to test for autocorrelation [67–69].

Plots of the residuals are shown in Figure A.1. Because these plots appear random with no correlations and data fairly evenly distributed above and below the zero line, we determine that we have independence of observation. Additionally, we check the Durbin-Watson statistic for autocorrelations, applied to the residuals of the ANCOVA fit. This statistic falls between 0 and 4, with 2 representing uncorrelated data, 0 representing strongly positively correlated data, and 4 representing strongly negatively correlated data [67–69]. Our value of 2.04 is close to 2, thus, we can state that our data follows the independence of observation assumption.

The second assumption of ANCOVA is homogeneity of variance. The variances of each population must be the same (in our case, the variance of post-test scores amongst the different majors and genders). This requirement is also known as homoscedasticity. To test for this, we examine the plot of unstandardized residuals versus the fitted model. A random display of points without patterns suggests that both this assumption and that of normality (discussed below) are met. We find that our data does conform to this, as shown in Figure A.2. If our data did not meet the assumption, we would expect to see a "fanning out" of data points (i.e., clustered along y=0 on the left and broadening out in the y-dimension towards the right) in this plot.

Thirdly, ANCOVA requires the residuals be normally distributed. ANCOVA is relatively robust to violations of this assumption, so only severe deviations from normality are cause for concern. Normality is tested via the Anderson-Darling test [25, 172], as well as determining the skewness and kurtosis (the third and fourth moments of the distribution) of the residuals. The Anderson-Darling test indicates normality to a significance of 1.0%. Additionally, both the skewness (-0.12 ± 0.12) and kurtosis (0.08 ± 0.13) were both near zero, an indication that neither of these effects is dominant in the residuals [98]. We further examine the plot of residuals versus fitted values (Figure A.2) visually as a test of normality. As there are no patterns in this plot (such as a parabolic pattern), normality has been met. Finally, a visual inspection of the Q-Q plot (shown in
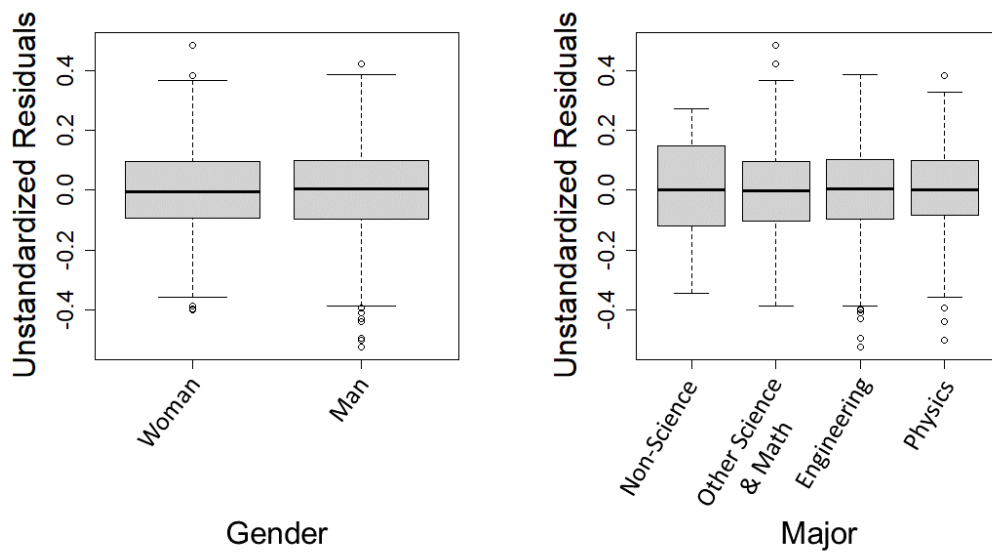
Figure A.1: Plots of the unstandardized residuals for the ANCOVA model. These show that our data conform to the independence of observation assumption as they fall randomly above and below the horizontal line at zero.
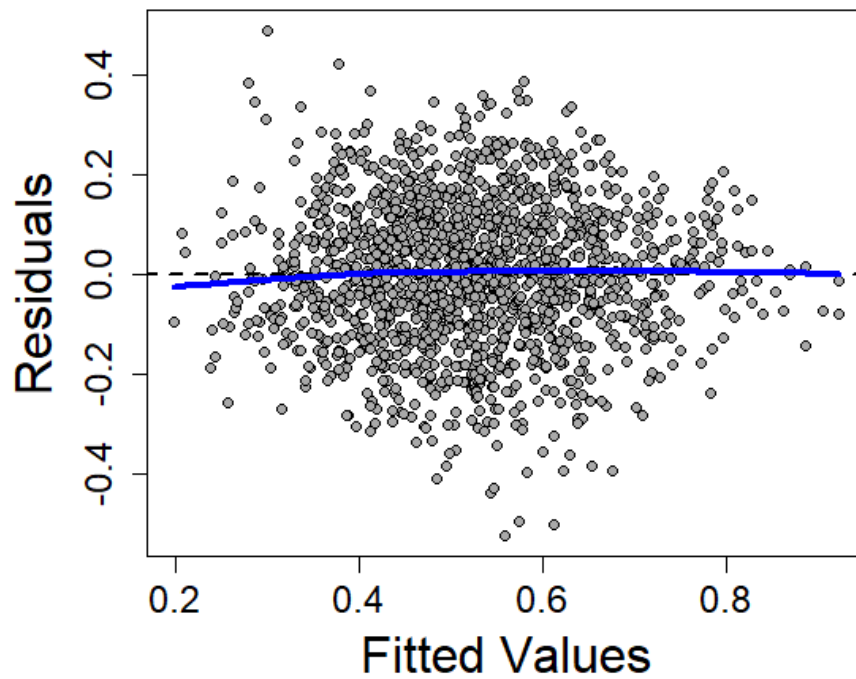
Figure A.2: Plot of the unstandardized residuals versus the fitted model. This shows that our data conform to the homogeneity of variance assumption. Shown as a solid blue line is the averaged data, which indicates no significant discernible patterns and is nearly perfectly aligned with the Residuals = 0 line (black dashed line), indicating normality and homogeneity of variance assumptions have been met.
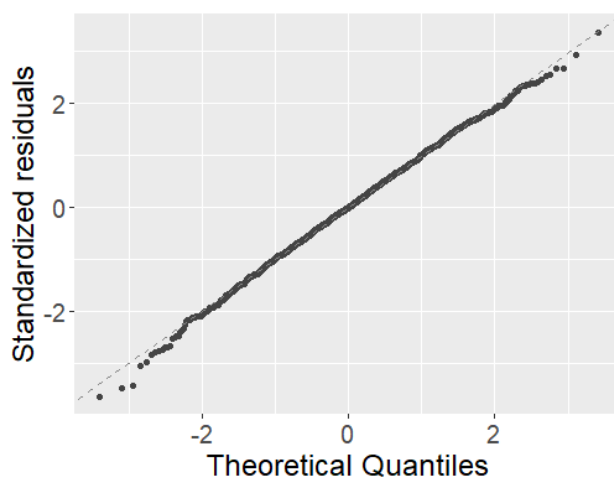
Figure A.3: Q-Q plot. These data indicate normally distributed residuals because the points adhere closely to the diagonal line; too many points deviating from this diagonal would indicate that the residuals are not normally distributed and may point to excess outliers.

