Comparing introductory and beyond-introductory students' reasoning about uncertainty

Emily M. Stump,[1] Mark Hughes,[2] Gina Passante,[2] and N. G. Holmes[1]
[1]*Laboratory of Atomic and Solid State Physics, Cornell University, 245 East Avenue, Ithaca, New York 14853, USA*
[2]*Department of Physics, California State University Fullerton, 800 N. State College Boulevard, Fullerton, California 92831, USA*

(Received 26 April 2023; accepted 21 August 2023; published 19 October 2023)

[This paper is part of the Focused Collection on Instructional labs: Improving traditions and new directions.] Uncertainty is an important concept in physics laboratory instruction. However, little work has examined how students reason about uncertainty beyond the introductory (intro) level. In this work we aimed to compare intro and beyond-intro students' ideas about uncertainty. We administered a survey to students at 10 different universities with questions probing procedural reasoning about measurement, student-identified sources of uncertainty, and predictive reasoning about data distributions. We found that intro and beyond-intro students answered similarly on questions where intro students already exhibited expert-level reasoning, such as in comparing two data sets with the same mean but different spreads, identifying limitations in an experimental setup, and predicting how a data distribution would change if more data were collected. For other questions, beyond-intro students generally exhibited more expertlike reasoning than intro students, such as when determining whether two sets of data agree, identifying principles of measurement that contribute to spread, and predicting how a data distribution would change if better data were collected. Neither differences in institutions, student majors, lab courses taken, nor research experience were able to fully explain the variability between intro and beyond-intro student responses. These results call for further research to better understand how students' ideas about uncertainty develop beyond the intro level.

DOI: 10.1103/PhysRevPhysEducRes.19.020147

I. INTRODUCTION

Laboratory instruction is a key part of the undergraduate physics curriculum, providing students with the opportunity to develop experimental skills and knowledge not covered in theory-focused courses [1]. One of these experimental skills is the ability to make decisions about interpreting data and drawing conclusions from an experiment [2–5]. Integral to developing this decision-making skill is understanding measurement uncertainty [6–8]. While many studies have probed introductory students' ideas about uncertainty, very little work has probed students' understanding of uncertainty beyond the intro level. This paper aims to bridge this gap by probing both intro and beyond-intro students' reasoning about different aspects of uncertainty.

A. Student understanding of uncertainty at the intro level

Most of the research on student understanding of uncertainty has focused on procedural reasoning: given some data, what (if any) additional data should they collect and/or how should they analyze the data they have. In a study of first-year undergraduate students, Séré *et al.* [9] found that most students struggled to use multiple measurements in interpreting their data. Most students would take only a single measurement in a lab experiment and would take more measurements only if explicitly prompted to do so. If required to take more measurements, students trusted the first measurement more than the subsequent ones and classified each measurement as "good" or "bad," rather than seeing the entire set of measurements as valuable. In a similar study, Coelho and Séré [10] found that high school students also emphasized individual measurements, believing that an experiment has a "true value" associated with it that they should be able to determine with a single perfect measurement. Because of this emphasis on an idealized single measurement, students tended not to consider uncertainty when evaluating whether two measurements were similar [9,10]. Other studies of intro undergraduate students found similar reasoning across science disciplines [11,12]. A significant portion of students believed that it is possible to make a perfect measurement of a true value with sufficient time and money. Many students also focused on the average value in comparing two datasets and did not consider the spread in the data.

These findings were replicated and extended by Allie, Buffler, Campbell, Lubben, and colleagues studying intro-level students at the University of Cape Town in South Africa [13–15]. These researchers developed a survey instrument to probe students' procedural reasoning related to measurement and uncertainty: the Physics Measurement Questionnaire (PMQ). Based on the student responses, they identified two main paradigms of student procedural reasoning: point and set reasoning. Point reasoners tend to emphasize individual measurements in interpreting data. They believe that any single measurement could yield the true value of a parameter and that deviation from the true value results from mistakes in the experiment. These students see taking multiple measurements as beneficial solely as a way to practice so that they can ultimately make a single perfect measurement. In contrast, set reasoners view each individual measurement as an estimate of the quantity of interest. They regard uncertainty as a natural part of experimentation and consequently rely on a *set* of measurements when interpreting their data. They use statistics such as the mean and standard deviation when reporting their data, as opposed to reporting an individual measured value. The set paradigm is seen as the expertlike approach to measurement and uncertainty, and, accordingly, the goal of instruction is argued to be to shift students' reasoning away from the point paradigm and toward the set paradigm [14,15].

These researchers used the PMQ and the point and set paradigms to probe intro-level students' reasoning, both before [15] and after [14] taking a lab course. Lubben *et al.* [15] found that prior to instruction, the students' responses were split approximately equally between point and set reasoning for questions about making more than one measurement, with point reasoners arguing a single data point was sufficient and set reasoners arguing in favor of collecting more data. In comparing two datasets, however, nearly all students relied on comparing the average values (point) rather than taking into account the spread in the data (set). After instruction, the majority of students exhibited set reasoning in the questions about taking more data, but most still struggled to consistently apply set reasoning in comparing datasets [14]. Like the students in Séré *et al.*'s [9] study, these students struggled to understand how uncertainty and spread in data should be used when interpreting experimental results.

Since then, additional researchers have measured the efficacy of various traditional lab courses at teaching set reasoning or have developed intro lab courses with the goal of shifting students away from point reasoning toward set reasoning, with mixed success. Although students have exhibited increases in set reasoning on the PMQ and similar questions from preinstruction to postinstruction, many (or, in some cases, most) students still apply "mixed" reasoning, using set reasoning in some contexts but point reasoning in others, particularly in comparing two sets of data [7,16–20]. These results indicate that a single intro-level lab course is likely insufficient for undergraduate students to master procedural reasoning about uncertainty.

Several studies have also probed students' ideas about sources of uncertainty and how these relate to students' procedural reasoning. For example, previous studies have found that while students are able to identify various sources of uncertainty in an experiment, they may struggle to appropriately quantify those sources [9,21]. Moreover, researchers have expressed concern about students' tendency to attribute uncertainty exclusively to something going wrong in the experiment or "human error" [5,6,22–24]. The concern centers on observations that this conception of uncertainty can lead students to believe that all uncertainty can be eliminated in an experiment, which is aligned with point reasoning. To address this issue, researchers have advocated for teaching uncertainty not as a list of mistakes to be fixed but rather as a fundamental aspect of experimentation that must be quantified and used to interpret data [6,8,12,20,24].

B. Student understanding of uncertainty beyond the intro level

To our knowledge, only a handful of studies have investigated students' ideas about uncertainty beyond the intro level. Hu and Zwickl [22] probed intro-level, upper-level, and Ph.D. students' views about uncertainty. They observed that upper-level students were more likely than intro-level students to identify uncertainty evaluation as important for evaluating whether an experimental result is trustworthy. They also found that Ph.D. students and upper-level students were more likely than intro-level students to view the purpose of uncertainty analysis as quantifying reliability, although only Ph.D. students flagged uncertainty as an inherent aspect of experimentation. Overall, these results suggest that students' views of measurement can change greatly throughout the physics curriculum.

Our previous work [25–29] has focused on upper-level students' ideas about sources of uncertainty and predictive reasoning about uncertainty in different experimental contexts. We found that students were more likely to identify physics principles as sources of uncertainty in quantum-mechanics experiments (e.g., the single-slit experiment) than in classical experiments (e.g., projectile motion) and in experiments with a theoretical expected distribution of outcomes (e.g., Brownian motion) than in experiments with a theoretical single outcome (e.g., projectile motion) [25,26,28]. We also asked students about how a data distribution would change if an experiment were performed by a larger group of students (more data) or if an experiment were performed by experts (better data) [25,27,29]. Most students indicated that more data would result in the same distribution (the correct answer), although a sizable minority indicated that more data would result in a narrower distribution. For better data, students indicated that experts would either measure the same distribution or measure a narrower distribution, with students more likely to answer "same" for quantum-mechanics experiments (e.g., the single-slit experiment) and in experiments with

a theoretical expected distribution of outcomes (e.g., Brownian motion).

C. Research aims

Given that many students leave intro-level lab courses with at least some pointlike ideas about uncertainty [7,16–20] and the dearth of research on student ideas about uncertainty beyond the intro level, our goal was to broadly characterize the reasoning of a diverse sample of intro and beyond-intro students using previously developed measures of student thinking. In particular, we probed three aspects of student reasoning: procedural reasoning [13–15], ideas about sources of uncertainty [28], and predictive reasoning about measuring more or better data [29]. We found that intro students were already expertlike in their reasoning on two of the five questions, but that beyond-intro students were more expertlike than intro students on the other three probes. We tested (and ruled out) several plausible explanations for the observed differences (selection based on major, lab course experience, research experience, and institutional variability). We use prior work to further situate these results and propose future research directions.

II. METHODS

In this section we describe the survey questions we analyzed and the data collection process. We also describe the coding schemes used to interpret open-ended questions and our approach to making quantitative claims.

A. Survey questions

The survey used in this work is adapted from surveys used in previous work probing student reasoning about uncertainty [13,28,29]. The survey questions analyzed here center around a single experimental scenario of a ball rolling down a ramp that was adapted from the PMQ [13]. Although student reasoning about measurement can vary significantly across different experimental contexts or in generalized questions [28–30], this scenario provides a snapshot of intro and beyond-intro students' reasoning in the context of a familiar experiment and allows us to compare our results to previous findings from the PMQ.

Students are first asked two questions from the PMQ: the Same Mean, Different Spread probe (SMDS) and the Different Mean, Same Spread probe (DMSS) (see Figs. 6, 7, and 8 in the Appendix). In the SMDS probe, two groups of students have each measured five data points such that the two groups have the same mean value but different spread in their data. The probe presents three possible viewpoints about which group has achieved better results and survey respondents are asked to identify with which viewpoint they agree and explain why. The DMSS probe is set up similarly: two groups have each measured five data points, but this time the groups have different mean values and similar spreads. The probe presents two possible viewpoints on whether the two groups' results agree and survey respondents are again asked to identify with which viewpoint they agree and why.

The survey then presents a fictitious histogram of measurements collected by 50 students in a lab class (see Ref. [28]). Students are asked to list sources of uncertainty that contribute to the spread in the data (Sources):

> What is causing the shape of the distribution? List as many causes as you can think of.

Finally, students are asked two closed-response questions about how the fictitious histogram would change if either 100 more students (More Data) or experts using the best possible equipment (Better Data) performed the experiment [29] (see Fig. 1).[1]

The questions were always asked in the same order: SMDS, DMSS, Sources, More Data, Better Data. Students could return to previous questions to change their answers as they progressed through the survey.

At the end of the survey, students were also asked a series of questions related to demographics (race and ethnicity, gender, major, etc.) and educational experiences. We asked students to report which types of lab courses they had taken:

> Are you currently taking or have you previously taken any of the following types of college physics lab class? Choose all that apply.
> (a) Introductory mechanics and/or E&M and/or waves/thermo
> (b) Upper division: electronics
> (c) Upper division: optics
> (d) Advanced lab
> (e) Other:_____

We also asked students to report whether they had research experience:

> Are you currently conducting or have you previously conducted research in any of the following areas? Choose all that apply.
> (a) Experimental physics or astrophysics
> (b) Theoretical physics or astrophysics
> (c) Computational physics or astrophysics
> (d) Experimental research in another science
> (e) Theoretical research in another science
> (f) Computational research in another science
> (g) Other:_____
> (h) No research experience

We use these questions to report on how many students had taken at least one lab course beyond the intro level (marking at least one of *upper division: electronics, upper*

[1]Sources, More Data, and Better Data are capitalized throughout the paper to denote the different questions in the survey.

More data question: If 100 more students were to perform the experiment using the same equipment, how would the shape of the distribution change (original distribution in grey; new distribution in blue)? Please explain your reasoning.

Better data question: If experts were to perform the experiment using the best possible equipment, how would the shape of the distribution change? Please explain your reasoning.

Multiple choice options for each question:

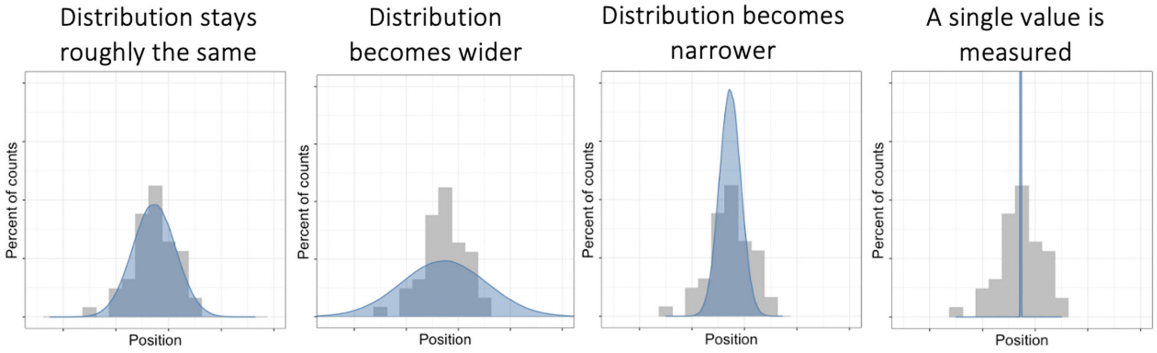


FIG. 1. Text and multiple-choice options for the More Data and Better Data questions [29].

division: optics, advanced lab, or *other* (if applicable)) and how many students had experimental research experience (marking at least one of *experimental physics or astrophysics* or *experimental research in another science*).

B. Data collection

The survey was administered online via Qualtrics during the second half of the Fall 2021 and Spring 2022 semesters. We targeted students who were either enrolled in introductory mechanics or electricity and magnetism courses ("intro" students) or who were enrolled in or had taken at least one quantum mechanics course ("beyond-intro" students). In total we received survey responses from students at 10 universities, with 427 completed responses from intro students and 158 completed responses from beyond-intro students. The universities include private universities, public universities, primarily white institutions, Hispanic-serving institutions, and a historically Black university. Students were recruited by their course instructors and were offered either course credit or a drawing for a $25 gift card for completing the survey. The numbers of participants from each university at each level are shown in Table I and the students' self-reported demographic information is shown in Table IV in the Appendix.

We also asked students questions related to their majors, lab courses taken, and research experience. The intro students were primarily non-physics majors, with 59% engineering majors and 14% other STEM majors; only 15% were physics majors. These students were either enrolled in an intro lab course (49%) or had not yet taken a physics lab course in college. Only 9% of intro students had experimental research experience. The beyond-intro students, in contrast, were primarily physics majors (92%). These students had more experience with lab courses and experimental physics compared to the intro students, with 56% having taken at least one lab course beyond the intro level and 41% having experimental research experience.

C. Coding schemes

The SMDS, DMSS, and sources questions were all open-ended. Thus, we used established coding schemes to characterize student responses to these questions.

The coding schemes for the SMDS and DMSS questions were developed based on the coding schemes in the original papers about the PMQ [13–15] and modified slightly to reflect the student responses in our sample. Responses were coded based on their alignment with the point and set paradigms. The point code was given to responses that focused on comparing only the average values for the datasets or that compared the individual data points in the

TABLE I. Number of responses from intro ($N = 427$) and beyond-intro ($N = 158$) students by university.

Institution	Intro	Beyond-intro
Auburn University	99	0
California State Polytechnic University Pomona	3	0
California State University Fullerton	3	0
California State University San Marcos	26	19
Cornell University	119	43
North Carolina A&T State University	89	0
San José State University	0	10
Texas A&M University	88	7
University of Colorado Boulder	0	78
University of Wisconsin Stout	0	1

TABLE II. Examples of responses receiving the point and set codes from the SMDS and DMSS probes.

Code	SMDS responses	DMSS responses
Point	*"As long as the average is the same, then the right procedures were followed"* (intro student)	*"Even though they are close, the averages are not the same."* (intro student)
	"I agree most with B because the point of an experiment is not to get the result that is better, but the one that is closely related to the values that the students are predicting. Even though the distribution of their data are different, they still have the same average so I would say that both are similar." (intro student)	*"They are extremely similar values, just a few different points were varied. The only difference in the ending values came from slight differences in measurements throughout the experiment, which is normal due to human error."* (intro student)
	"By only looking at the data, I don't believe group B has better results than group A or vice versa. I say this because they both have the same average of 435 mm." (beyond-intro student)	*"The two averages have a percentage difference of 0.46%, so they agree."* (beyond-intro student)
	"I don't really understand what any of these groups mean by 'better.' These are just data from an experiment and just because the data has a greater standard deviation (or spread) it does not mean that it is any better or worse than another set of data." (beyond-intro student)	*"Although the average is not the same, they are very close, averages improve over time and do not have to be equal for agreement"* (beyond-intro student)
Set	*"The data for A is less spread because when looking a the table, the highest and lowest values have a smaller difference compared to B's. A's difference is only 20 mm and B's is 50."* (intro student)	*"The distribution of each group's data is similar enough to confidently assume that the data is reflecting the same phenomena."* (intro student)
	"Although the average is the same, the lower range of the first group is superior." (intro student)	*"Both groups should construct confidence intervals. They appear as though they will overlap and therefore agree with each other."* (intro student)
	"I agree with A, because the numbers when measured are closer together, which means their measurements are more precise." (beyond-intro student)	*"The difference in their means is within the range allowable by the variance of their data"* (beyond-intro student)
	"Group A had more consistent data points with a lower deviation. This indicates that the experiment was done more consistently and carefully, resulting in a lower error than group B. Because group A has a lower estimated error than group B, group A can be more confident in their end average than group B, if each group uses only their own data." (beyond-intro student)	*"The standard error of the mean of each group is about the same, about 6 mm. The averages are within one standard deviation of one another, so the results agree."* (beyond-intro student)

two datasets rather than considering the spread in the data. For example, point responses may argue that the spread is irrelevant to the quality of data or use the percent error between the means to decide whether two datasets agree. The set code was given to responses that mentioned the spread in the data as the justification for the selected viewpoint. For example, set responses may argue that the group with the smaller spread in their data had a better result in the SMDS probe or that the two means in the DMSS probe agreed because there was significant overlap in the spreads of the two datasets. Responses that could not be coded as either point or set or that contained elements of both point and set reasoning were coded as unclear as in Refs. [13–15]. Example point and set responses are shown in Table II.