Figure A.3) also indicates that the residuals are normally distributed.

The fourth assumption of ANCOVA is overall linearity of data; since ANCOVA is a linear regression, we require that the regression of post-test score on pre-test score is linear. We test this by plotting post-test score versus pre-test score and fitting a line. We don't require a perfect line for this assumption, but rather that our data is linear enough, meaning that it tends towards linear rather curvilinear or uncorrelated. A plot of post-test score vs. pre-test score is shown in Figure A.4. We fit a line to these data and determine $R^2 = 0.402$. Visual analysis of our data shows that it conforms to the assumption of linearity.

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

The sixth assumption is that the covariate and independent variables ideally would be independent from one another. In practice, in assessments, this is not the case; it is extremely common for the covariate and independent variable(s) to share variance. For example, a student's major and their pre-test score are inherently linked. We accept that we must violate this assumption to
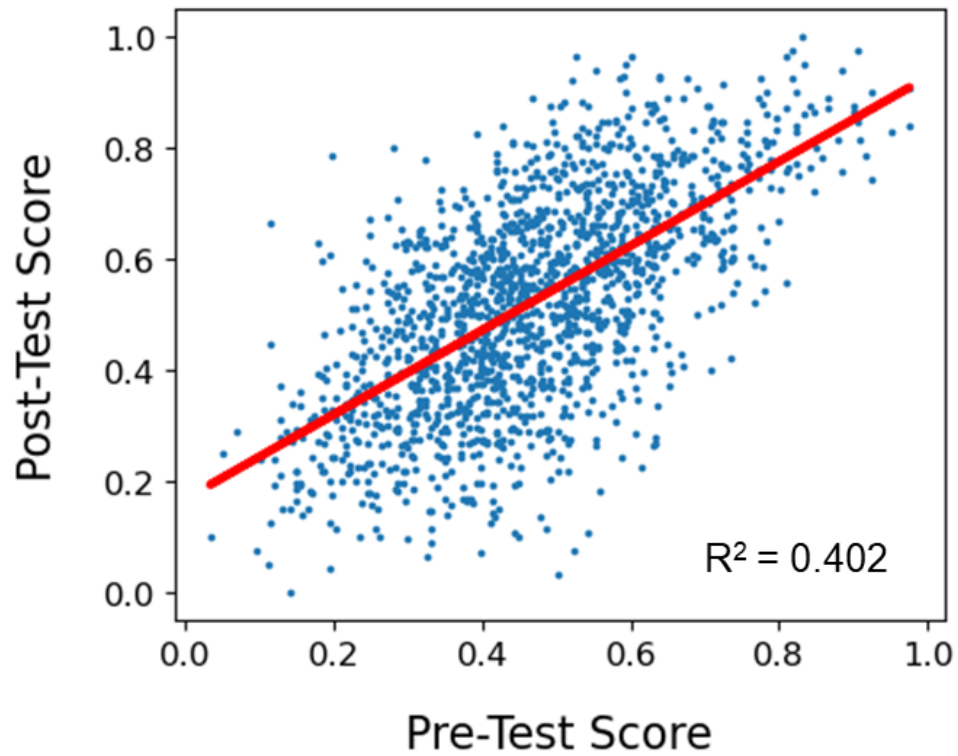
Figure A.4: Plot of Post-test vs. Pre-test overall scores (blue points) with linear fits (red line). This plot shows good evidence of linearity, therefore providing evidence that our data meets this fourth assumption for ANCOVA to be applied.

some extent, with the outcome being that all partial η^2 values that describe the amount of variance explained by each of the predictors must be considered a lower bound instead of considered a particular value [166]. Thus, we might be underestimating some of the variance explained by major or gender when we report values from our ANCOVA analysis.

The next assumption is that the covariate (i.e., pre-test score) must also be measured without error. In terms of assessment, this means that the assessment itself must be proven to have reliability and validity. In Chapter 4, we have shown that SPRUCE has high reliability and validity, and therefore this assumption is met.

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

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

Appendix B

Ordinal Logistic Regression Assumptions and Adherence

The assumptions required to perform for ordinal logistic regression are [100, 175]:

(1) Independence of observations

(2) Noncollinearity of independent variables

(3) Independent variables are linear on the logit

(4) "Perfect" measurement

(5) Non-sparseness of data

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

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

In our case, this means that the pre-test score can not be highly collinear with gender, major, or importance of the AO. We determine our data meets this requirements by performing an OLS

regression and examining the variance inflation factor (VIF) for each AO. VIF values greater than 10 indicate a violation [99]. VIF is defined as:

$$\text{VIF} = \frac{1}{1 - R^2}, \tag{B.1}$$

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

Our results show that our variance inflation factors are all less than 10, meaning that pre-test is not collinear with major, gender, or importance, and therefore, we meet this requirement. The variance inflation factors themselves are presented in Table B.1.

Further, in order to appropriately utilize logistic regression, our independent variables must be linear on the logit, meaning that they must vary linearly with the logit of the dependent variable. However, this assumption is only a requirement for continuous predictors [99], and our model for logistic regression does not have these. Therefore, this assumption is not relevant for our data.

Additionally, logistic regression requires "perfect" measurement. Typically, this means that we measure both our independent and dependent variables without error. In terms of assessment, this means that the assessment itself must be proven to have reliability and validity. In Chapter 4, we have shown that SPRUCE has high reliability and validity, and therefore this assumption is met. Further, since students are self-reporting the demographics used in this analysis, we can assume that we measure this without error as well.

Finally, we aim for the data to not be sparse – that is, we aim to not have any full category cells that are empty (i.e., all categorical data intersections have at least one data point). For example, we hope that our data includes cases of male physics majors at all possible scores on a particular AO for pre- and post-test. In our case, we do have some cells that are not populated. However, in these cases, the effect is that the error on the coefficients is larger, and therefore underestimate the significance of some results; because we are underestimating significance (rather

Table B.1: Variance inflation factors for pre-test with gender, major, or importance. These results show that they are not collinear, as all variance inflation factors are less than 10. We do not have the data for importance of D4, so it is excluded here.

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

than overestimating), and because we can not fix the issue of sparseness without collecting more data, we allow this condition to not be met with the caveat that in our logistic regression models, we might, in some cases, underreport significance.

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

Appendix C

Lab Taxonomy: Lab Title Codebook and Results

In this appendix, we present the codebook and results from a qualitative coding of the lab titles submitted by instructors who participated in the survey. We provide the definitions of all codes used to qualitatively code the lab titles in Table C.1. Codes were developed emergently, where each lab title was first categorized on fine-grained scale and then codes were collapsed to create the final categories. In some cases, titles might be double- or triple-coded as they fall into two or three clear categories. For example, positron emission tomography (PET) is both a particle physics experiment but also one with medical applications and therefore is coded in both categories.

Table C.1: Definitions of codes used to qualitatively code the lab titles as well as the number of courses and number of lab titles coded for each. We include in some cases specifically items which are not included as part of the code.

Code	Definitions	Num. Courses	Num. Lab Titles
Advanced materials and solid state	crystals (including 2D crystals), ferrite hysteresis, fluorophore characterization, magnetic hysteresis, nanoparticles, photovoltaics, plasmon resonance, PN junctions, quantized conduction, quantum dots, quantum Hall effect, semiconductors, solar power, superconducting quantum interference device (SQUID), superconductivity, surface physics, surface roughness via advanced microscopy techniques, thermionic emission, tribology	17	30
Arduino		3	3
Blackbody radiation	blackbody radiation, Planck radiation, thermal radiation	5	5
Charge-to-mass ratio of electron		8	8
Density Measurement	Archimedes' principle, measuring the density of liquids and/or solids	8	9

Dynamics (mechanics)	Atwood machine, collisions, conservation laws (energy, momentum), energy, drag, forces, friction, Maxwell wheel, measuring gravitational constant G (NOT measuring gravitational acceleration g), Newton's laws, orbits, torque, work; NOT pendulum, NOT springs	33	56
Electric fields/electrostatics	2D electric potential [V(x,y)], capacitance, Coulomb's law, current balance, dielectric properties of materials, forces between capacitor plates	16	24
Electron diffraction		6	6
Electronics (advanced)	adders, central processor creation, counters, decoders, digital circuits, digital microchips, drivers for hardware devices, electronic oscillator, flip-flops, Fourier transform, harmonic oscillator circuit implementation, logic gates, multiplexors, Nyquist-Shannon sampling theorem, registers, serial adders, small radio construction, stopwatch with OLED display, triodes	10	30

Electronics (intermediate)	chaotic circuits, coaxial transmission line, diodes, electric engines, electronic feedback and/or control, electronic hysteresis, impedance, IV characteristics (NOT Ohm's law), light-emitting diodes (LEDs), lock-in amplifier, magnetoresistive effects, motors, nonlinear circuits, operational amplifiers (op-amps), passive filters, power factor measurement, RF spectrum analyzer, RLC circuits, Thevenin circuits, toggle circuits, torsion magnetometer, transients, transistors, Wheatstone bridge	29	69
Electronics (simple)	AC circuits, capacitors, DC circuits, inductors, internal resistance, Kirchoff's laws, material resistivity and/or wire resitivity, Ohm's law, resistors, RC circuits, RL circuits, series and parallel circuits, voltage sources	29	56
Fluids	aerodynamics, Bernoulli's equation, Brownian motion, diffusion, fluid flow, Hagen-Poiseuille's law, liquids, rheological behavior, stable Kaye effect, Stokes law, superfluid helium, surface tension	15	21
Franck-Hertz Experiment		6	6
Hall Effect		10	11
Interferometry	Fabry-Perot, Mach-Zehnder, Michelson	12	14

Introduction to measurement and uncertainty	generic "introduction to equipment" or "introduction to measurement", control of variables, statistics lectures, uncertainty analysis and/or error propagation	25	36
Kinematics	center of mass, free fall, gyroscope, inclined plane, measuring gravitational acceleration, g (NOT gravitational constant, G, and NOT the use of a pendulum), moment of inertia, motion analysis, parabolic motion, projectile motion, rotational motion	36	55
Lasers	diode laser, fiber laser, laser pulses, Nd:YAG laser; NOT HeNe lasers, NOT laser spectroscopy	11	12
Magnetic fields	Earth's magnetic field, eddy currents, Faraday's law and/or induction, Helmholtz coils, induced electromotive force, solenoids	15	18
Magnetism	force-distance relationship, Lorentz force, magnetic domains, magnetic force on conductor	11	12
Materials (simple)	anelasticity of solids, bending a bar, deformation, elasticity, elastic torsion, elongation of a wire, plasticity of solids, Young's modulus	10	14
Mechanical oscillations	coupled oscillators, damping, forced mechanical oscillator (with and without friction), harmonic motion, mass/spring, normal modes, resonance	12	16