Two of the authors independently coded a random sample of 50 responses from the SMDS probe and 50 responses from the DMSS probe. We quantified interrater reliability using Cohen's kappa, achieving values of 0.92 and 0.8 for the SMDS and DMSS probes, respectively, which indicates almost perfect agreement [31]. The two researchers then split the remaining responses and independently coded them.

Responses to the Sources question were coded using a previously developed coding scheme [28]. This coding scheme classifies student-listed sources of uncertainty as limitations or principles and is based on the Modeling Framework for Experimental Physics [32,33]. The limitations code was applied to sources of uncertainty related to imperfections in the experimental procedure or setup, such as human error in conducting an experiment, environmental factors such as air resistance, or the precision limit of a measurement device.

The principles code encapsulates both statistical principles and theoretical physics principles. The first includes the idea that experimental measurement is fundamentally probabilistic and, therefore, uncertainty must be modeled using statistical principles. The second includes sources of uncertainty that are due to principles of theoretical physics,

TABLE III. Examples of responses receiving the limitations and principles codes.

Code	Examples
Limitations	*"Equipment accuracy/precision (depending on what is used to take measurements)"* (intro student) *"Different masses of balls"* (intro student) *"Slight shifting of the measuring paper before or after marking"* (beyond-intro student) *"Air currents in the room."* (beyond-intro student)
Principles	*"The distribution is roughly symmetric, and the sample size of 50 is relatively large, greater than 30. Central Limit Theorem explains that this distribution for d is approximately normal, explaining the symmetry."* (intro student) *"Error is typically normally distributed"* (intro student) *"in any real system, there are bound to be differences in measurements creating a normal distribution"* (beyond-intro student) *"the results are random which result in a gaussian distribution (what we observe)"* (beyond-intro student)

such as the Heisenberg uncertainty principle, which places limits on the precision of measurement due to the nature of quantum-mechanical systems. In the context of this study, virtually all responses that received the principles code were related to modeling uncertainty using statistical principles. Examples of student-listed sources of uncertainty that were coded as limitations and principles are shown in Table III.

Any sources that were too vague to classify as limitations or principles were coded as unclear. Most students listed at least one source of uncertainty that we were able to code as either limitations or principles (85% of intro students and 92% of beyond-intro students).

The Sources coding scheme was validated in previous work [28] in which two researchers achieved a Cohen's kappa value of 0.85, indicating almost perfect agreement [31]. One of these researchers coded all of the responses in this study.

As with any analysis of open-response surveys, both of our coding schemes are limited in scope. The codes we used are fairly broad and may fail to capture some interesting nuance in students' responses. For example, the point or set coding scheme treats each student's response as aligned either with novicelike (point) or expertlike (set) reasoning. However, there may be a range of sophistication and expertlike thinking within responses classified as point, set, or unclear that our coding scheme fails to capture. Similarly, our limitations code includes a wide variety of sources, from actionable or quantifiable sources, such as varying force applied while dropping a ball or the instrumental precision of a ruler, to the more vague and unproductive human error [5,6,22–24]. Within this work, we were comparing student responses across five survey questions, leading to a large number of comparisons that increases the chances of seeing an effect due to random chance. As a result, we chose not to subdivide these codes further within this paper. We leave it to future work to address these nuances within our codes to further categorize student reasoning about uncertainty.

D. Data analysis

Our goal in this work was to compare the reasoning exhibited by intro and beyond-intro students and to provide possible explanations for any differences observed. For the SMDS and DMSS probes, we evaluated the fraction of intro and beyond-intro students whose response was given a code of point, set, or unclear. For the Sources question, students could list multiple sources of uncertainty, each of which received a code of limitations, principles, or unclear. We therefore compared the fractions of intro and beyond-intro students who listed at least one source that we coded as limitations and, separately, the fraction of students who listed at least one source that we coded as principles. For the More Data and Better Data probes, we evaluated the fraction of intro and beyond-intro students who chose each of the possible predicted distributions (the same, wider, narrower, or single value).

To make quantitative comparisons between groups of students, we relied on graphical representation of the 95% confidence interval for each proportion (estimated using the Wilson score interval [34]). We evaluated the distinguishability of pairs of proportions based on the relative overlap of the 95% confidence intervals. We did not quantify p values due to the large number of comparisons being made and the various concerns about p values in the literature [35–37].

Although there are various ways in which students may productively answer some of our survey questions, others have a clear alignment with expertlike reasoning. For the SMDS and DMSS questions, responses that are given the point code are considered to be novicelike, while responses coded as set are considered to align with expertlike reasoning [13–15]. For the Sources question, we expect that identifying principles sources of uncertainty related to the probabilistic nature of measurement is aligned with an expertlike view of measurement [6,8,12,20,24].[2] Similarly, we consider the answer that a single value is measured for either the More Data or Better Data question to be novicelike, as it aligns with point reasoning and is in opposition to a probabilistic understanding of measurement [13–15]. Finally, the most correct answer to the More Data question is that the distribution remains the same, as additional students employing similar methods should measure the same distribution of

[2]Given that very few students mentioned sources related to theoretical physics principles, we do not interpret these responses.

results as the original students. A common incorrect answer is that the distribution becomes narrower, as students may use a "more data is better" heuristic, assuming that because collecting a large amount of data is important for reducing uncertainty in the estimate of a parameter that the distribution of measurements will also become narrower [29].

III. RESULTS

In this section we first present comparisons between intro and beyond-intro students' responses to our three types of measurement uncertainty probes. We then explore possible explanations for the differences we observed between the two populations.

A. Comparing intro and beyond-intro students' reasoning

We probed students' procedural reasoning about measurement, ideas about sources of uncertainty, and predictive reasoning about data distributions. In this section, we report on similarities and differences in reasoning between these two groups of students.

1. Procedural reasoning

We first asked students two questions from the PMQ [13]: the SMDS probe and the DMSS probe. The SMDS probe asks respondents to evaluate which of two data distributions with the same mean but different spreads is the better result, while the DMSS probe asks respondents to decide whether two data distributions with different means but the same spread agree. Student responses to these probes were coded as exhibiting either *point* or *set* reasoning; here we report on the fraction of intro and beyond-intro students' responses that received each code (see Fig. 2).

For the SMDS probe, intro and beyond-intro students were indistinguishable in the rates at which each group applied point and set reasoning. Both intro and beyond-intro students

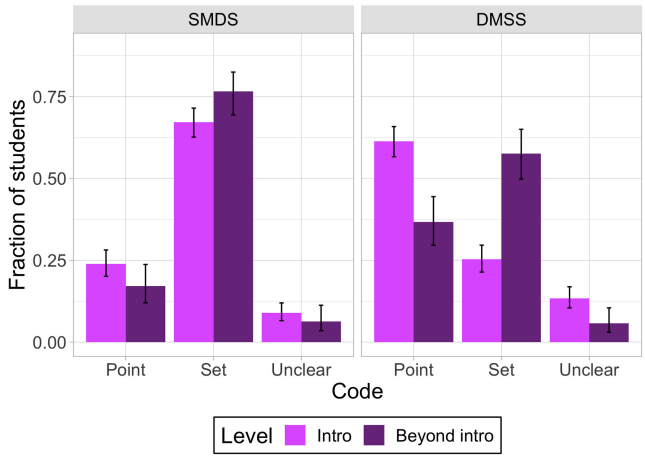


FIG. 2. Point and set codes applied to intro and beyond-intro students' responses to the SMDS and DMSS probes from the PMQ [13]. Uncertainty bars represent the 95% confidence interval.

primarily displayed set reasoning in their explanations (67% and 77%, respectively). These explanations tended to argue that the data distribution with a narrower spread was the better result. For example, a beyond-intro student wrote, *"The standard deviation and error bars will be smaller for group A."* Similarly, an intro student argued, *"Although the average is the same, the lower range of the first group is superior."*

Fewer students applied point reasoning in their responses (24% of intro students and 17% of beyond-intro students). Some of these students concluded that both sets of data were equally good because their means were identical, for example *"Averages are how we determine the accuracy of things in physics. The standard deviation/ uncertainty of group B may be larger than for group A, but that doesnt make their result any 'worse' than group A"* (beyond-intro student). Another line of point reasoning argued that because uncertainty is always a part of measurement, reducing uncertainty is not important: *"Variety is common in physics experiments. It is not always exact. Therefore, if the experiment is done correctly, there should not be any discussion about which is better because variety is common"* (intro student).

For the DMSS probe, however, we observed distinguishable differences in the rates of point and set reasoning for intro and beyond-intro students. Intro students were more likely to use point reasoning in their explanations (61%) compared to beyond-intro students (37%). The students who gave point responses applied varying approaches to comparing the mean values of the two distributions but did not discuss the spread in the data. For example, some of these students argued a difference in means was large or small with no clear justification for the judgement: *"I think that they agree because of how close their averages are"* (beyond-intro student). Other students used the percent difference to make a comparison, for example *"Although their averages are not the same, they are fairly close, and the difference is only a very small percent"* (intro student).

Correspondingly, beyond-intro students were more likely to apply set reasoning (58%) than intro students (25%). Set reasoning responses tended to use measures of the variability in the data, such as the standard deviation or the spread, to determine whether the two datasets agreed. For example one beyond-intro student wrote, *"Both groups do not have precise data, and it would reason that the confidence interval of both groups would have an overlap of the other group's data values."* Similarly, an intro student used the standard deviation to conclude that the two data distributions agreed: *"The average values are within a standard deviation of each other."*

2. Sources of uncertainty

After the PMQ probes, we then asked students the open-ended Sources question: "What is causing the shape of the distribution? List as many causes as you can think of." We coded student-listed sources of uncertainty as either *limitations* in the experimental apparatus or procedures or as *principles* of the measurement process and here report on

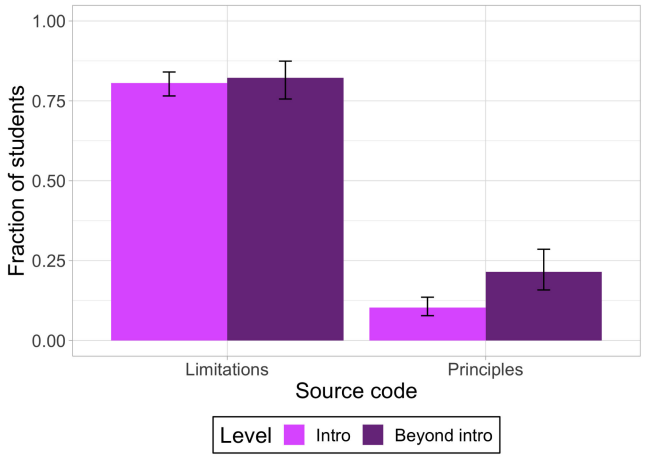


FIG. 3. The fraction of intro and beyond-intro students who listed at least one source of uncertainty coded as limitations and principles. Uncertainty bars represent the 95% confidence interval.

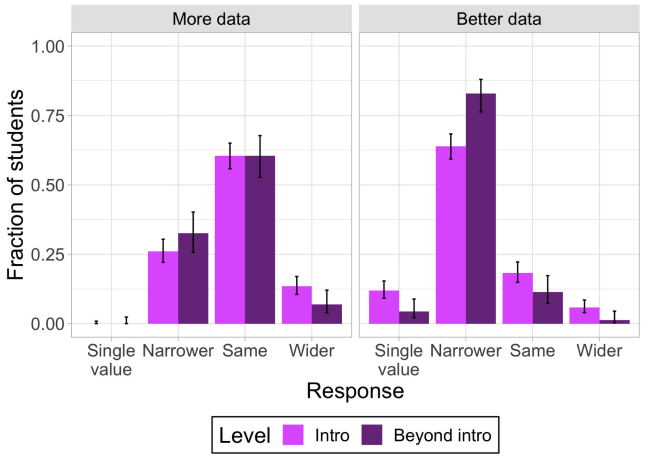


FIG. 4. Distribution of intro and beyond-intro students' responses to the More Data and Better Data questions. Uncertainty bars represent the 95% confidence interval.

the fraction of intro and beyond-intro students who listed at least one source that received each code (see Fig. 3).

The majority of both intro and beyond-intro students listed a source coded as limitations (81% and 82%, respectively). These limitations sources included a variety of experimental factors. Some sources were attributed to errors made by the students, such as *"Due to mistakes, there will be some outliers"* (intro student) and *"Error in reading (i.e., the measuring stick was off center, the ball bounced and they are reading the second impact, etc.)"* (beyond-intro student). Other sources highlighted aspects of the setup that would be more difficult for the students to control, such as *"Air currents in the room"* (beyond-intro student), *"Difference of friction between the ball and the ramp caused by blemishes on the ball"* (intro student), and *"Instrumental error"* (beyond-intro student). The rates at which intro and beyond-intro students identified limitations sources of uncertainty were indistinguishable.

Both intro and beyond-intro students mentioned sources that received the principles code much less frequently than the limitations sources. Furthermore, more beyond-intro students (22%) than intro students (10%) listed at least one principles source of uncertainty. These students tended to mention principles sources related to the inherent statistical nature of measurement, for example *"Principle of normal distributions (data shaped like a bell curve)"* (intro student) and *"Overall general Gaussian distribution is due to expected random error"* (beyond-intro student).

3. Predictive reasoning

The final set of questions in the survey asked respondents to identify what would happen to the data distribution if 100 more students (More Data) or experts (Better Data) were to conduct the experiment. Respondents were given four choices: a single value is measured, distribution becomes narrower, distribution stays roughly the same, and distribution becomes wider. Here we report on the fraction of intro and beyond-intro students who chose each of these four answer options for the two predictive reasoning questions (see Fig. 4).

For the More Data question, the fractions of intro and beyond-intro students choosing each option were indistinguishable. A majority of students indicated that the distribution would remain the same if 100 additional students were to collect data (61% of intro students and 60% of beyond-intro students), which we consider to be the correct answer. The second most common response for both intro and beyond-intro students was that the distribution would become narrower (26% of intro students and 33% of beyond-intro students), which we consider to be an incorrect response.

For the Better Data question, the ordering of answer popularity was the same for intro and beyond-intro students, but the fraction of students who gave each of these responses varied between the two groups. For both groups of students, the distribution becoming narrower was the most common response (64% of intro students and 83% of beyond-intro students), followed by the distribution remaining the same (18% of intro students and 11% of beyond-intro students) and a single value being measured (12% of intro students and 4% of beyond-intro students). However, more beyond-intro students indicated that the distribution would become narrower (83%) compared to intro students (64%), while more intro students indicated that a single value would be measured (12%) compared to beyond-intro students (4%), with the differences beyond the 95% confidence intervals.

4. Summary

Across the five survey questions we analyzed, we observed some instances of similarity between intro and beyond-intro students' answers, while for other questions these two groups answered quite differently. Intro and beyond-intro students primarily applied set reasoning on the SMDS probe, tended to identify limitations sources of uncertainty, and tended to indicate that taking more data

would not change the data distribution width. On the other hand, beyond-intro students were more likely to apply set reasoning on the DMSS probe, more likely to list principles sources of uncertainty, and more likely to answer that better data would result in a narrower distribution (and less likely to answer that a single value measurement would result) compared to intro students.

B. Possible explanations for differences in responses

We consider two types of hypotheses for the differences in survey responses between the intro and beyond-intro students and perform appropriate analyses of our data to test them. One possible hypothesis is that the two groups come from different overall populations, characterized by, for example, their majors or differences in the institutions represented at each level. Another possible hypothesis is that the two groups differ only in their educational experience within the physics curriculum. In our analysis, we test each possible hypothesis individually, although we acknowledge that some of the differences we observed may be explained by combinations of variables rather than individual variables.

1. Population differences

We first consider the possibility that variability in responses could be attributable to differences in the populations of the two groups, as distinct from the additional educational physics experiences that the beyond-intro students have had compared to the intro students.

One population difference relates to the different institutions represented in the samples of intro and beyond-intro students. That is, the data include some institutions that are represented in one group but not the other or that make up different proportions of the sample in each group. Thus, we wanted to confirm that the differences between intro and beyond-intro students were not exclusively explained by institution differences. To do so, we made the same comparisons discussed in Sec. III A within a single university (i.e., holding institution constant), Cornell University, as Cornell was the only individual institution where we had large enough sample sizes at both the intro and beyond-intro levels to draw meaningful conclusions. With this subset of the data, we observed the same trends between the intro and beyond-intro students' responses identified in Sec. III A (see Fig. 9 in the Appendix). These results are discussed further in Appendix C. This finding suggests that the differences we observed in the full dataset between intro and beyond-intro students were not exclusively explained by the differences in the universities represented in each group. We note also that the Cornell University data do not make up the majority of the full dataset at either level, so this analysis is not due to Cornell University driving the trends in the full dataset.