Medical applications	doppler sonography, electrocardiogram (ECG), eye optics, fluids (blood, sweat, tears), imaging sonography, myography, optical coherence tomography, optical computed tomography (CAT) scan, positron emission tomography (PET), radioactivity & health, ultrasound	14	19
Microscopy	atomic force microscopy (AFM), evanescent light scattering, magnetic force microscopy (MFM), scanning electron microscopy (SEM), scanning probe microscopy (SPM), scanning tunneling microscopy (STM), transmission electron microscopy (TEM)	13	26
Millikan oil drop	determining charge of electron	7	7
Nuclear magnetic resonance	electron spin resonance, Larmor precession, nuclear magnetic resonance	10	14

Optics (advanced)	acousto-optic modulator (AOM), Berry phase (Pancharatnam phase), critical opalescence, dynamic light scattering, Fabry-Perot cavity, fluorescence correlation spectroscopy, Fourier optics, heterodyning, ion traps, laser interference lithography, magneto-optical trap (MOT), magneto-optic effects, optical fibers (NOT fiber lasers), optical pumping, optical trapping and/or tweezers, photocarrier grating, photoluminescence quantum yield, photomultiplier tube (PMT), photon transfer functions, pump-probe (including femtosecond), quantum cryptography, quantum experiments, Raman-Nath diffraction (acousto-optic diffraction, AOD), single photon correlation, single photon detectors, sonoluminescence, spatial light modulator, wavefront shaping, Zeeman effect	17	49
Optics (intermediate)	birefringence, diffraction, greenhouse effect, holography, interferometry and interference (including Young's experiment), microwave diffraction, microwave reflection, microwave scattering, Newton's rings, photometry, physical optics, prisms, refraction (Snell's law), schematic diagrams, spatial filtering, spectroscopy (optical), thin films, wavelengths of visible light	68	108

Optics (simple)	alignment, building a light microscope and/or Kohler's illumination principle, geometric optics, HeNe lasers, lamps, lenses, light sources, mirrors, polarization and/or Brewster's angle, rail optics, ray optics, telescopes and/or galileoscopes	23	39
Particle physics	accelerator physics, alpha rays, angular correlation, beta rays, chain reactions, cloud chamber, coincidence measurements, compton scattering, cosmic ray muons, cross-section, dark matter detection, Fe57 metastable state lifetime, gamma absorption and/or attenuation, gamma spectroscopy, mass of neutron, Mössbauer effect and/or spectroscopy, muon lifetime, nuclear power, particle tracking, positron emission tomography (PET), relativistic electrons, spectroscopy, strangeness, Z0 decays	18	44
Pendulum	chaotic, coupled, Kater's, physical, Pohl, reversion, simple, torsional	23	33
Photoelectric effect		9	9
Plotting	graphical presentation of measurements, graphing motion, graphing with Excel, graphs, plotting, presenting data	5	5
Radioactivity	radioactivity, half-life, attenuation	14	17
Solar cells		6	6

Spectroscopy	atomic spectra, Balmer series, dynamic light scattering, fluorescence correlation spectroscopy, Fourier transform infrared (FTIR) spectroscopy, laser spectroscopy, mass spectroscopy and/or spectrometry, optical grating spectroscopy, Raman spectroscopy, rubidium saturation spectroscopy, saturation spectroscopy, spectroscopy, time-resolved absorption spectroscopy, time-resolved fluorescence spectroscopy; NOT gamma spectroscopy, NOT Mössbauer spectroscopy	28	37
Speed of light		5	5
Speed of sound	measurement of Doppler effect, properties of sound waves, speed of sound in air and in materials	8	8
Springs	Hooke's law, spring constant	8	9
Stern-Gerlach experiment		3	3
Test and measurement equipment	calibration, drift chambers, field programmable gate arrays (FPGAs), image processing, lock-in amplifiers, microcontrollers, micrometers, oscilloscopes, Palmer caliper, periodic signals, reading seismic data, slider caliper, strain gauges, thermocouples/thermometers/thermistors, transducers, Vernier calipers	27	30

Thermodynamics	adiabatic experiments, Boltzmann constant, Boyle's law (Boyle-Mariotte law), calorimetry, critical point, evaporation in a vacuum, heat capacity, heat capacity ratio (Cp/Cv), heat conduction, heat engine, heat pump, heat transfer, heat of combustion, heat of fusion, heat of vaporization (including of liquid nitrogen), ideal gas, linear expansion, Newton's law of cooling, phase transitions, Piston effect, solar cooking box, specific heat, Stirling cycle, temperature dependence of surface tension, thermal expansion, triple points, water vapor	26	56
Viscosity	determination of viscosity of fluid, free fall of sphere in viscous fluid	9	11
Waves	electrical waves, mechanical vibrations, properties of sound waves, resonance of electromagnetic waves, sound frequency measurement, sound resonance in open-end tube, standing electromagnetic waves, standing waves, thermal waves, traveling waves, vibrating strings, vibrations, water waves	17	22
X-ray experiments	X-ray diffraction, X-ray experiments	12	15

Lab Taxonomy Survey, Adapted

Laboratory Taxonomy Survey

Lab Taxonomy Project

Thank you for participating in this survey! We are hoping to collect information from all physics **undergraduate** lab courses around the world. **Please feel free to send the survey link to any other physics lab course instructors you know.**

In this survey, we will ask you about the course(s) you teach. **Fill out this survey for ONLY ONE course at a time - if you teach more than one course, fill out the survey once for EACH course you teach.** There is a text box at the end of the survey for additional comments you feel were not captured by the survey questions.

Characteristics of the Institution and Course

Institution Name (no abbreviations): _____

What country is the institution located in: *Dropdown List of Countries*

What is the highest level of degree that is awarded by your institution?

○ Associate's degree (2-year college, community college)

○ Bachelor's degree

○ Master's degree

○ PhD

Please answer the following questions for ONE course. If you teach more than one course, please take this survey once per course.

Course Name (ex. Introduction to Mechanics): _____

Course Number, if available (ex. PHYS1140): _____

What is the level of your course?

Select **introductory** if your course is part of an introductory physics sequence.

Select **beyond introductory** if your course is **NOT** part of the introductory physics sequence **AND** has at least one prerequisite physics lab course.

◯ Introductory

◯ Beyond Introductory

Approximately how many students are in the entire course in a typical term/semester (all sections)? _____

How many students in total are present in the lab rooms at one time as part of this course (i.e., number of students in a lab section/session)? _____

What is the modality of teaching?

◯ Mostly/only in person

◯ Mostly/only online/remote

◯ Hybrid of in person and online

What topics in physics are included in the lab portion of the course? **Select all that apply.**

☐ Astronomy/Astrophysics

☐ Biophysics

☐ Classical Mechanics

☐ Condensed matter/solid state/materials

☐ Electricity and Magnetism

☐ Electronics (analog and/or digital)

☐ Fluid Mechanics

☐ Geophysics

☐ Modern physics

☐ Nuclear physics

☐ Optics/ laser physics

☐ Particle physics

☐ Plasma physics

☐ Quantum mechanics

☐ Quantum Information

☐ Thermodynamics

☐ Waves

☐ Other: _____

Is the lab course part of a specific physics theory course?

    ○ Yes

    ○ No

Does the lab course include lectures on statistics, data analysis, or experimental techniques required for the lab?

    ○ Yes

    ○ No

Does the lab course meet weekly?

    ○ Yes

    ○ No

If yes to 'does the lab course meet weekly':

    How many weeks does the lab course run for each term/semester? *Dropdown 1 - 52*

    How many hours **per week** are students scheduled to be in the lab room (working on a lab activity)? *Dropdown 1 - 40*

    How many hours **per week** do you estimate students spend in the lab room (working on a lab activity) **beyond the scheduled time**?

        ○ 0 hours   ○ More than 6 hours

        ○ 1-3 hours   ○ Unknown

        ○ 4-6 hours   ○ Not allowed

If no to 'does the lab course meet weekly':

    Please tell us about the frequency of lab course meetings, number of hours students are scheduled to spend in the lab working on experiments during the course, etc.: _____

Do students have a choice from which experiments they complete?

    ○ Yes, for ALL parts of the course

    ○ Yes, for SOME parts of the course

    ○ No

Which of the following describes your lab course? A student does... (**select all that apply**)

    ☐ Multiple experiments per individual lab meeting

    ☐ One experiment per individual lab meeting

    ☐ One experiment for multiple individual lab meetings

    ☐ Multi-session open-ended/partially open-ended project

If selected 'Multi-session open-ended/partially open-ended project' above:

    Thinking specifically about the project component of the course...

        How many weeks are spent on project work? *Dropdown 1 - 52*

        Do students choose their own question to investigate?

            ○ Yes

            ○ Sometimes

            ○ No

        Do students design their own experimental procedure?

            ○ Yes

            ○ Sometimes

            ○ No

        Do students build their own experimental setup/apparatus?

            ○ Yes

            ○ Sometimes

            ○ No

Students in the Course

Students in this course are typically earning a degree in (**select all that apply**):

☐ Physics

☐ Astrophysics/Astronomy

☐ Physics/Astronomy teaching/pedagogy

☐ Engineering

☐ Mathematics

☐ Computer Science

☐ Chemistry

☐ Another science (e.g., biology, geology)

☐ Other teaching/pedagogy

☐ Other degree than above: _____

What fraction of students in the course are earning a degree in physics/astrophysics?