The second possibility is that the responses vary due to the different student majors represented in the samples of intro and beyond-intro students. Physics majors comprise 92% of our beyond-intro sample but only 15% of our intro sample. We compared the responses of intro-level physics majors to intro-level nonphysics majors. For four of the five survey questions (namely, SMDS, Sources, More Data, and Better Data) we observed that intro physics majors and intro nonphysics majors responded similarly (see Fig. 10 in the Appendix and Fig. 5). For the DMSS probe, we observed a small difference in reasoning based on major. More intro nonphysics majors than intro physics majors exhibited point reasoning in their responses (65% and 50%, respectively; see Fig. 5). However, the fractions of nonphysics majors and physics majors who used set reasoning were indistinguishable (23% and 32%, respectively). These results indicate that the differences in majors between the intro and beyond-intro students may partly explain the lower rate of point reasoning in beyond-intro students' responses but cannot explain the higher rate of set reasoning in beyond-intro students' responses.

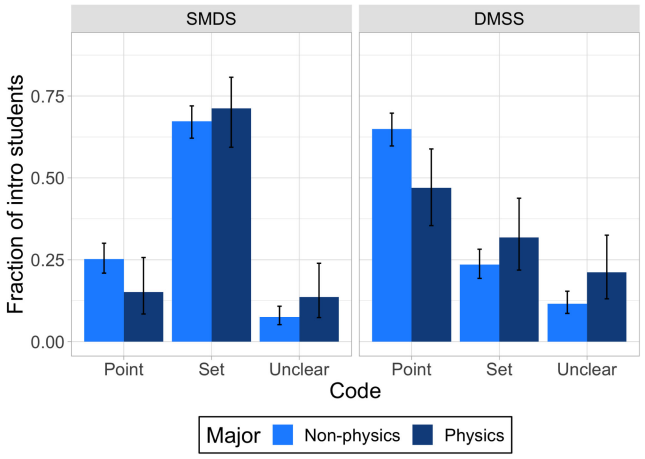


FIG. 5. Codes applied to intro physics majors' and intro nonphysics majors' responses to the SMDS and DMSS probes. Uncertainty bars represent the 95% confidence interval.

2. Educational experiences

The above results indicate that population differences based on institution or major cannot fully explain the observed differences in intro and beyond-intro student reasoning. The differences in reasoning, therefore, are likely also due to differences in physics educational experiences between the intro and beyond-intro students. Here we consider two types of educational experiences that may impact student reasoning about uncertainty: lab courses and research experience.

One of the places we would expect beyond-intro students to learn more about uncertainty is in lab courses taken beyond the intro level. To test this explanation, we compared beyond-intro students' responses to the five survey questions based on whether they had taken only intro lab courses or had taken (or were currently taking) at least one lab course beyond the intro level. We found no differences in student responses to any of the survey questions based on whether

students had taken a beyond-intro lab course (see Fig. 11 in the Appendix).

Another environment in which we might expect beyond-intro students to learn more about uncertainty is conducting research in an experimental laboratory. To test this explanation, we compared beyond-intro students' responses to the five survey questions based on whether they had research experience in an experimental context. We found no differences in student responses to any of the survey questions based on research experience (see Fig. 12 in the Appendix). Overall, we found no evidence that either lab courses taken or research experience could explain the differences in intro and beyond-intro students' responses.

IV. DISCUSSION

In this study, we probed intro and beyond-intro students' ideas about uncertainty using five survey questions related to procedural reasoning, sources of uncertainty, and predictive reasoning. We found that intro and beyond-intro students gave similar answers for the SMDS and More Data questions and in identifying limitations sources of uncertainty but different answers for the DMSS and Better Data questions and in identifying principles sources of uncertainty.

We found that intro and beyond-intro students answered similarly on questions where both groups were mostly using expertlike thinking. Both intro (67%) and beyond-intro (77%) students primarily used set reasoning in their responses to the SMDS probe. This result aligns with previously reported rates of set thinking on the SMDS probe after intro lab instruction.[3] For example, Pillay *et al.* [20] found that 64% of intro students at the University of Cape Town used set reasoning on this probe after taking a lab course designed to help students develop set thinking. Similarly, Wilson *et al.* [38] reported that approximately 70% of responses in a sample of mixed pre and post surveys from the University of Colorado Boulder included set reasoning. For the More Data question, intro and beyond-intro students also answered similarly, with most students giving the correct answer that the distribution would remain the same (61% and 60%, respectively). The high rates of expertlike reasoning among intro-level students may explain why we observed no difference between intro and beyond-intro students' responses: most intro students have mastered the relevant ideas about uncertainty in the context of these questions, so there is limited room for improvement from intro to beyond-intro levels.

Another similarity in intro and beyond-intro students' responses was in identifying limitations sources of uncertainty. The majority of both intro (81%) and beyond-intro (82%) students listed at least one limitation in the procedures or physical setup of the experiment when asked the Sources question. This is unsurprising, as prior work has found that intro students are able to identify a wide variety of sources of uncertainty in a lab setting [5,9,21]. Unlike the more data and SMDS items discussed above, we cannot characterize students listing limitations sources as expertlike or novicelike. Students can list a wide variety of limitations in an experiment, ranging from actionable or quantifiable sources of uncertainty, such as varying force applied while dropping a ball or the instrumental precision of a ruler, to the more vague and unproductive human error [5,6,22–24]. Furthermore, students often struggle to quantify uncertainty associated with limitations in their experiment [9,21], which means listing limitations alone does not demonstrate expertise. Future work should disentangle different productive and unproductive modes of reasoning about limitations in experiments.

For the DMSS probe and Better Data question, in contrast, we observed that beyond-intro students exhibited more expertlike reasoning than intro students. For the DMSS probe, our results indicate low levels of set reasoning among intro students (25%), which aligns with the preinstruction rates of set reasoning reported in prior work [13,20]. Notably, rates of set reasoning were much higher in Pillay *et al.*'s [20] postinstruction survey (75%) and in Wilson *et al.*'s [38] mixed pre and post dataset (approximately 60%) than in our data. This contradiction suggests that the intro lab courses in our dataset may be less effective in teaching set reasoning compared to the lab courses in Pillay *et al.*'s and Wilson *et al.*'s studies. In spite of this apparent shortcoming in intro lab courses, however, the majority of beyond-intro students in our study exhibited set reasoning on the DMSS probe (58%), though still less frequent than in Pillay *et al.*'s study.

For the Better Data question, we observed that beyond-intro students (83%) were more likely than intro students (64%) to indicate that experts would measure a narrower distribution, corresponding to a lower fraction of beyond-intro students (4%) than intro students (12%) who indicated that experts would measure a single value. Believing that experts would measure a single value is aligned with point reasoning, implying that all uncertainty in an experiment can be eliminated and that a single measurement can produce the true value [11–15]. Although this response was fairly uncommon for both intro and beyond-intro students, the lower fraction of beyond-intro students responding "single value" suggests that beyond-intro educational experiences may be effective for eliminating this view of uncertainty from students' understanding of measurement.

One aspect of student reasoning that may help explain the differences in the DMSS probe and Better Data question is students' ideas about sources of uncertainty. Although rare, more beyond-intro students (22%) than intro students (10%) described uncertainty as a principle inherent to the measurement process. Given that previous work has argued that holding this conception of uncertainty can help students apply set reasoning [6,8,12,20,24], this may help

[3]Other work has also looked at student responses to both the SMDS and DMSS probes but either collapsed results for the SMDS and DMSS probes into a single category of data comparison [14,15] or did not report the frequency of point or set codes at the student level [18]. As a result, we cannot compare our results for individual probes directly to these previous findings.

explain why more beyond-intro than intro students exhibited set reasoning on the DMSS probe and why fewer beyond-intro than intro students indicated that experts would measure a single value on the Better Data question. Additional research is necessary to understand how reasoning about sources of uncertainty is connected to point and set reasoning on specific PMQ probes and predictive reasoning questions.

We attempted to determine what educational experiences might explain the differences between intro and beyond-intro students' responses on the survey. We tested whether taking at least one lab course beyond the introductory level or having experimental research experience could explain the differences in student reasoning. We found no evidence, however, that either of these educational experiences alone could explain the differences in intro and beyond-intro students' reasoning. Overall, our results suggest that students' educational experiences beyond the intro level may be enhancing students' reasoning about some aspects of measurement, such as considering uncertainty when comparing two datasets and recognizing that uncertainty is a fundamental aspect of experimental measurement and cannot be eliminated. However, current beyond-intro educational experiences may be less effective for changing other aspects of students' reasoning about uncertainty, such as teaching students that smaller spread in data is desirable or that collecting more data will not change the shape of a data distribution. More research is necessary to identify what specific aspects of beyond-intro students' educational experiences are effective for shifting students' reasoning about uncertainty and evaluate how interventions and course transformations can be used to improve lab instruction related to uncertainty beyond the intro level. In the future, we intend to administer this survey pre and post to students in a variety of physics courses to better understand how specific pedagogical practices impact student reasoning about uncertainty.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139899 and National Science Foundation Grants No. DUE-1808945 and No. DUE-1809178. We thank Courtney White, Matthew Dew, and Andy Schang for their contributions to our previous investigations of student reasoning about uncertainty.

APPENDIX A: SURVEY QUESTIONS

A visual representation of the SMDS and DMSS survey questions is shown in Figs. 6, 7, and 8.

An experiment is being performed by students in the physics laboratory.
A wooden slope is clamped near the edge of a table. A ball is released from a height *h* above the table as shown in the diagram. The ball leaves the slope horizontally and lands on the floor a distance *d* from the edge of the table. Special paper is placed on the floor on which the ball makes a small mark when it lands.

The students have been asked to investigate how the distance *d* on the floor changes when the height *h* is varied. A meter stick is used to measure *d*.

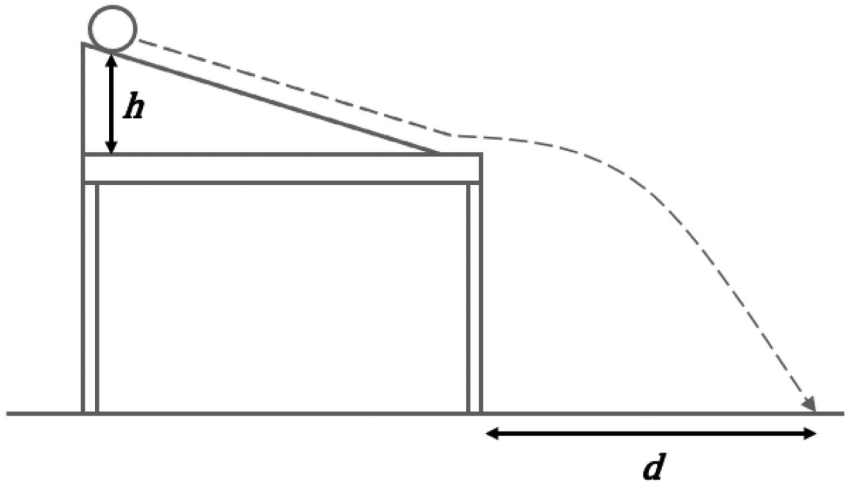


FIG. 6. The overall experimental scenario, modified from [13].

Two groups of students compare their results for *d* obtained by releasing the ball at *h* = 400 mm. Their results for five releases are below.

Release	Group A *d* (mm)	Group B *d* (mm)
1	444	441
2	432	460
3	424	410
4	440	424
5	435	440
Average:	**435**	**435**

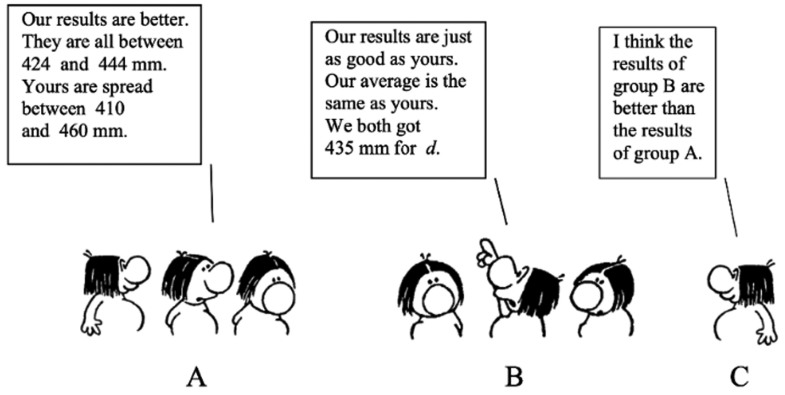


FIG. 7. The SMDS probe, modified from [13].

Two other groups of students compare their results for *d* obtained by releasing the ball at *h* = 400 mm. Their results for five releases are shown below.

Release	Group D *d* (mm)	Group E *d* (mm)
1	440	432
2	438	444
3	433	426
4	422	433
5	432	440
Average:	**433**	**435**

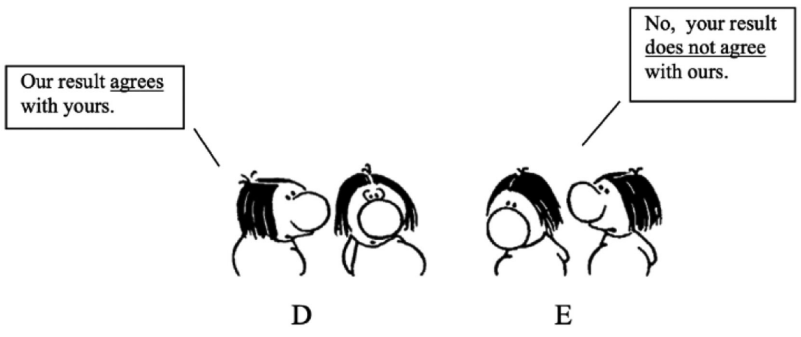


FIG. 8. The DMSS probe, modified from [13].

TABLE IV. Demographic information self-reported by the students included in this study (427 intro students and 158 beyond-intro students). Students who marked two or more races are counted in each race category they chose.

	Intro	Beyond-intro
Year of college		
First year (freshman)	191	1
Second year (sophomore)	156	28
Third year (junior)	42	63
Fourth year + (senior)	23	56
Graduate student	0	2
Unspecified	15	8
Gender		
Female	163	37
Male	241	110
Non-binary	2	4
Unspecified	21	7
Race/ethnicity		
American Indian or Alaska Native	6	3
Asian or Asian American	78	34
Black or African American	83	4
Hispanic or Latinx	58	21
Native Hawaiian or other Pacific Islander	4	2
Prefer to self-describe	3	4
White	214	105
Unspecified	18	9
First-generation status		
First-generation college student	77	26
Not first-generation college student	327	125
Unspecified	23	7

APPENDIX B: DEMOGRAPHIC INFORMATION

The self-reported demographic information for the survey participants is shown in Table IV.

APPENDIX C: ADDITIONAL COMPARISONS ACROSS STUDENT SURVEY RESPONSES

First, we wanted to confirm that the differences between intro and beyond-intro students were not exclusively explained by institution differences. To do so, we compared intro and beyond-intro students' responses within a single university, Cornell. These comparisons are shown in Fig. 9. We observed that most of the trends present in the full dataset were reflected in the Cornell results. Within the Cornell population, we saw no differences in intro and beyond-intro students' responses to the SMDS probe, listing of limitations sources of uncertainty, and responses to the More Data question, as in the full dataset. We saw differences in student's responses to the DMSS probe and listing of principles sources of uncertainty similar to the full dataset. The only discrepancy in conclusions we would

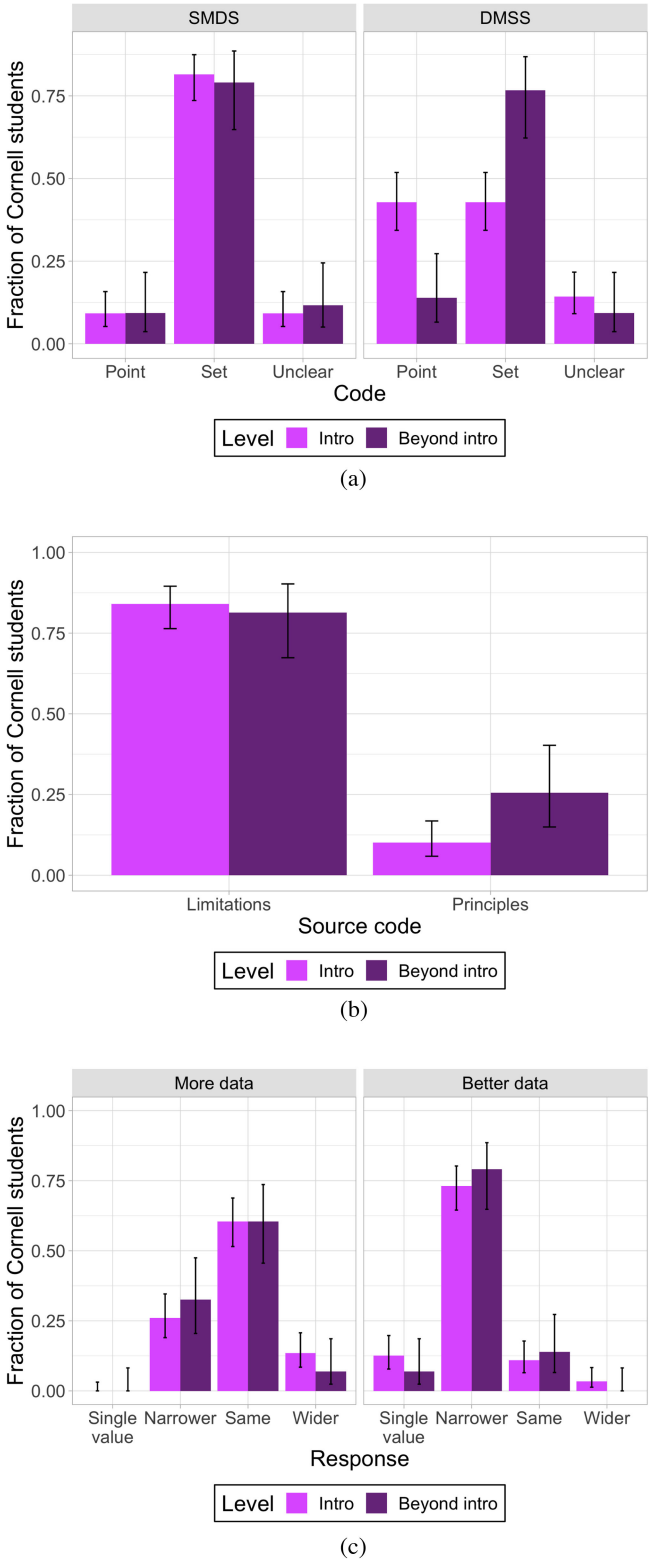
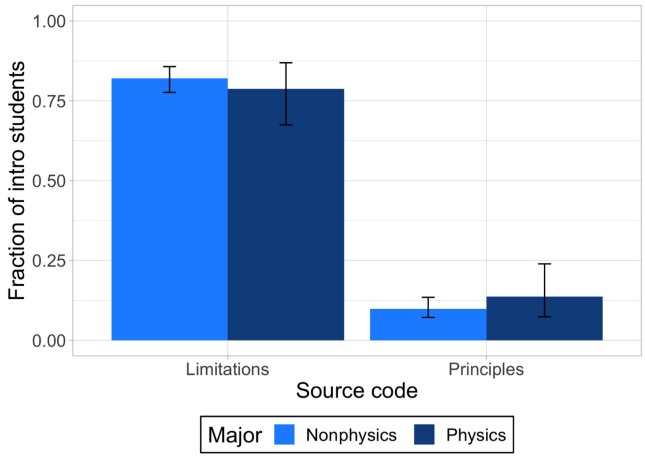
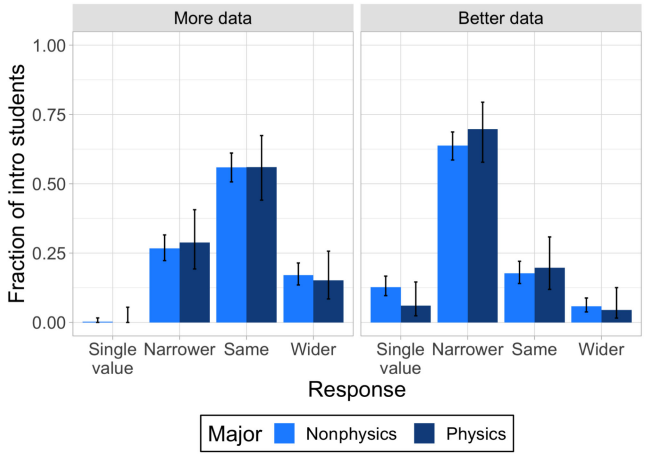


FIG. 9. Comparison of intro and beyond-intro Cornell students' responses to the PMQ probes (a), Sources question (b), and More Data and Better Data questions (c). Uncertainty bars represent the 95% confidence interval.



(a)



(b)

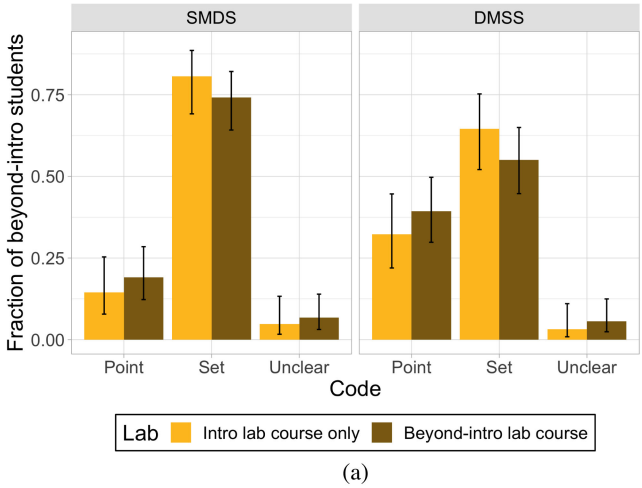
FIG. 10. Comparison of intro physics majors' and intro nonphysics majors' responses to the Sources question (a) and More Data and Better Data questions (b). Uncertainty bars represent the 95% confidence interval.

draw in the Cornell-only dataset compared to the full dataset is for the Better Data question. In the full data, we observed a difference between intro and beyond-intro students' responses, but within the Cornell data these two groups' responses are indistinguishable. However, the trends in the observed fractions for each answer response align with the full-data results, even if the fractions are indistinguishable within uncertainty: a larger fraction of intro students than beyond-intro students indicated that experts would measure a single value (13% and 7%, respectively), while a smaller fraction of intro students than beyond-intro students indicated that experts would measure a narrower distribution (73% and 79%, respectively). Overall, therefore, the Cornell-specific results agree with the full-data results.

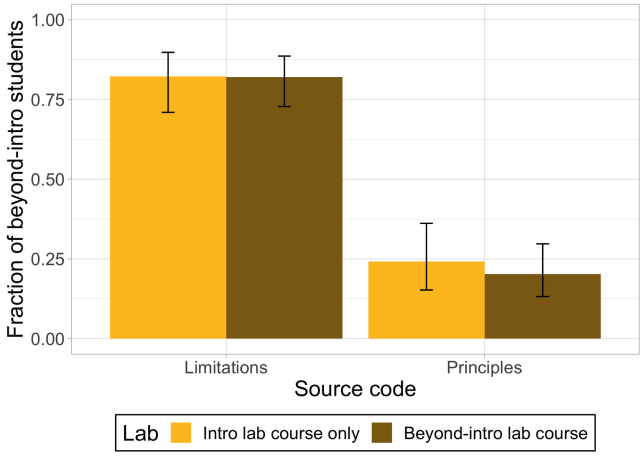
To test whether differences between intro and beyond-intro students' responses were due to differences in student major, we compared intro physics majors' and intro nonphysics majors' responses. These comparisons for the Sources, More Data, and Better Data questions are shown in Fig. 10. For these three questions, we observed no differences in intro students' responses based on major.

To test whether differences between intro and beyond-intro students' responses were due to differences in what lab courses students had taken, we compared beyond-intro students' responses based on whether they had taken only intro-level lab courses or had taken (or were currently taking) at least one beyond-intro lab course. These comparisons are shown in Fig. 11. We observed no differences in beyond-intro students' responses based on lab courses taken.

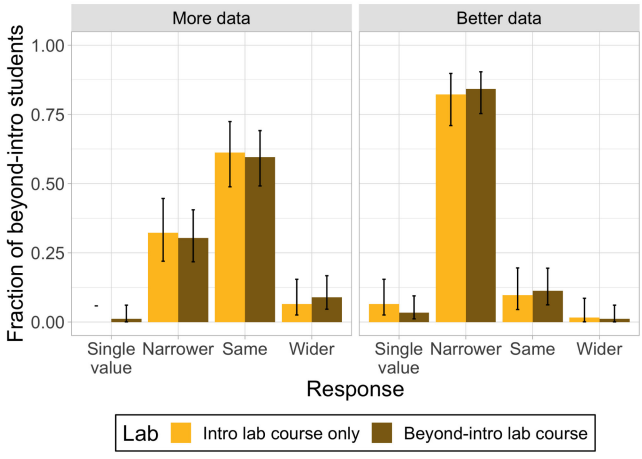
To test whether differences between intro and beyond-intro students' responses were due to differences in students' experience conducting research in an experimental lab setting, we compared beyond-intro students' responses based on whether they had experimental research experience. These comparisons are shown in Fig. 12. We observed no differences in beyond-intro students' responses based on research experience.



(a)

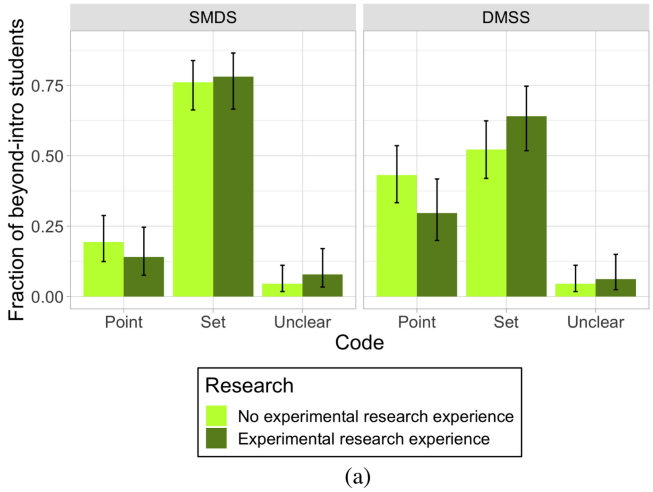


(b)

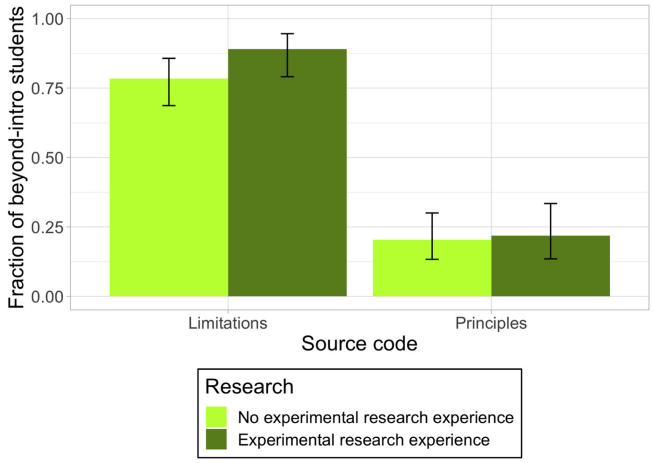


(c)

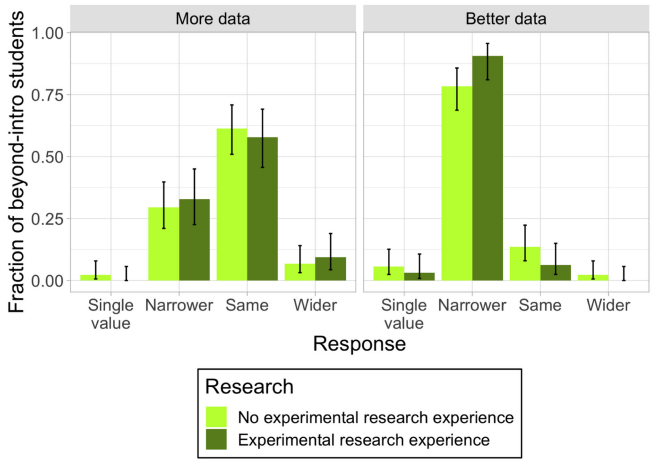
FIG. 11. Comparison of beyond-intro students' responses to the PMQ probes (a), Sources question (b), and More Data and Better Data questions (c) based on what lab courses students had taken. Uncertainty bars represent the 95% confidence interval.



(a)



(b)



(c)

FIG. 12. Comparison of beyond-intro students' responses to the PMQ probes (a), Sources question (b), and More Data and Better Data questions (c) based on research experience. Uncertainty bars represent the 95% confidence interval.

[1] J. Kozminski, N. Beverly, D. Deardorff, R. Dietz, M. Eblen-Zayas, R. Hobbs, H. Lewandowski, S. Lindaas, A. Reagan, R. Tagg, J. Williams, and B. Zwickl, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum, Tech. Rep.* (AAPT, College Park, MD, 2014).

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

[3] N. G. Holmes and E. M. Smith, Operationalizing the AAPT learning goals for the lab, Phys. Teach. **57**, 296 (2019).

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

[5] N. G. Holmes and C. Wieman, Assessing modeling in the lab: Uncertainty and measurement, in *2015 Conference on Laboratory Instruction Beyond the First Year of College* (BFY Conference, College Park, MD, 2015), pp. 44–47, https://dx.doi.org/10.1119/bfy.2015.pr.011.

[6] S. Allie, A. Buffler, B. Campbell, F. Lubben, D. Evangelinos, D. Psillos, and O. Valassiades, Teaching measurement in the introductory physics laboratory, Phys. Teach. **41**, 394 (2003).

[7] R. L. Kung, Teaching the concepts of measurement: An example of a concept-based laboratory course, Am. J. Phys. **73**, 771 (2005).

[8] A. Buffler, S. Allie, and F. Lubben, Teaching measurement and uncertainty the GUM way, Phys. Teach. **46**, 539 (2008).

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

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

[11] J. Leach, R. Millar, J. Ryder, M.-G. Séré, D. Hammelev, H. Niedderer, and V. Tselfes, Survey 2: Students' images of science as they relate to labwork learning, Technical Report, European Commission – Targeted Socio-Economic Research Programme, 1998.

[12] D. Evangelinos, O. Valassiades, and D. Psillos, Undergraduate students' views about the approximate nature of measurement results, in *Proceedings of the 2nd Second International Conference of the European Science Education Research Association* (European Science Education Research Association (ESERA), Kiel, Germany, 1999).

[13] S. Allie, A. Buffler, B. Campbell, and F. Lubben, First-year physics students' perceptions of the quality of experimental measurements, Int. J. Sci. Educ. **20**, 447 (1998).

[14] A. Buffler, S. Allie, and F. Lubben, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).

[15] F. Lubben, B. Campbell, A. Buffler, and S. Allie, Point and set reasoning in practical science measurement by entering university freshmen, Sci. Educ. **85**, 311 (2001).

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

[17] T. Wan, Investigating student reasoning about measurement uncertainty and ability to draw conclusions from measurement data in inquiry-based university physics labs, Int. J. Sci. Educ. **45**, 223 (2022).

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

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

[20] S. Pillay, A. Buffler, F. Lubben, and S. Allie, Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics lab, Eur. J. Phys. **29**, 647 (2008).

[21] E. Etkina, A. Karelina, and M. Ruibal-Villasenor, How long does it take? A study of student acquisition of scientific abilities, Phys. Rev. ST Phys. Educ. Res. **4**, 020108 (2008).

[22] D. Hu and B. M. Zwickl, Examining students' views about validity of experiments: From introductory to Ph.D. students, Phys. Rev. Phys. Educ. Res. **14**, 010121 (2018).

[23] M.-G. Séré, M. Fernandez-Gonzalez, J. A. Gallegos, E. D. Manuel, F. J. Perales, and J. Leach, Images of science linked to labwork: A survey of secondary school and university students, Res. Sci. Educ. **31**, 499 (2001).

[24] D. Evangelinos, D. Psillos, and O. Valassiades, An investigation of teaching and learning about measurement data and their treatment in the introductory physics laboratory, in *Teaching and Learning in the Science Laboratory*, edited by D. Psillos and H. Niedderer (Springer Netherlands, Dordrecht, 2002), pp. 179–190, https://doi.org/10.1007/0-306-48196-0_19.

[25] M. M. Stein, C. White, G. Passante, and N. G. Holmes, Student interpretations of uncertainty in classical and quantum mechanics experiments, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Stein.

[26] E. M. Stump, C. White, G. Passante, and N. G. Holmes, Student reasoning about sources of experimental measurement uncertainty in quantum versus classical mechanics, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.Stump.

[27] C. White, E. M. Stump, N. G. Holmes, and G. Passante, Student evaluation of more or better experimental data in classical and quantum mechanics, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.White.

[28] E. M. Stump, M. Dew, G. Passante, and N. G. Holmes, Context affects student thinking about sources of uncertainty in classical and quantum mechanics, arXiv:2306.14994.

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

[30] J. Leach, R. Millar, J. Ryder, and M.-G. Séré, Epistemological understanding in science learning: The consistency of representations across contexts, Learn. Instr. **10**, 497 (2000).

[31] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33**, 159 (1977).

[32] B. M. Zwickl, D. Hu, N. Finkelstein, and H. J. Lewandowski, Model-based reasoning in the physics laboratory: Framework and initial results, Phys. Rev. ST Phys. Educ. Res. **11**, 020113 (2015).

[33] D. R. Dounas-Frazer and H. J. Lewandowski, The modelling framework for experimental physics: Description, development, and applications, Eur. J. Phys. **39**, 064005 (2018).

[34] E. B. Wilson, Probable inference, the law of succession, and statistical inference, J. Am. Stat. Assoc. **22**, 209 (1927).

[35] J. Cohen, The earth is round ($p < .05$), Am. Psychol. **49**, 997 (1994).

[36] G. Cumming, The new statistics: Why and how, Psychol. Sci. **25**, 7 (2013).

[37] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, The preregistration revolution, Proc. Natl. Acad. Sci. U.S.A. **115**, 2600 (2018).

[38] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, Phys. Rev. Phys. Educ. Res. **18**, 010141 (2022).