◯ 0-25%

◯ 25-50%

◯ 50-75%

◯ 75-100%

Students in the course are typically in their (**select all that apply - undergraduate only**):

☐ 1st year

☐ 2nd year

☐ 3rd year

☐ 4th year

☐ 5th year or higher

Grouping

Do students typically work on lab activities/experiments...

◯ Alone

◯ With at least one other student

If 'With at least one other student' above:

How many students are typically in each group:

      ○ 2

      ○ 3

      ○ 4

      ○ 5 or more

Which of the following best describes the groups in the course?

      ○ Most students stay in the same group for the entire course

      ○ Most students change groups at least once during the term/semester

Students (**select all that apply**):

      ☐ Choose their own group members

      ☐ Are assigned group members

Instructional Staff Role

Here we are asking about the role of teachers in the lab. These are often referred to using different terms, e.g., instructors, demonstrators, teachers, teaching assistants...

Number of instructional staff present in the lab at one time:

    Faculty/Instructor: *Dropdown 1 - 15+*

    Lab Technicians: *Dropdown 1 - 15+*

    Postdoc/PhD/Master's teaching assistants: *Dropdown 1 - 15+*

    Undergraduate teaching assistants: *Dropdown 1 - 15+*

If 'Postdoc/PhD/Master's teaching assistants' and/or 'Undergraduate teaching assistants' are greater than zero above:

Training provided to student teaching assistants (**select all that apply**):

☐ Pedagogy (teaching practices)

☐ Familiarization with lab equipment (e.g., rehearsing the experiment)

☐ How to grade students' work

☐ Other: _____

Frequency of training provided to student teaching assistants:

☐ Never

☐ Once per academic year

☐ Once per term/semester

☐ Weekly

☐ Other: _____

Goals

Which of the following are goals for your lab course?

	Major Goal	Minor Goal	Not a Goal	Future Goal
Reinforcing physics concepts previously seen in lecture (confirming known results / seeing theory in an experiment)	◯	◯	◯	◯
Learning/discovering physics concepts not previously seen in lecture	◯	◯	◯	◯
Developing technical knowledge and skills (e.g., making measurements and hands-on manipulation of equipment)	◯	◯	◯	◯
Designing experiments	◯	◯	◯	◯
Develop mathematical model(s) of experimental results	◯	◯	◯	◯
Learning how to analyze and interpret data (e.g., linear regressions, uncertainty)	◯	◯	◯	◯
Learning how to visualize data (e.g., plotting)	◯	◯	◯	◯
Developing lab notebook keeping skills	◯	◯	◯	◯
Developing scientific writing skills (e.g., lab reports)	◯	◯	◯	◯
Developing other communication skills (e.g., oral presentations, poster presentations)	◯	◯	◯	◯

Making quick and simple approximations to predict experimental outcomes (e.g., back of the envelope calculations)	◯	◯	◯	◯
Developing expert-like views about the nature of the process of doing experimental physics (e.g., experimentation is iterative, not linear)	◯	◯	◯	◯
Developing collaboration and teamwork skills	◯	◯	◯	◯
Reflecting on and evaluating one's own learning and knowledge (metacognition)	◯	◯	◯	◯
Enjoying experimental physics and/or the course	◯	◯	◯	◯

Activities

Which, if any, of the following **officially branded** approaches to lab instruction do you use in whole or part in your class? **Select all that apply.**

If you don't recognize these, please select "None of the above".

Feel free to click the links to learn more about each type of instruction. *Note: in the actual survey, each of these options is a link to a page with more information about each of these types of instruction*

☐ Course-based undergraduate research experience (CURE)

☐ ISLE Physics

☐ SCALE-UP

☐ Thinking Critically in Physics Labs (Cornell Designed Labs)

☐ Modeling Instruction

☐ Studio Physics

☐ Other: _____

◯ None of the above

Do you often use any of the following research-based assessments to evaluate teh course? **Select all that apply.**

Feel free to click the links to learn more about each assessment. *Note: in the actual survey, each of these options is a link to a page with more information about each assessment.*

☐ Colorado Learning Attitudes about Science Survey, Experimental Physics

☐ Modeling Assessment for Physics Laboratory Experiments

☐ Physics Lab Inventory of Critical thinking

☐ Survey of Physics Reasoning on Uncertainty Concepts in Experiments

☐ Other: _____

◯ None of the above

On average for the course, how often do students engage with the following activities?

Data Analysis and Visualization

	Never	Would like to use in the future	1-2 times per semester	Somewhat frequently	Very frequently
Quantify uncertainty in a measurement	◯	◯	◯	◯	◯
Calculate uncertainty using error propagation	◯	◯	◯	◯	◯
Use software to aid with data analysis and visualization	◯	◯	◯	◯	◯
Fit a mathematical model to data (e.g., linear regression)	◯	◯	◯	◯	◯
Relate curve fitting parameters to physical quantities	◯	◯	◯	◯	◯
Write their own code to analyze data (e.g., in Python, Mathematica, Matlab)	◯	◯	◯	◯	◯
Determine the mathematical model to best represent the data	◯	◯	◯	◯	◯
Determine if two measurements (with uncertainty) are consistent with each other	◯	◯	◯	◯	◯

Communication

	Never	Would like to use in the future	1-2 times per semester	Somewhat frequently	Very frequently
Give oral presentations as an individual	○	○	○	○	○
Give oral presentations as a group	○	○	○	○	○
Write lab reports	○	○	○	○	○
Design and present a poster	○	○	○	○	○
Maintain an individual lab notebook (paper or electronic)	○	○	○	○	○
Maintain a group lab notebook (paper or electronic)	○	○	○	○	○
Do a literature search	○	○	○	○	○
Read scientific papers	○	○	○	○	○
Write a proposal for designing or conducting an experiment	○	○	○	○	○
Complete a worksheet/fill-in-the-blank template	○	○	○	○	○
Provide peer feedback to other students in the course	○	○	○	○	○

Student Decision-Making

	Never	Would like to use in the future	1-2 times per semester	Somewhat frequently	Very frequently
Develop their own research questions	◯	◯	◯	◯	◯
Design their own procedures for data collection	◯	◯	◯	◯	◯
Design their own apparatus	◯	◯	◯	◯	◯
Choose their own analysis methods	◯	◯	◯	◯	◯
Refine experimental apparatus or procedure to reduce uncertainty (statistical and/or systematic)	◯	◯	◯	◯	◯

Materials and Resources

	Never	Would like to use in the future	1-2 times per semester	Somewhat frequently	Very frequently
Build their own apparatus using equipment provided	○	○	○	○	○
Build their own apparatus, using equipment they have at home	○	○	○	○	○
Use pre-constructed apparatus	○	○	○	○	○
Use a step-by-step lab manual to complete experiments	○	○	○	○	○
Use a semi-guided (less guidance than step-by-step) lab manual to complete experiments	○	○	○	○	○
Use a scientific paper to guide lab experiments	○	○	○	○	○
Write code to collect data during experiments (e.g., LabVIEW/Python)	○	○	○	○	○
Use commercial educational equipment (e.g., Vernier, Pasco, Teachspin)	○	○	○	○	○

Modeling and Other Activities

	Never	Would like to use in the future	1-2 times per semester	Somewhat frequently	Very frequently
Use mathematical or conceptual models to make predictions	○	○	○	○	○
Calibrate measurement tools	○	○	○	○	○
Confirm results previously known to students	○	○	○	○	○
Determine results already known to instructors but not known by students	○	○	○	○	○
Determine results unknown to both students and instructors before the labs (e.g., coefficient of friction of unknown surfaces)	○	○	○	○	○
Troubleshoot problems with the setup or apparatus	○	○	○	○	○
Watch a video (e.g., LabVIEW/Python)	○	○	○	○	○
Complete safety training	○	○	○	○	○
Engage with PhET simulations	○	○	○	○	○

Evaluating Student Learning

If 'With at least one other student' is selected in the question about whether students work alone or with others:

Are students graded/evaluated individually?

◯ Yes

◯ Yes, but some part of the grade is the same for all members of a group

◯ No, they are given a grade only for their group

Select all of the following that are marked/graded to determine the final course grade (**select all the apply**):

☐ Before the lab (prelab) calculations, questions, and/or worksheet

☐ Taking a quiz at home before the lab

☐ Writing a measurement and/or analysis plan before the lab

☐ Watching a video before the lab

☐ Quiz/interview in the lab prior to being allowed to work

☐ Attendance and/or participation

☐ Lab notebooks

☐ Worksheets/fill-in-the-blank template completed during the lab

☐ Lab report

☐ Partial lab report (e.g., methods section only, results section only)

☐ Accuracy and/or precision of experimental results

☐ Oral presentation

☐ Poster presentation

☐ Written exam

☐ Practical exam (i.e., hands-on exam)

☐ Peer feedback on other students' work

☐ Observation of students (e.g., use of equipment or teamwork)

☐ Interview/meeting - discussion of lab between student and instructor

☐ Other:

_____

When grading students, do you use a rubric (a set of guidelines about how something is graded)?

        ○ Yes, for all parts that are graded

        ○ Yes, for some parts that are graded

        ○ No

Optional Long Entry

(Optional) Enter the titles of your lab experiments (use any language):

(Optional) Additional comments about your lab course that might not have been captured above: