Impact of a course transformation on students' reasoning about measurement uncertainty

Benjamin Pollard[1,2,*] Alexandra Werth,[1,2] Robert Hobbs,[3] and H. J. Lewandowski[1,2]

[1]*Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA*
[2]*JILA, National Institute of Standards and Technology, Boulder, Colorado 80309, USA*
[3]*Department of Physics, Bellevue College, Bellevue, Washington 98007, USA*

Physics lab courses are integral parts of an undergraduate physics education, and offer a variety of opportunities for learning. Many of these opportunities center around a common learning goal in introductory physics lab courses: measurement uncertainty. Accordingly, when the stand-alone introductory lab course at the University of Colorado Boulder (CU) was recently transformed, measurement uncertainty was the focus of a learning goal of that transformation. The Physics Measurement Questionnaire (PMQ), a research-based assessment of student understanding around statistical measurement uncertainty, was used to measure the effectiveness of that transformation. Here, we analyze student responses to the PMQ at the beginning and end of the CU course. We also compare such responses from two semesters: one before and one after the transformation. We present evidence that students in both semesters shifted their reasoning in ways aligned with the measurement uncertainty learning goal. Furthermore, we show that more students in the transformed semester shifted in ways aligned with the learning goal, and that those students tended to communicate their reasoning with greater sophistication than students in the original course. These findings provide evidence that even a traditional lab course can support valuable learning, and that transforming such a course to align with well-defined learning goals can result in even more effective learning experiences.

I. INTRODUCTION

Lab courses are an important part of physics undergraduate curricula [1,2]. These courses offer opportunities for learning that is critical to becoming a physicist in many different career paths. For example, lab courses are natural settings for students to acquire experimental skills [3,4], practice scientific communication in a wide range of formats [3,5,6], and develop sophisticated beliefs and epistemologies around the nature of science [7]; these goals are not often a primary focus in lecture or theory-focused courses [3,8]. Recent research has shown that there are lab courses that meet some of these learning goals [9–12], but that there is still room for improvement [13,14]. As such, an increasing number of lab educators are considering the variety of learning goals possible in lab courses, and working to align their courses to better achieve these goals. Hand in hand, education researchers need to better understand the range of learning that occurs in lab courses, and identify teaching strategies that are effective at facilitating such learning.

A. Measurement uncertainty as a learning goal

In this work, we focus on measurement uncertainty as a learning goal of introductory physics lab courses. Uncertainty analysis is a common learning goal in physics lab courses [15], thus the specifics of how it is taught are as varied as physics labs themselves. Here, we highlight a few lab curricula that are discussed in literature that describe a focus on measurement uncertainty, as well as some research studies around learning of measurement uncertainty in labs.

The Scientific Community Laboratory (SCL), developed at the University of Maryland, centers around a series of research questions that aim to teach students how to produce, analyze, and evaluate scientific evidence [16]. The SCL elevates measurement concepts to the same level of importance as physics concepts, recognizing them as critical for those broader skills. In particular, the SCL focuses on sources of variation in data and the generalizability of results based on statistical significance, and also includes uncertainty considerations in experimental design [17].

The Student-Centered Activities For Large Enrollment Undergraduate Programs (SCALE-UP) Project at North Carolina State University includes labs with uncertainty considerations as learning goals [18]. These activities focus

on including uncertainties when reporting results, and using uncertainty when comparing data. A test developed specifically for SCALE-UP further explored student ideas and approaches around measurement uncertainty [19].

The Investigative Science Learning Environment (ISLE), developed at Rutgers University, includes inquiry activities that focus on sources of experimental uncertainty and ways to minimize them in the context of experimental design and iteration [20,21]. ISLE integrates measurement uncertainty ideas, particularly around systematic uncertainties, in design of and reflection on laboratory experiments.

More recently, the Structured Quantitative Inquiry Lab (SQILab) at the University of British Columbia aims to teach measurement uncertainty in the context of critical thinking [22]. SQILab includes explicit instruction around skills and concepts related to distributions, and extends to the formalisms comparing results using statistical tests. Research on SQILab also includes attitudes and beliefs related to measurement uncertainty [23].

Lastly, the physics labs that are part of the introductory calculus-based physics sequences at Cornell University have been transformed with measurement uncertainty as a learning goal. The transformation frames its goals explicitly in the context of the AAPT lab guidelines [3], and identifies both statistical and systematic uncertainty learning outcomes in the context of modeling and experimental design [24]. While the course included conceptual introductions to measurement uncertainty before it was transformed, research has shown that the more integrated and explicit approach in the transformed course resulted in more students viewing uncertainty as important when deciding if a result is trustworthy [25].

B. Measuring learning of measurement uncertainty

In addition to developing physics lab curricula, physics education researchers have also studied student learning and student ideas around measurement uncertainty, often in conjunction with curriculum development [26–30]. Central to those efforts is the development of several research-based assessment tools related to measurement uncertainty [31,32]. The Concise Data Processing Assessment (CDPA) was developed around a decade ago to measure student understanding of both measurement uncertainty and mathematical models of measured data [33]. It has since been used to study pedagogical scaffolding [34] and gender differences in physics labs [35]. Around the same time that the course transformation project at CU was initiated, the Laboratory Data Analysis Instrument (LDAI) was developed to measure data analysis skills within the context of a single lab report [36]. While the LDAI does not focus on measurement uncertainty exclusively, it includes many aspects of measurement uncertainty as they relate to data analysis. More recently, the Physics Lab Inventory of Critical Thinking (PLIC) was developed to measure a range of skills under the umbrella of critical thinking [37]. Measurement uncertainty concepts are represented in the PLIC in the context of this broader range of experimental practice.

For this work, we use the Physics Measurement Questionnaire (PMQ) [38] to study the introductory lab course at the University of Colorado Boulder (CU), as both the course and the PMQ focus on statistical measurement uncertainty concepts at the introductory physics level. We first describe the course in Secs. II A and II B. We then describe the history and philosophical perspective of the PMQ in Sec. II C, and the particular items (or probes) of the PMQ on which this work focuses in Sec. III A.

II. BACKGROUND

A. Transformation of an introductory lab course

In the broader context of improving physics lab education, the introductory lab course at the University of Colorado Boulder was recently transformed and studied. We describe the course and the transformation process here; more details can be found in Refs. [10–12,32,39,40].

The introductory physics lab course at CU is a stand-alone course typically taken by students in their second or third semester of study at CU. For most students, it is the first physics lab course that they take at the college level. The course, both before and after it was transformed, consists of a series of lab activities involving basic concepts from mechanics, electricity and magnetism, and other topics from introductory physics. Students meet weekly in 2 h lab sessions to work through each activity, and occasionally attend additional lecture sessions on background topics. There are only five to six lectures throughout the course, making them a relatively minor focus compared to the lab activities themselves. The course has no midterms nor a final exam.

Beginning in 2016, author H. J. L. began teaching the introductory physics lab course at CU. At the same time, she initiated a project to transform this course. First, professors in the physics department and various departments in the College of Engineering and Applied Science were surveyed, and engaged in group discussions, in order to identify learning goals for the course. These goals included an alignment of students beliefs and epistemologies about experimental physics with that of expert physicists, positive attitudes about the course and about experimental physics more generally, the ability to create quality graphs, and an understanding of measurement uncertainty [12]. Based on these learning goals, H. J. L., B. P., R. H., and others created a new set of lab activities for the course, with corresponding apparatus, analysis software, lab guides, grading rubrics, prelab videos, and lectures. The transformed course was first taught in Fall 2018, and continues to the present. H. J. L. continued to teach the course throughout this process, including all the semesters studied below.

Here, we highlight some salient aspects added in the course transformation. In the transformed course, there are

short prelab videos that students view before each activity, which include embedded questions for students to respond at particular points in each video [11]. Before the transformation, students were instructed only to read the lab guide ahead of time. In the original course, students wrote lab reports after their in-person activities, and were graded on these reports. In the transformed course, instead of lab reports, students keep an electronic lab notebook while they work, which they upload for grading and feedback at the end of each activity. Graduate teaching assistants grade these lab notebooks using rubrics provided to them by the instructor [12].

While the transformed course was designed to meet the identified learning goals, it is still distinct from the ideal course that the designers would have wanted. This mismatch is due mostly to logistical constraints such as those arising from working with 20–30 graduate teaching assistants, and the logistics of scheduling 35–45 separate weekly lab sections in a single instructional space. Thus, the transformed course still operates in many ways as a traditional introductory physics lab course. For the context of this work, we see traditional lab courses as highly guided and prescriptive, focusing on conceptual rather than skills-based learning, and consisting of verification experiments. In particular for our transformed course, the lab activities remained quite prescriptive, guiding students through procedures with significant scaffolding throughout. Nonetheless, most of the activities in the transformed course focused on skills-based learning, and none were verification experiments.

B. Transformation aspects related to measurement uncertainty

In this work, we focus on the course transformation learning goal concerning measurement uncertainty. Both before and after the transformation, the lectures and lab activities concerned measurement uncertainty concepts in addition to other topics. However, the transformed introductory lab course at CU includes several aspects that intentionally and specifically support learning around measurement uncertainty beyond the original course. First, each lab activity in the transformed course involves students measuring a quantity or outcome that they would not know before completing the measurement. These activities are different than *verification labs*, in which students are measuring a value that they learned in lecture or could look up in a textbook. In the course before transformation, five out of six activities were verification labs, in our judgment. In addition to there being no verification labs in the transformed course, many of the lab activities ask students to use measurements they made previously to make predictions about their present experiment. Then, after making a measurement, many of the lab activities ask students to discuss their result with their peers in the classroom, comparing data to decide if their different

results agree with each other. These discussions provide repeated opportunities for students to consider and communicate both the value and the uncertainty of a result, and to discuss these results in the context of their choices involving data collection and procedure.

Beyond the lab activities themselves, four out of the six lectures in the transformed course focus entirely on measurement uncertainty concepts, and a fifth includes additional discussion of measurement uncertainty along with a discussion of professional norms concerning graphs. Lecture topics around measurement uncertainty in the transformed course include the importance of measurement uncertainty, estimating uncertainty from single and from multiple measurements, the concepts of standard deviation and standard deviation of the mean (also called standard error), distributions in general and the normal distribution in particular, sigma values and significance, making comparisons between measurements, and systematic errors in comparison to random uncertainties. The sixth and final lecture in the transformed course concerns Snell's law, and largely does not focus on measurement uncertainty. Overall across these lectures, students first learn about distributions and the act of measurement as sampling from a distribution. The idea of uncertainty in measurement is presented as a measure of such underlying distributions.

In contrast, while measurement uncertainty was included in lectures before the transformation, it was not as much a focus, and was presented with less of a conceptual underpinning, focusing more on the mechanisms of error propagation and the proper structures for reporting results. Lecture topics around measurement uncertainty in the original course also included Poisson statistics, which was absent in the transformed course. The original course lectures also included many of the same topics as the transformed course, including how to compare measurements, estimating uncertainty from single and from multiple measurements, systematic errors in comparison to random uncertainties, the concepts of standard deviation and standard deviation of the mean, distributions in general and the normal distribution in particular, and sigma values and significance. However, these topics tended to be presented in a somewhat mechanistic way, rather than focusing on conceptual underpinnings.

C. The Physics Measurement Questionnaire

The PMQ originated from studies by researchers in York, UK with primary school students age 9–16 [41,42]. This work stemmed from a need to evaluate a new national curriculum that included school laboratory programs [43], and resulted in a model for how students progressed in their ideas about measurement that categorized students' ideas concerning experimental data as a progression through eight levels. [44]. Soon after, researchers in Cape Town, ZA attempted to use the materials from York in their physics classes for first-year university students at the University of

Cape Town. They found, however, that the materials were not suitable in their context, so they created the PMQ by adapting the instruments developed in York [38].

Similarly, as the Cape Town researchers interpreted preliminary responses from their students, they extended and adapted the frameworks from York to develop the point and set paradigms [45,46]. The point and set paradigms characterize two philosophical perspectives pertaining to the statistical uncertainty of measured quantities. They are described in detail in Refs. [38,47].

The point paradigm represents the idea that it would be possible for a single measurement trial to completely represent the true value of a physical quantity or measurand, where deviations from that true value are due to mistakes in the data taking procedure or unaccounted-for effects in the measurement apparatus. The point paradigm would maintain that the overall goal of a good measurement procedure is to prevent, or identify and eliminate, all mistakes and unaccounted-for effects, allowing for a measurement trial that perfectly captures the measurand. Thus, in the point paradigm, the results of individual trials can be considered independently of each other as long as all factors leading to deviation are taken into account.

In contrast, the set paradigm represents the idea that each individual measurement trial reveals some information about the measurand, but that no individual measurement can yield its true value. Thus, multiple trials must be considered as a distribution, with each successive trial revealing more information about the measurand. However, perfect knowledge of the measurand with zero uncertainty is impossible under the set paradigm. The set paradigm stems from a probabilistic approach to measurement uncertainty [46], and is often considered to be more aligned with expertlike reasoning than the point paradigm.

It is also worth noting what is not captured by the point and set paradigms. The paradigms, and by extension the PMQ itself, were designed to characterize reasoning around *statistical* measurement uncertainty. Discussions around systematic errors, that is, any unwanted or unaccounted-for effect that would not "average out" with repeated trials, are outside the scope of the point and set paradigms. While students' responses in the PMQ often involve such reasoning, those elements are irrelevant in the framework of the point and set paradigms. Additionally, skills and concepts concerning the propagation of uncertainty throughout a calculation are also outside the scope of the paradigms and the PMQ. There are also more subtle distinctions to be made when discussing statistical measurement uncertainty, such as the differences between frequentist and Bayesian perspectives, that are more complex than the distinctions captured by the point and set paradigms. Despite its limited scope, the PMQ is still a valuable tool for studying student learning around measurement uncertainty, especially at an introductory level. Likewise, the point and set paradigms

are nonetheless useful constructs for understanding overarching trends in learning regarding measurement uncertainty at the introductory physics level.

D. This work

In this work, we use the PMQ to measure the effectiveness of the introductory lab course at CU at facilitating student learning around statistical measurement uncertainty. Such learning is directly related to one of our course transformation's learning goals. In full, this goal was stated as, "Students should demonstrate a set-like reasoning when evaluating measurements," where "setlike" is a reference to the set paradigm discussed in Sec. II C.

We aim to answer the research question, (Q1) *Did students respond to the PMQ in ways more aligned with the set paradigm after taking the introductory lab course, compared to when they began the course?* In answering (Q1), we consider both the original and transformed course, despite the fact that the original course did not have explicitly stated learning goals, to investigate whether an entirely traditional physics lab course can achieve such a learning outcome.

Furthermore, we use the PMQ to evaluate the effectiveness of the transformation at achieving its learning goal around measurement uncertainty. We aim to answer the research question, (Q2) *Did student responses to the PMQ after the transformation show greater change towards the set paradigm than responses before the transformation?*

In Sec. III, we describe the probes of the PMQ that we use in this work, as well as the students who take the introductory physics lab course and our approach to collecting and analyzing responses from them. Section IV presents results from our analysis of these PMQ responses, comparing responses from the start of the course (pre) and after completing the course (post), and from before the transformation and after it. We finish by discussing these results in the broader context of physics lab courses in Secs. V and VI.

III. METHODS

A. Probes of the PMQ

The entirety of the PMQ is based on an experiment involving rolling a ball down a slope and then measuring the distance it travels in free fall. Each item, or probe, of the PMQ concerns a decision at one step in the measurement process, from taking data to comparing the analyzed results. While responses to each probe can be interpreted using the point and set paradigms, the different contexts for each probe mean that each probe measures reasoning around measurement uncertainty in a distinct way. Each probe asks students to make a choice, usually between two or three multiple-choice options, and then to explain their choice in open-response format.

In this work, we analyze student responses to four particular probes of the PMQ: RD, UR, SMDS, and DMSS [48]. We chose to exclude the other probes of the PMQ from our analysis for a combination of reasons. Some probes (RT, DMOS, and DMSU) only appear on the pre-test or the post-test version of the PMQ, but not on both, so it was not possible to directly compare students' responses to these probes between pre and post tests. Additionally, when a fellow researcher in our group adapted the PMQ from the original pen-and-paper format to an online format using the Qualtrics online survey platform [49], the SLG probe required significant adaption to the online format, so we decided not to code responses to that probe. Finally, when we consulted with one of the creators of the PMQ, they recommended that we omit one of the probes (RDA) from our analysis, as they gained relatively little insight from that probe in their work.

As an example, the RD probe is shown in Fig. 1. The RD probe concerns data collection, and measures reasoning in that context [38]. RD stands for "repeating distance," referring to whether to repeat a trial that measures the distance that a ball travels. The probe presents three stances: to repeat the trial several more times, to move on after performing only a single trial, or to repeat the trial exactly one more time. Respondents are asked to choose with which stance they agree, and to explain their choice in a text box.
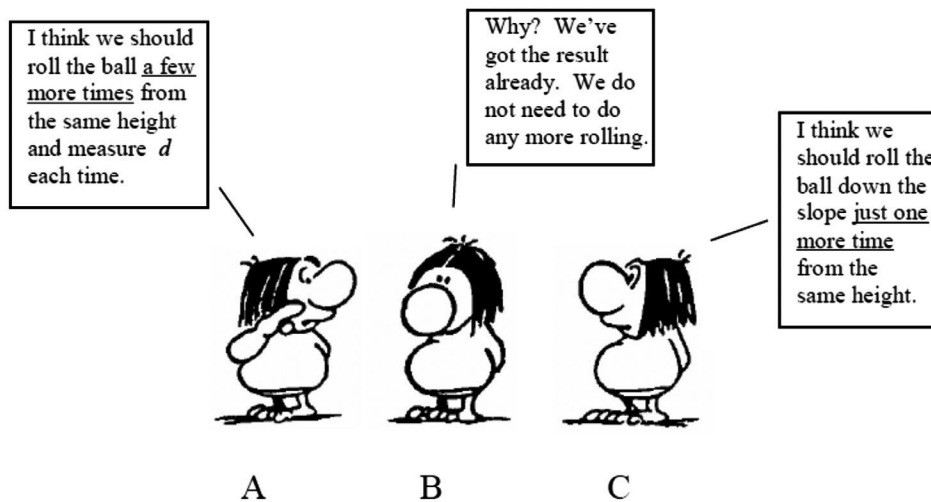
The other probes of the PMQ have the same general format as RD, and are shown in Appendix. The UR probe, which stands for "using repeats," asks how to analyze data to produce a final result. Thus, it probes reasoning in the context of data processing [38] The SMDS probe, which stands for "same mean, different spread," concerns data comparison, asking respondents to decide which of two datasets is better. The two datasets have the same mean, but different spread. Similarly, the DMSS probe, which stands for "different mean, same spread," also concerns data comparison. However, the DMSS probe concerns two datasets that have different means but the same spread. These latter two probes share the context of comparison of results [38], but differ in how the comparison is presented, with SMDS focusing more on means and DMSS focusing more on spreads.

B. Coding scheme development

Because every probe in the PMQ involves an open-response component, responses must first be classified qualitatively before additional analysis is performed. To classify PMQ responses, the creators of the PMQ

The students work in groups on the experiment. Their first task is to determine d when $h = 400$ mm. One group releases the ball down the slope at a height $h = 400$ mm and, using a metre stick, they measure d to be 436 mm.

The following discussion then takes place between the students.

> I think we should roll the ball <u>a few more times</u> from the same height and measure d each time.

> Why? We've got the result already. We do not need to do any more rolling.

> I think we should roll the ball down the slope <u>just one more time</u> from the same height.

A   B   C

With whom do you <u>most closely</u> agree?   (Circle ONE): | A | B | C |

Explain your choice.

FIG. 1.   The RD probe of the PMQ. Reproduced from Ref. [47].

developed a coding scheme based on responses from their students [42]. The coding scheme consists of a different set of codes for each probe, and aims to capture the types of reasoning students draw upon when reasoning about uncertainty in the various stages of measurement in the PMQ. This coding scheme was developed concurrently with the paradigm model described above, and in the current version each code is associated with either the point paradigm, the set paradigm, or an "unknown" designation if the code represents reasoning that does not unambiguously align with one paradigm.

The first set of PMQ responses that were analyzed at CU came from students in the course in Fall 2016, before any of the responses that we present here. That early work encountered a range of student reasoning in the 2016 dataset that was not captured by the coding scheme developed by the creators of the PMQ, and thus B. P., R. H., H. J. L., and others expanded upon it to describe responses from our different national, institutional, and course context. These expanded codes were subsequently consolidated into common themes by B. P., H. J. L., and others, which were then reframed as code definitions to create a new coding scheme for the PMQ. This new coding scheme was then refined based on CU student responses from Fall 2017, with a subset of those responses used to check interrater reliability between two independent coders (R. H. and B. P.). The process of creating and refining the new coding scheme is described in more detail in Ref. [10].

After our coding scheme was developed and verified, we applied it to data from the Spring 2017 and Spring 2018 semesters. We first matched pre- and postresponses by student, and removed the responses that were not matched from the dataset to make direct comparisons of pre- and postdistributions straightforward. Then, for each probe, all of the pre- and post-test responses from those two semesters were anonymized, combined, and shuffled into a single dataset, which RH coded without knowing from which semester and pre- and postdesignation each response came. After the codes were assigned, the data were separated back into their respective categories for further analysis.

C. Coding scheme

The new coding scheme we created consists of 12–16 codes per probe. Each code is denoted by a letter and a number, for example, "U3." When it is necessary to disambiguate between codes pertaining to different probes, we prepend the probe's acronym and a hyphen, for example, "RD-U3."

The letter is either S, P, or U, signifying whether the code falls under the set paradigm, the point paradigm, or unknown reasoning that does not unambiguously align with one paradigm. Across all probes, there are 21 P codes, 22 S codes, and 13 U codes. The codes within each paradigm further differentiate between student reasoning at a finer-grained level. The number designation of each code

distinguishes between them, though we do not intend for the numbers to be interpreted as an ordering. The relative merits of each code are not inherent to the reasoning they represent, and in practice will depend on the context of how the results of analysis are interpreted. For example, when we use the coding scheme here to measure the success of a course, we compare codes in terms of the extent to which each code's reasoning aligns with the learning goals of the course.

Sometimes, student responses contain multiple distinct lines of reasoning, and thus we allow multiple codes to be assigned to a single response. In the datasets analyzed below, 87.9% of the responses were assigned a single code, 11.9% were assigned two codes, and 0.3% were assigned three codes. For the purposes of classifying a response into a single paradigm, if a response was assigned multiple codes from different paradigms, S and P codes both took precedence over U codes. For example, if a response was assigned a P code and a U code, the response was considered pointlike overall. If both an S code and a P code were assigned to a single response, which happened in 1.7% of the responses in the dataset analyzed below, we classify that response's paradigm as U.

The complete code books of the new coding scheme, one for each of the four probes we analyzed, are reproduced in full in the Supplemental Material [50] accompanying this work. Here, we present a subset of these codes in Table I. As an illustration of the scope and depth of the coding scheme, we discuss here three codes for the RD probe: S4, P2, and U1. Each code represents a reason to perform more than one trial, aligned with response A or C of the RD probe (Fig. 1).

The S4 code argues that multiple measurements should be performed in order to reduce the uncertainty of a mean value, implying that the mean is the result that matters. This argument aligns with the set paradigm. On the other hand, the P2 code represents the idea that multiple measurements are beneficial because they allow the experimenter to identify outliers or mistakes in data collection. This argument aligns with the point paradigm.

Lastly, the U1 code represents responses that merely state that more data is needed. In this case, the respondent did not write a sufficient explanation to classify it into one paradigm or the other. It is possible that, if discussing the probe with the respondent in person, their underlying reasoning would become apparent. It is also possible that the respondent lacked the language to express their reasoning, or that they had not considered their reasoning to any greater depth. It is even possible that the student was merely pressed for time when completing the survey, and otherwise they would have provided an explanation that aligned well with another code. In any case, the PMQ coder has only the written response to interpret, and as such, is forced to assign a code such as U1 regardless of these hypothetical cases.

We note that, from the perspective of an expert experimental physicist, there is validity behind the reasoning

TABLE I.   Selected codes from the new PMQ coding scheme.

Probe	Identifier	Name	Definition: "Argument is that…"
RD	S4	Reduce uncertainty of mean	…multiple measurements will be used to reduce the error or uncertainty of the mean or average.
RD	P1	Measure the true value	…the experimenter could measure the correct value in a single measurement.
RD	P2	Identify the outliers after all measurements	…repeated measurements are needed in order to know which measurements were mistakes or outliers, after all measurements are taken. This code includes the idea that the experimenter must get the same result at least twice for it to be correct.
RD	U1	Just take more data	…experimenter needs to take more data. No statistical reasoning apparent.
UR	S1	Simply average	…I averaged, do the average, average is best, or it is the average, but does not elaborate. Includes statements that simply say what the reported value is.
UR	S3	Why average is appropriate in this case	…reporting the average is best because all of this data matters, or because the spread of this data is small enough. Includes reporting all data as well as the average.
UR	S4	Report average and spread	…experimenter should report the average and the uncertainty, range, or spread.
UR	S5	How to compute	…how to compute the average.
UR	P1	Choose single value	…experimenter should choose a single value to report (for any reason).
SMDS	S2	Smaller spread is better, no mention of external factors	…a smaller spread, uncertainty, range is better, more accurate, more precise, etc. The response does not mention external factors, outliers, human error, etc.
SMDS	S3	Smaller spread is better, due to external factors	…a smaller spread, uncertainty, range is better, more accurate, more precise, etc. The response mentions external factors, outliers, human error, etc.
SMDS	P1	The means are the same	…the groups agree because the means are the same.
SMDS	P4	Differences in carefulness	…differences in the spread are due to differences in how carefully the measurements were performed.
DMSS	S3	Similar means and spreads, mentions overlap	…the groups agree because the means and spreads are similar. Argument considers the overlap between the means and/or spreads of the two datasets.
DMSS	S4	Chose A, blank explanation	Respondent chose "A" but left the explanation blank.
DMSS	P2	Means must match	…the groups do not agree because the means are not the same (no mention of spread)
DMSS	P3	Means close enough, treats average as point	…the groups agree because the means are close enough
DMSS	U1	Not about statistics	…only nonstatistical things, such as systematics, are mentioned.
DMSS	U3	Misc.	Argument that doesn't fit into any of the other codes.

represented by both the S4 and the P2 codes, and even the U1 code cannot be said to be incorrect. Therefore, we do not intend to assign an inherent ranking or hierarchy to the codes in our code book. They aim only to classify common lines of reasoning used when responding to a given PMQ probe. Even at the paradigm level, though the set paradigm has been identified as more expertlike than the point paradigm in previous work, that does not mean that pointlike reasoning is not sometimes included in expert approaches more generally. For example, outliers are often the impetus for proposing causes and enacting revisions in the larger context of modeling, and identifying systematic errors is sometimes merely a more nuanced framing for catching mistakes in the measurement procedure [51]. When using our coding scheme to measure the effectiveness of a learning experience, as we do in this work, we intend for our codes to be compared to each other only in terms of how well they align with the goals of that learning experience. In general, interpreting results from our coding scheme by ranking codes relative to each other is best done in the context of a learning goal.

D. Student-level paradigms

In addition to interpreting the codes assigned to each students' response to any particular probe, we also consolidate students' responses to each of the four probes we analyzed into a single designation of that student's reasoning overall [32]. Because of the coarse-grained nature of this consolidation, we characterize student's reasoning only at the paradigm level. These student-level paradigms correspond to the number of probe-level paradigms emerging from each probe, as defined in Table II. A student's overall response is *pointlike* if their responses to the probes are represented only by P and U codes and no S codes. Conversely, an overall *setlike* response comes from

TABLE II. Definitions of overall student paradigms. Reproduced from Ref. [32].

Student paradigm	Number of P's	Number of S's
pointlike	≥ 1	0
setlike	0	≥ 1
mixed	≥ 1	≥ 1
mixed	0	0

responses to the probes that are represented only by S and U codes and no P codes. A third designation, *mixed*, refers to an overall response that includes both P and S codes, or to when the response is entirely represented by U codes. However, a response represented entirely by U codes only occurred 0.2% of the time in the datasets analyzed here.

E. Statistical methods

After responses were coded using the coding scheme described above, we analyzed the distributions of the assigned codes. Note that in this work, we compare distributions of students within a semester and PMQ administration (pre or post), rather than comparing matched responses student by student. We compare distributions of codes and demographic characteristics using Fisher's exact test [52], with a significance threshold of $p < 0.05$. When considering the significance of multiple tests at once, we apply the Holm-Bonferroni method [53] to correct for the problem of m multiple comparisons. We compute these tests using the base package included in the R programming language, version 3.6.2 [54]. For an additional visual indication of the uncertainty in counts or percentages of codes or paradigms, we plot the binomial proportion confidence interval at the 95% confidence level. When plotting the difference between the number of postresponses and preresponses for each code, we propagate these confidence intervals for the plotted uncertainty bar of the calculated difference. Finally, when two similarly measured proportions are both statistically significant based on these methods, we estimate the degree to which they are different by calculating an effect size using Cohen's h [55].

F. Course context

In this work, we compare two semesters of the introductory physics lab course at CU, one before the transformation of that course (Spring 2017, the "original course") and one after the course was transformed (Spring 2018, the "transformed course"). We compare two spring semesters, though the course is also taught in fall semesters, to avoid a range of factors that influence students of differing backgrounds enrolling in the fall versus in the spring. There were 641 students who completed the course at CU in Spring 2017, and 722 in Spring 2018. Of these students, 539 and 499, respectively, completed both the pretest and post-test, and were included

TABLE III. Self-reported gender, race or ethnicity, major, and year of students enrolled in the course in both Spring 2017 and Spring 2018. "Engineering" excludes the major Engineering Physics, which is included in "Physics." The p values from Fisher's exact test comparing the two semesters appear next to each dimension heading.

Gender	$p = 0.31$
Female	23.6%
Male	75.1%
Other Gender	1.3%
Race or ethnicity	$p = 0.81$
American Indian or Alaska Native	0.9%
Asian	14.4%
Black or African American	2.2%
Hispanic or Latino	8.8%
Native Hawaiian or other Pacific Islander	0.7%
White	69.0%
Other race or ethnicity	4.0%
Major	$p = 0.48$
Physics	17.2%
Engineering	44.8%
Other STEM	35.1%
Other disciplines	3.0%
Year	$p = 0.30$
First year	48.9%
Second year	31.5%
Third year	11.4%
Fourth year	6.2%
Fifth year and above	2.0%

in the dataset analyzed here. The self-reported gender, race and/or ethnicity, major, and year of students in these two semesters, collected using another research-based assessment that was administered at the same times as the PMQ, are shown in Table III. We include this information for various reasons [56], including to provide context for our research findings, as well as to enable metastudies that combat normative whiteness and highlight inequities in research [57]. We compared the proportions of students identifying with each of these categories in Spring 2017 and Spring 2018 as an indication of the similarity of the students entering the course during these two semesters. The resulting p values from Fisher's exact test are shown by each category heading in Table III. Along each of these dimensions, the populations of students in the two courses were statistically equivalent ($p > 0.05$).

The PMQ was administered electronically to students at the beginning (pre) and at the end (post) of the course during both semesters. Students were sent an internet link to complete the PMQ independently, and as an incentive for completing the questionnaire, were offered a small amount of participation-based course credit totaling 1–2% of their final grade in the course. In the original course, students completed the pretest and post-test as in-class assignments.

However, due to unavoidable scheduling circumstances, students in the transformed course completed the surveys outside of class. In that semester, students received the pretest link five days after the course started, and were required to complete their responses within seven days. Previous research on other online research-based assessments of student learning has shown little difference in matched student responses when completed outside or during class time, showing at most a small positive increase from taking the assessment in class overall [58].

Additionally, because of the change in timeline in the transformed course, the first lecture in that course occurred before 64% of the respondents completed their presurvey response. The content of that lecture touched on aspects related to measurement uncertainty, specifically the importance of every measurement having an associated uncertainty and how that uncertainty is used for comparing measurements. For the post-tests, students received the link close to the end of the semester, after they had completed all activities for the course, and were required to complete their response before the semester ended.

IV. RESULTS

Here, we present quantitative results from applying the PMQ coding scheme developed at CU to two semesters of the introductory physics lab course at CU: Spring 2017 (referred to as the original course, or before the transformation) and Spring 2018 (referred to as the transformed course, or after the transformation). For each semester, we compare the distribution of responses from the pretest to equivalent distributions from the post-test, as a "pre-to-post" comparison. We first present these comparisons at the student paradigm level, as the most simplified interpretation of our results. We then break down the results probe by probe, still interpreting responses at the paradigm level.

Lastly, we consider each probe at a level beyond the paradigms, comparing distributions of the codes that make up the paradigms. At each level, we note how each finding aligns or runs counter to the learning goal of the transformed course. We further discuss this alignment more broadly in the context of our research questions in Sec. V.

A. Paradigm-level results

Figure 2 shows the percentage of students whose responses to the PMQ fell into each student-level paradigm, *pointlike*, *mixed*, or *setlike*. The left panel shows the semester before the transformation, while the right show the semester after the transformation. Light gray bars represent the pretest, while darker gray represent the post-test.

The error bars in Fig. 2 suggest that, for both semesters, there were significant differences between pre and post for the number of *mixed* and the number of setlike responses. Fisher's exact test confirms those differences, all with $p \ll 0.05$. However, the proportions of pointlike responses were statistically similar ($p = 1$ for 2017 and $p = 0.34$ for 2018). Overall, at the student level students shifted predominantly from the mixed to setlike paradigm, both before and after the transformation.

Moving to the probe level, shifts between pre- and postparadigms for each of the four probes we analyzed are shown in Fig. 3. As before, the left panel shows the semester before the transformation, while the right shows the semester after the transformation. Within each panel, the four probes are represented on the vertical axis. The horizontal axis represents the proportion of students whose responses were coded with either S or P codes. Solid markers denote the proportion with an S code on the pretest, while open markers denote the proportion with a P code on the pretest. The end of the corresponding arrows
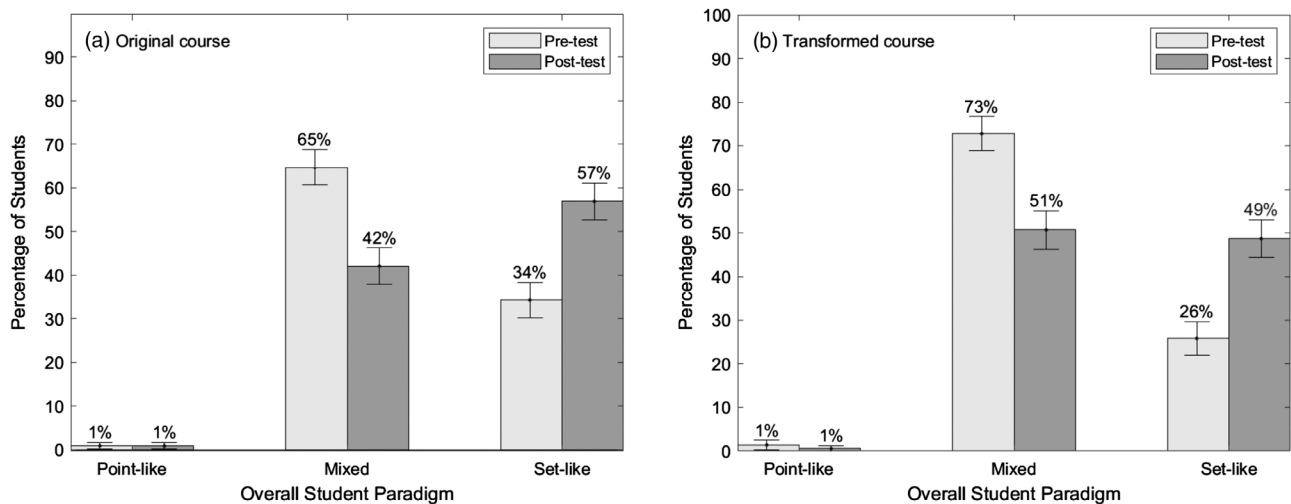

FIG. 2. Pre-post shifts at the student paradigm level. (a) Before transformation; (b) after transformation. Error bars are the binomial proportion confidence interval at the 95% confidence level.
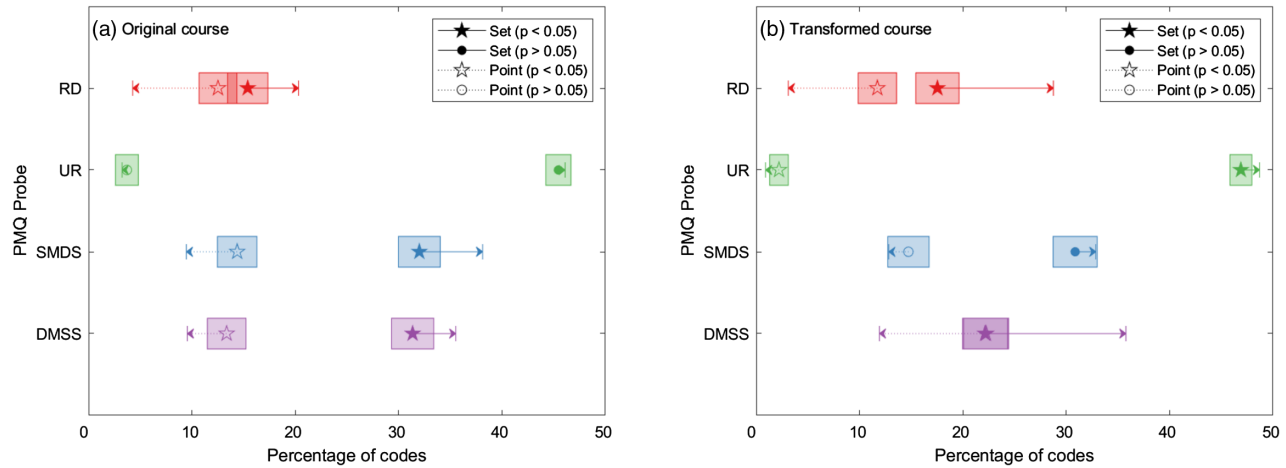
FIG. 3. Pre-post shifts at the probe paradigm level. (a) Before transformation; (b) after transformation. The horizontal axis represents the proportion of students responding in either the point or set paradigm, on each of four probes along the vertical axis. The shapes at the start of each arrow represent the pre-test proportion, while the location of the arrowhead represent the post-test proportion. Solid shapes represent the set paradigm proportion, while open shapes represent the point paradigm proportion. Star shapes represent a statistically significant pre-to-post shift using Fisher's exact test at the 95% confidence level; circle shapes represent a shift that is not significant by that same test. Shaded boxes represent the pre-test binomial proportion confidence interval at the 95% confidence level, as a guide to the eye.

show the corresponding proportions on the post-test. The significance of these shifts is also indicated.

Across all probes, students shifted to more S reasoning and less P reasoning during the semester. However, only some of these shifts were statistically significant. In the RD probe, the transformed course showed larger shifts towards S than the original course (with Cohen's h effect sizes of $h = 0.45$ and $h = 0.21$, respectively), while in the SMDS probe, the transformed course showed no significant shifts at all. The UR probe showed little practical significance in either semester due to the large proportion of S responses, even on the pretest. We have seen this saturation effect before in the UR probe when analyzing responses from CU students at the paradigm level [10,40].

The DMSS probe shows a larger shift towards S reasoning in the transformed course than in the original course. However, this shift is due to a difference in the proportion of S responses in the pretests, rather than the post-tests. Such a difference stands in contrast to the other three probes, which show similar pretest proportions between the two semesters. Furthermore, this difference in pretest responses stands in contrast to all of the information we have available on the distribution of students who enrolled in the class for these two semesters, which would suggest that the two groups of students are similar. Our best guess as to the cause of this difference in the DMSS pretest proportions concerns the differences in the timelines of the two semesters. We speculate that responses in the transformed semester to the DMSS probe in particular were affected by the first lecture of the course, for the 64% of students who completed the survey after that lecture. That lecture touched on the idea that uncertainty is

used for comparing measurements in a generalized way, an idea that relates to the DMSS probe. However, it also relates to the SMDS probe, from which the paradigms of preresponses seem similar before and after the transformation. Given this uncertainty, we proceed with caution when further analyzing DMSS responses from these two semesters, remembering that the full story around this portion of the dataset remains unclear.

B. Code-level results

We now analyze results beyond the level of paradigms, considering the individual codes themselves in the context of the course transformation. For each probe and semester, we plot the difference between the number of responses to each code on the post-test and on the pretest. In addition to the error bars that represent the uncertainty of these differences, we use blue bars to represent codes in which the pre- and post-test distributions are statistically different using Fisher's exact test ($p < 0.05$), and yellow bars for the codes in which the pre- and post-test distributions are statistically the same ($p > 0.05$). We apply the Holm-Bonferroni correction, with m as the number of codes in the given probe's code book, to the p values from Fisher's exact test before determining statistical significance.

1. The RD probe

A comparison for the codes in the RD code book is shown in Fig. 4. In both semesters, the code with the largest change from pre to post was S4. The prominent increase in S4 is encouraging, as it is aligned with the learning goals of the course, in particular the idea that all numbers have an
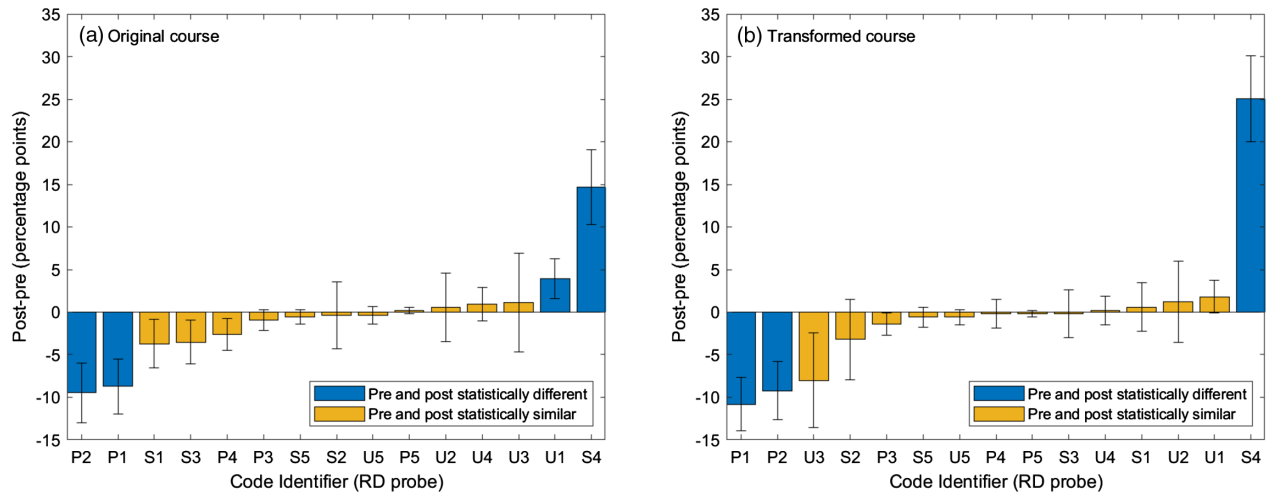
FIG. 4. Pre-post differences in code counts for the RD probe, for the original course (a) and the transformed course (b). Blue bars are statistically significant differences using Fisher's exact test at the 95% confidence interval, adjusted using the Holm-Bonferroni method. Orange bars are not significant by that same test. Error bars are the binomial proportion confidence interval at the 95% confidence level.

uncertainty, including the mean of a set of data. Before the transformation, the other code that increased was U1, while that code did not significantly change after the transformation. U1 represents a response that does not display more sophisticated reasoning than the other codes. Therefore, the lack of an increase in U1 after the transformation compared to before could suggest that students articulated their reasoning with greater sophistication in the transformed course.

Considering the codes that decreased from pre to post, there were two codes that showed a significant pre-to-post decrease, and they were the same codes in both the transformed and the original course. As they were both pointlike, that change is aligned with the goals of the transformation.

2. The UR probe

Figure 5 shows a comparison of the codes in the UR code book. In both semesters, the most prominent change was an increase of a single setlike code. This consolidation phenomenon in UR responses is discussed in Ref. [10]. In the original course, the code into which students consolidated was S1, while in the transformed course the code as S4. S1 represents reporting the average as the result of a set of measurements, while S4 represents reporting an average as well as a spread. S4 aligns with the transformation's learning goals, as it recognizes the importance of the spread of a distribution and aligns with the idea that all numbers have an uncertainty.
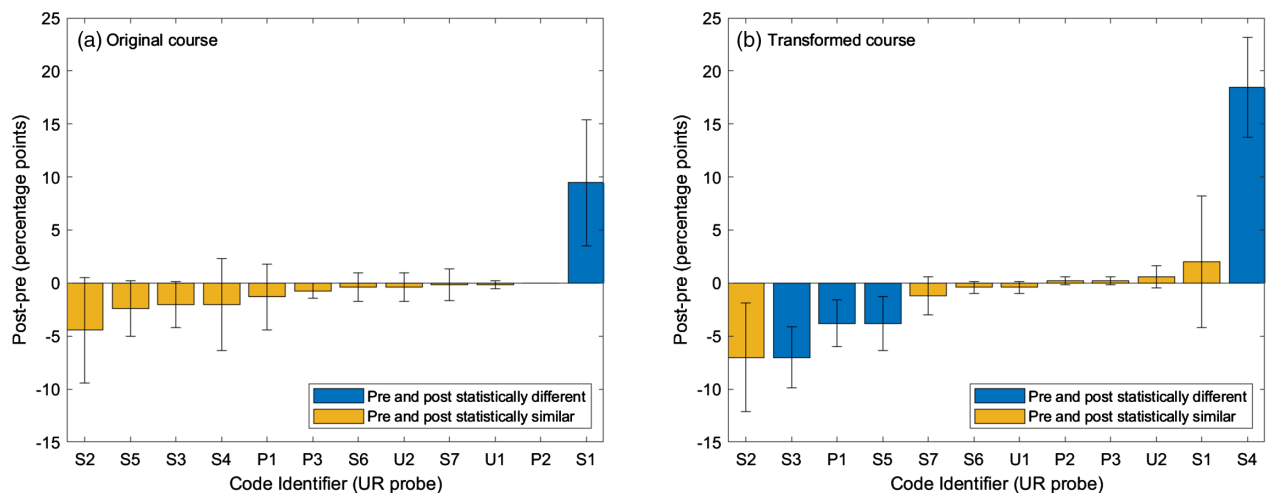

FIG. 5. Pre-post differences in code counts for the UR probe, for the original course (a) and the transformed course (b). (b) Reproduced from Ref. [10]. Blue bars are statistically significant differences using Fisher's exact test at the 95% confidence interval, adjusted using the Holm-Bonferroni method. Orange bars are not significant by that same test. Error bars are the binomial proportion confidence interval at the 95% confidence level.

There were no other statistically significant pre-post decreases in UR responses in the original course. However, there were three significant decreases in the transformed course: P1, S3, and S5. P1 represents canonical pointlike reasoning, that one should choose the value from a single trial to represent the result of an experiment. A decrease in this code is aligned with the goals of the transformation. The other two codes that decreased in the transformed course were setlike. One of them, S3, discusses the purpose of reporting an average, representing conceptual reasoning around the role of averages in experimentation. The second, S5, states the mathematical process of calculating an average, and represents the basic skill or practice of reporting an average. The transformation's goals included the reasoning represented by both of these codes, suggesting that the decreases in S3 and S5 may represent further room for improvement.

3. The SMDS probe

We compare responses to the SMDS probe in Fig. 6. Overall, the magnitude of the shifts on any of these codes are relatively small, suggesting that the type of reasoning solicited by the SMDS probe is relatively stable in our population of students. In fact, in the transformed course, no single probe showed a statistically significant change pre to post.

Before the transformation, the code that significantly increased was S3, which states that a smaller spread is better because of external factors such as "air resistance" or "human error." While the recognition of spread playing a role in data comparison is aligned with setlike reasoning, and thereby the goals of the transformation, the focus on external factors over inherent statistical variation is more pointlike than setlike. The code that decreased pre to post

before the transformation was P4, which talks about differences in carefulness between the two experimenters. This idea aligns with the point paradigm if the lack of carefulness manifests as mistakes in individual trials. However, there is a subtle difference between this line of reasoning and the idea that the spread of a dataset overall is affected by differing tendencies of experimenter care. Taken together, the SMDS trends observed in the original course contain elements that are both closer and further away from setlike reasoning.

4. The DMSS probe

Lastly, a comparison of DMSS codes appears in Fig. 7. In both semesters, the most prominent pre-to-post increase was S3, with an especially large increase in the transformed course. S3 represents the most complete way to compare results under the set paradigm, by looking for overlap between the means and spreads of the two datasets.

The S4 code, choosing the best multiple choice answer but leaving the explanation field blank, also increased significantly in the original course. There could be many reasons for that difference between the two semesters, including time limitations stemming from the different settings in which the survey was administered. The difference in S4 responses yields little insight into student learning or the transformation. Similarly, the U3 code, which decreased in both semesters, represents miscellaneous reasoning, and yields little insight without further qualitative interpretation of these responses. Likewise the U1 code, which decreased in the transformed course, represents non-statistical reasoning and itself offers little insight. That is not to say that these responses, which likely contain sophisticated reasoning about systematic effects and other experimental considerations, are not worthy of
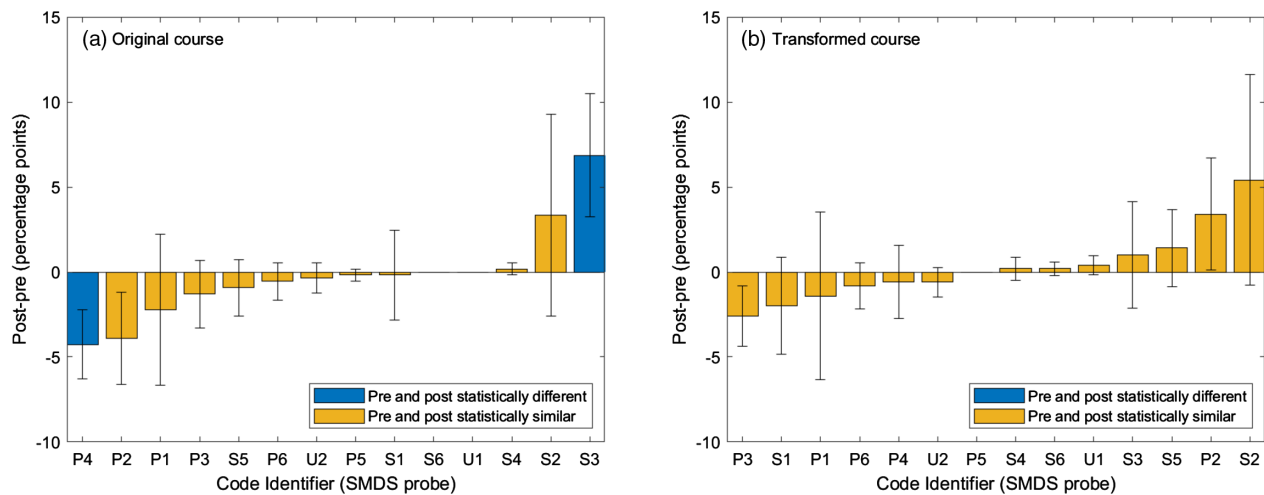

FIG. 6.   Pre-post differences in code counts for the SMDS probe, for the original course (a) and the transformed course (b). Blue bars are statistically significant differences using Fisher's exact test at the 95% confidence interval, adjusted using the Holm-Bonferroni method. Orange bars are not significant by that same test. Error bars are the binomial proportion confidence interval at the 95% confidence level.
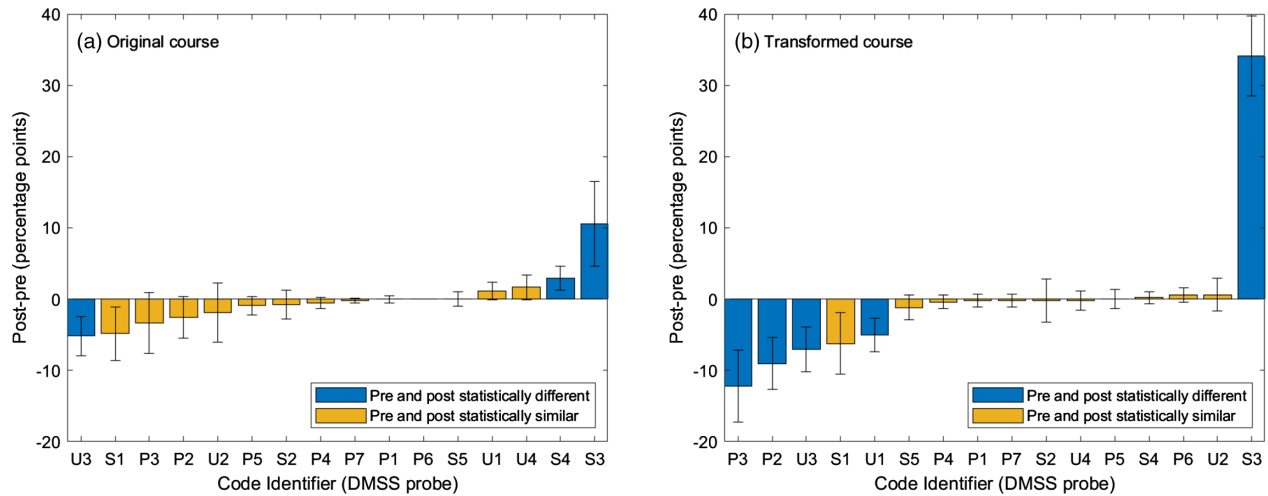
FIG. 7. Pre-post differences in code counts for the DMSS probe, for the original course (a) and the transformed course (b). Blue bars are statistically significant differences using Fisher's exact test at the 95% confidence interval, adjusted using the Holm-Bonferroni method. Orange bars are not significant by that same test. Error bars are the binomial proportion confidence interval at the 95% confidence level.

study. They are simply outside the scope of the point and set paradigms and the focus of the PMQ.

Two remaining codes decreased in prevalence during the transformed course: P2 and P3. Both consider only the mean of the two data sets when comparing them, P2 concluding that the means must match for the results to agree, and P3 concluding that the means are close enough to agree. Both responses lack reasoning around distributions or statistical uncertainty. A decrease in their prevalence after the transformation, given the absence of such a decrease before the transformation, is an indication that the transformation was effective.

V. DISCUSSION AND CONCLUSIONS

In this section, we synthesize the results presented above through the lens of the course transformation's learning goal around measurement uncertainty, which was, "Students should demonstrate a setlike reasoning when evaluating measurements." We apply the lens of that learning goal to both the original and the transformed course. We consider the original course, in addition to the transformed course, through that lens for two reasons: first, as a baseline of comparison for the transformed course, and second as an example of an entirely traditional physics lab course that nonetheless achieved measurable and desirable learning outcomes.

A. (Q1) Effectiveness of course overall

Regarding (Q1), *Did students respond to the PMQ in ways more aligned with the set paradigm after taking the introductory lab course, compared to when they began the course?*, on each of the levels of analysis presented here, both the original and the transformed course met the

learning goal to some extent. At the most simplified level of student paradigms, our analysis shows increases in setlike reasoning and decreases in mixed reasoning in both semesters, which is aligned with the learning goal. At a finer level of detail looking at paradigms probe by probe, each of the four PMQ probes that we analyzed showed, in both semesters, pre-to-post increases in setlike reasoning and pre-to-post decreases in pointlike reasoning. In all but two cases, these increases were statistically significant. Finally, in the most fine-grained interpretation of the results, looking at individual codes beyond their paradigms, there were pre-to-post changes in each of the four probes that aligned with the learning goal to some extent. In particular, significant increases in RD-S4 and DMSS-S3, and significant decreases in RD-P1 and RD-P2, were observed in both semesters, and unambiguously align with the learning goal.

We speculate that these two probes, RD and DMSS, showed significant and consistent changes in both semesters because of the particular contexts they concern. The RD probe centers around data collection, and thus touches on larger ideas of experimental design, epistemology, and the nature of measurement. This broad scope perhaps made RD more open ended than the contexts of the other probes, and thus allowed for a more pronounced pre-to-post change in responses, as students were able to draw on the content of the course (specifically the uncertainty of the mean in RD-S3) in the post test. On the other hand, the DMSS probe presents a specific example of data comparison, and thus has a more narrow conceptual scope than the RD probe, in a sense. Furthermore, the comparison presented in DMSS highlights the role of spread in making comparisons, as opposed to the other data comparison probe, SMDS, which focuses more on mean values. Again, we speculate that

students were readily able to draw on the content of the course in the context of the DMSS probe (specifically concepts around spreads and distributions in DMSS-S3) in the post test, leading to the pronounced changes we observed in this probe.

B. (Q2) Effectiveness of transformation

Regarding (Q2), *Did student responses to the PMQ after the transformation show greater change towards the set paradigm than responses before the transformation?*, there were several indications that the transformed course achieved the learning goal to a greater extent than the original course. There were also some indications that it did not, suggesting directions for future improvement. However, all of these indications lie at finer levels of analysis than that of student paradigms, in which the semesters before and after transformation appear similar in all respects.

At the level of paradigms for each probe, the RD probe showed a striking pre-to-post increase in S codes after the transformation, with an effect size of $h = 0.45$, as compared to the corresponding increase before the transformation with an effect size of $h = 0.21$. This difference suggests that the course transformation was especially successful in the context of evaluating choices in data collection.

On the other hand, in the SMDS probe, we observed significant pre-to-post shifts in the original course, but not in the transformed course. This suggests that there is more for students to learn in the transformed course around data comparison, particularly in the context presented by this probe: when the two datasets being compared have identical means but different spreads. In such a situation, deciding whether the results agree is very simple, and once can entirely ignore the spreads of the two distributions. Accordingly, the most common SMDS pointlike code across both semesters was SMDS-P1 (18% of all responses in the data set), which represents this simple approach.

However, there is more to consider when deciding not whether the results agree, but which result is better overall. The SMDS probe asks respondents to do this latter task. With this broader task in mind, the fact that one result has a smaller spread becomes relevant, as represented by the SMDS-S2 code, the most common code of any paradigm assigned to SMDS responses (53% of all responses in the dataset). Nonetheless, it is possible that the probe does not prompt setlike reasoning as directly as other probes, as the same means encourage students to stop there without considering the dataset at a deeper level. For the transformed course to improve further, results from the SMDS probe suggest that students could be better supported in using setlike thinking all the time, not just when the situation lends itself to it. Perhaps including more focused or nuanced discussions around what makes a dataset better or worse would result in more favorable SMDS

responses, and more importantly, further improve physics laboratory instruction.

Regarding the other two probes, less can be drawn from the paradigm-level results. While pre-to-post shifts in the UR probe were statistically significant only in the transformed course, and in directions aligned with the learning goal, the overwhelming prevalence of S responses in all cases makes this result have little practical significance. The effect sizes of the DMSS probe are also more favorable in the transformed course, but because this difference is due to differences in pretest proportions rather than post-test proportions, we hesitate to interpret it further.

Finally, analyzing pre-to-post differences in each individual code yields further insight into the success of the transformation. In the RD probe, an absence of an increase in RD-U1 after the transformation indicates not only that more students communicated in alignment with the learning goal (as established earlier in this section), but also that the transformation allowed them to do so with greater sophistication. In the UR probe, a consolidation of responses into UR-S4 in the transformed course, rather than UR-S1 in the original course, is additional evidence of more sophisticated reasoning, this time regarding the idea that every number has an uncertainty. However, decreases in UR-S3 might suggest that the transformed course also deemphasized a more sophisticated conceptual understanding of the role of means, which could be a focus of further improvement. In the SMDS probe, there were no clear messages from analyzing pre-post differences code by code, underscoring the inherent consistency of responses to this probe. While the changes in the original course seem at first to align with the learning goal, further consideration of the reasoning they represent complicates this picture (as discussed above). More qualitative study is needed to better understand how students are interpreting, reasoning, and responding to the SMDS probe.

In the DMSS probe, the observed decrease in DMSS-P2 and DMSS-P3 in the transformed course would suggest that the transformation was successful at encouraging setlike reasoning around data comparison. However, the irregularities around the DMSS probe, as discussed above and in the next section, cast doubt on the full implications of this finding.

C. Successes and limitations

We start by noting a success regarding research methodology, noting that qualitatively different insight emerged as we proceeded to each deeper level of analysis detail. Indications that the transformation was successful at meeting the learning goal around measurement uncertainty emerged only when considering responses probe by probe, and evidence about the depth of that learning emerged only when investigating responses code by code. More generally, these observations are merely a reminder that there is far more to learning around measurement uncertainty than

is captured by the point and set paradigms. However, this bigger picture also comes with a limitation: here, we can study only the reasoning prompted by the probes of the PMQ, which is a smaller scope than all that is important to student learning around measurement uncertainty.

Furthermore, studying responses to surveys come with limitations in general, as the administration of surveys precludes the ability to ask follow-up questions. When applying our coding scheme in this work, we had only the written survey responses to interpret, complete with any ambiguities in what the student was actually thinking. These ambiguities often forced us to assign U codes to such responses. For those students, more interactive methods (such as interviews) would allow us to better distinguish their ability to communicate reasoning from the nature of their reasoning itself.

We also note that the responses we analyze came from a single institution, CU, which is a large, research-focused, highly resourced, primarily white institution of a type that is overrepresented in literature [57]. A broader range of students and institutions is needed to determine whether these findings hold beyond CU. Additionally, we studied student reasoning at the introductory undergraduate level, and we expect these findings to apply only to students at similar academic levels.

While we did not directly compare pre- and postresponses student by student in this analysis, we still included in our dataset only matched responses, that is, only those students who completed both pre- and post-test. We did this because we focused our analysis on measuring the effectiveness of our course, rather than making any direct comparisons to other courses or institutions. This framing is distinct from the motivations outlined in Ref. [59], which calls for using statistical methods to model student responses and predict the missing responses from students with unmatched data. However, our comparisons between the original and the transformed course, given the similarity in student demographics between the two semesters, and that they come from the same institutional context and instructor, are less affected by the bias identified in that reference. Furthermore, it would require a larger dataset than currently available to apply the techniques described in Ref. [59] to the nominal data of paradigm or coding designations. Nonetheless, and unavoidably, there could be some bias in the changes we observe based on exogenous factors that affect both a student's reasoning around measurement uncertainty and their likelihood to complete both the pre- and post-tests.

Because of the straightforward pretest and post-test methodology we employ in this study, and the lack of any data between these two survey administrations, we can only speculate on which particular aspects of the transformation resulted in the changes we observed, or when during the semester they occurred. Moreover, given that learning occurs as interactions between many factors that persist and evolve throughout a course, connecting individual factors to measured learning outcomes would require a more careful design study in which changes to the course were made one at a time. Thus, while it is tempting to tie changes to particular aspects of the transformation, we leave such rich and compelling questions to further studies. Here, we aim only to attribute the changes we measured to the course transformation as a whole.

Lastly, the differences in timing of the pre-test between the original and transformed courses casts some doubt if the pretest in the transformed course is a valid baseline to which to compare the post-test of that semester. With 64% of students completing the pretest after the first lecture in the transformed semester, this could potentially bias, but not eliminate, any measured learning outcomes from that first lecture. However, considering the focus on measurement uncertainty throughout the course, the first lecture is a very small portion of all of the learning opportunities that the students experienced throughout the course. Furthermore, when answering (Q2) by comparing the transformed course to the original one, additional instructional opportunities before the pretest in the transformed course would cause measured pre-to-post changes to have a smaller effect size than otherwise, assuming the instruction has an overall effect aligned with the learning goal. Results from (Q1) suggest that instruction does indeed shift students toward the set paradigm overall, thus, the difference in timing would result in a decrease in apparent shifts towards setlike reasoning in the transformed course. Given that we observe evidence for the opposite effect, the timing difference is less of a concern. However, irregularities in the transformed course pretest results, specifically in the DMSS probe, remain a mystery, and require further investigation before results from that probe can be taken at face value.

VI. SUMMARY

Here, we used the PMQ to measure the effectiveness of the introductory lab course at CU, and a recent transformation of that course. We aimed to answer two research questions: (Q1), *Did students respond to the PMQ in ways more aligned with the set paradigm after taking the introductory lab course, compared to when they began the course?*, and (Q2), *Did student responses to the PMQ after the transformation show greater change towards the set paradigm than responses before the transformation?* With regards to (Q1), we see strong evidence of PMQ responses that are more aligned with the set paradigm in the post-tests from both semesters, compared to the corresponding pretests, and we see this evidence at all levels of analysis, from the coarsest to the finest grain sizes. With regards to (Q2), we see evidence that PMQ responses in the transformed course shifted pre to post towards more prevalent setlike reasoning compared to those from the original course. Furthermore, we also see some evidence that the responses from the transformed course tend to shift

towards more sophisticated reasoning than in the original course. We also identified specific aspects of a sophisticated understanding of measurement uncertainty that were less apparent in responses from the transformed course than from the original course, suggesting areas for further improvement. These findings add to the growing body of evidence that physics lab courses, even traditional ones, have value by creating opportunities for students to learn important aspects of conceptual physics and to develop expert physics practices.

APPENDIX: PMQ PROBES

The first probe of the PMQ is the RD probe, which appears in Fig. 1.

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 <u>horizontally</u> 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 metre stick is used to measure *d* and *h*.
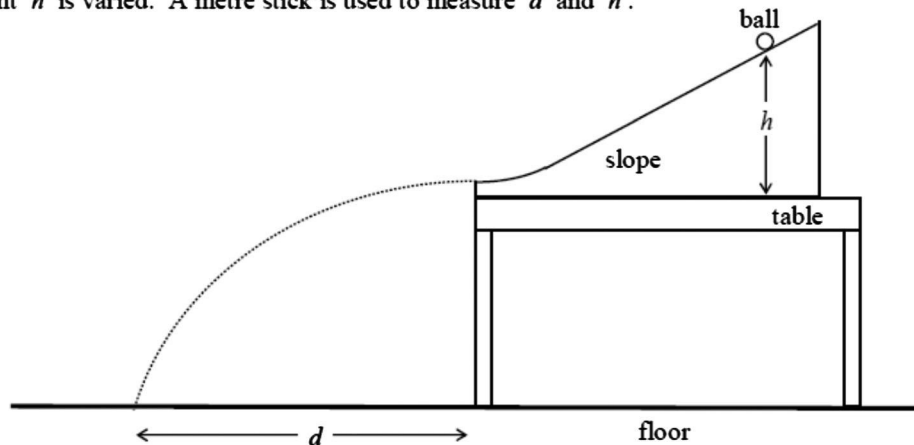

FIG. 8. Contextual information for the PMQ. This information precedes the probes themselves. Reproduced from Ref. [47].

The students continue to release the ball down the slope at a height $h = 400$ mm.
Their results after five releases are:

Release	d (mm)	
1	436	
2	425	[426 in post questionnaire]
3	440	[438]
4	425	[426]
5	434	

The students then discuss what to write down for d as their final result.
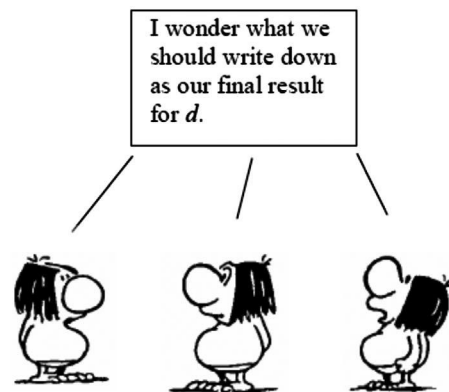
Write down what you think the students should record as their final result for d.

Explain your choice.

FIG. 9.   The UR probe of the PMQ. Reproduced from Ref. [47].

Two 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 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

Our results are better. They are all between 424 mm and 444 mm. Yours are spread between 410 mm and 460 mm.

Our results are just as good as yours. Our average is the same as yours. We both got 435 mm for d.

I think the results of group B are better than the results of group A.

A   B   C

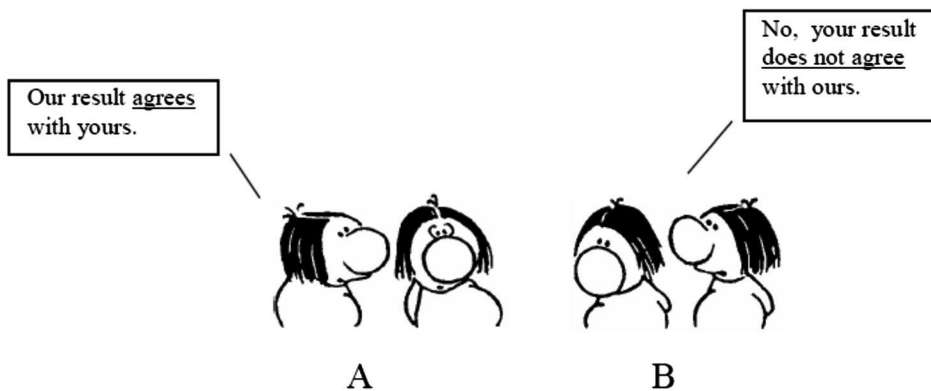With which group do you <u>most closely</u> agree?   (Circle ONE): | A | B | C |

Explain your choice.

FIG. 10.   The SMDS probe of the PMQ. Reproduced from Ref. [47].

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 A d (mm)	Group B d (mm)
1	440	432
2	438	444
3	433	426
4	422	433
5	<u>432</u>	<u>440</u>
Average:	**433**	**435**

No, your result <u>does not agree</u> with ours.

Our result <u>agrees</u> with yours.

A　　　　　　　B

With which group do you <u>most closely</u> agree?　(Circle ONE):　　A　B

Explain your choice.

FIG. 11.　The DMSS probe of the PMQ. Reproduced from Ref. [47].

[1] PCAST STEM Undergraduate Education Working Group, *Engage to Excel: Producing One Million Additional College Graduates with Degrees in STEM* (Executive Office of the President, Washington, DC, 2012).

[2] National Research Council, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (National Academy Press, Washington, DC, 2012).

[3] AAPT Committee on Laboratories, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (American Association of Physics Teachers, College Park, MD, 2014).

[4] Joint Task Force on Undergraduate Physics Programs, *Phys21: Preparing Physics Students for 21st Century Careers* (American Association of Physics Teachers, College Park, MD, 2016).

[5] C. Moskovitz and D. Kellogg, Inquiry-based writing in the laboratory course, Science **332,** 919 (2011).

[6] B. M. Zwickl, N. Finkelstein, and H. J. Lewandowski, The process of transforming an advanced lab course: Goals, curriculum, and assessments, Am. J. Phys. **81,** 63 (2013).

[7] B. R. Wilcox and H. J. Lewandowski, A summary of research-based assessment of students' beliefs about the nature of experimental physics, Am. J. Phys. **86,** 212 (2018).

[8] E. M. Smith, N. Chodkowski, and N. G. Holmes, Context of authority may affect students' evaluations of measurement, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2020).

[9] E. M. Smith, M. M. Stein, C. Walsh, and N. Holmes, Direct Measurement of the Impact of Teaching Experimentation in Physics Labs, Phys. Rev. X **10,** 011029 (2020).

[10] B. Pollard, R. Hobbs, D. R. Dounas-Frazer, and H. J. Lewandowski, Methodological development of a new coding scheme for an established assessment on measurement

uncertainty in laboratory courses, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2020).

[11] H. J. Lewandowski, B. Pollard, and C. G. West, Using custom interactive video prelab activities in a large introductory lab course, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2020).

[12] H. J. Lewandowski, D. R. Bolton, and B. Pollard, Initial impacts of the transformation of a large introductory lab course focused on developing experimental skills and expert epistemology, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2018).

[13] N. Holmes, J. Olsen, J. L. Thomas, and C. E. Wieman, Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content, Phys. Rev. Phys. Educ. Res. **13**, 010129 (2017).

[14] N. G. Holmes and C. E. Wieman, Introductory physics labs: We can do better, Phys. Today **71**, No. 1, 38 (2018).

[15] M. F. J. Fox, A. Werth, J. R. Hoehn, and H. J. Lewandowski, Teaching labs during a pandemic: Lessons from Spring 2020 and an outlook for the future, arXiv:2007.01271.

[16] R. F. Lippmann, Students' understanding of measurement and uncertainty in the physics laboratory: Social construction, underlying concepts, and quantitative analysis, Ph.D. thesis, University of Maryland, College Park, 2003.

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

[18] R. J. Beichner, J. M. Saul, D. S. Abbott, J. J. Morse, D. L. Deardorff, R. J. Allain, S. W. Bonham, M. H. Dancy, and J. S. Risley, The Student-Centered Activities For Large Enrollment Undergraduate Programs (SCALE-UP) project, *Research-Based Reform of University Physics*, edited by E. F. Redish and P. J. Cooney (American Association of Physics Teachers, College Park, MD, 2007), Reviews in PER Vol. 1, http://www.per-central.org/document/ServeFile.cfm?ID=4517.

[19] D. S. Abbott, Assessing student understanding of measurement and uncertainty, Ph.D. thesis, North Carolina State University, 2003.

[20] E. Etkina and A. V. Heuvelen, Investigative Science Learning Environment—A science process approach to learning physics, in *Research-Based Reform of University Physics*, edited by E. F. Redish and P. J. Cooney (American Association of Physics Teachers, College Park, MD, 2007), Reviews in PER Vol. 1, http://www.per-central.org/document/ServeFile.cfm?ID=4988.

[21] E. Etkina, A. Karelina, M. Ruibal-Villasenor, D. Rosengrant, R. Jordan, and C. E. Hmelo-Silver, Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories, J. Learn. Sci. **19**, 54 (2010).

[22] N. G. Holmes, Structured Quantitative Inquiry Labs: Developing critical thinking in the introductory physics laboratory, Ph.D. thesis, The University of British Columbia, 2014.

[23] L. E. Strubbe, J. Ives, N. G. Holmes, D. A. Bonn, and N. K. Sumah, Developing student attitudes in the first-year physics laboratory, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 340–343.

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

[25] E. M. Smith, M. M. Stein, C. Walsh, and N. G. Holmes, Direct Measurement of the Impact of Teaching Experimentation in Physics Labs, Phys. Rev. X **10**, 011029 (2020).

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

[27] R. Lippmann Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, NorDiNa **4**, 40 (2006).

[28] N. G. Holmes and D. A. Bonn, Doing science or doing a lab? Engaging students with scientific reasoning during physics lab experiments, in *Proceedings of the 2013 Physics Education Research Conference, Portland, OR*, edited by P. V. Engelhardt, A. D. Churukian, and D. L. Jones (AIP, New York, 2014), pp. 185–188.

[29] N. Majiet and S. Allie, Student understanding of measurement and uncertainty: Probing the mean, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2018).

[30] M. M. Stein, C. White, G. Passante, and N. G. Holmes, Student interpretations of uncertainty in classical, and quantum mechanics experiments, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2020).

[31] A. Madsen, S. B. McKagan, E. C. Sayre, and C. A. Paul, Resource Letter RBAI-2: Research-based assessment instruments: Beyond physics topics, Am. J. Phys. **87**, 350 (2019).

[32] B. Pollard, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab course on students' understanding of measurement uncertainty, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by L. Ding, A. Traxler, and Y. Cao (AIP, New York, 2017), pp. 312–315.

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

[34] N. G. Holmes, J. Day, A. H. K. Park, D. A. Bonn, and I. Roll, Making the failure more productive: scaffolding the invention process to improve inquiry behaviors and outcomes in invention activities, Instr. Sci. **42**, 523 (2014).

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

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

[37] C. Walsh, K. N. Quinn, C. Wieman, and N. Holmes, Quantifying critical thinking: Development and validation of the Physics Lab Inventory of Critical Thinking, Phys. Rev. Phys. Educ. Res. **15,** 010135 (2019).

[38] B. Campbell, F. Lubben, A. Buffler, and S. Allie, 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), http://www.phy.uct .ac.za/sites/default/files/image_tool/images/281/people/ buffler/physics_education/Monograph%202005.pdf.

[39] B. Pollard and H. J. Lewandowski, Transforming a large introductory lab course: Impacts on views about experimental physics, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2019).

[40] H. J. Lewandowski, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and B. Pollard, Student reasoning about measurement uncertainty in an introductory lab course, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by L. Ding, A. Traxler, and Y. Cao (AIP, New York, 2018), pp. 244–247.

[41] F. Lubben and R. Millar, Children's ideas about the reliability of experimental data, Int. J. Sci. Educ. **18,** 955 (1996).

[42] T. S. Volkwyn, First year students' understanding of measurement in physics laboratory work, Ph.D. thesis, University of Cape Town, 2005.

[43] R. Millar, F. Lubben, R. Gott, and S. Duggan, Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance, Res. Papers Educ. **9,** 207 (1994).

[44] R. Millar, R. Gott, F. Lubben, and S. Duggan, Children's performance in investigative tasks in science: A framework for considering progression, in *Progression in Learning*, edited by M. Hughes (Multilingual Matters Ltd., Clevedon, 1996), pp. 82–108.

[45] 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).

[46] A. Buffler, S. Allie, F. Lubben, and B. Campbell, Evaluation of a research-based curriculum for teaching measurement in the first year physics laboratory, in *Proceedings of the 4th Conference of the European Science Education Research Association (ESERA) in Noordwijkerhout, the Netherlands on 19-23 August 2003* (2003).

[47] 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).

[48] For readability, we omit the meanings behind these probe names here, describing those that are relevant to this work later in this section. A full description of the probes of the PMQ can be found in Table 2 of Ref. [38].

[49] Qualtrics, https://www.qualtrics.com (2005).

[50] See Supplemental Material at http://link.aps.org/ supplemental/10.1103/PhysRevPhysEducRes.16.020160 for the complete code books of the new coding scheme for four PMQ probes.

[51] D. R. Dounas-Frazer and H. J. Lewandowski, The Modelling Framework for Experimental Physics: Description, development, and applications, Eur. J. Phys. **39,** 064005 (2018).

[52] R. A. Fisher, On the interpretation of $\chi 2$ from contingency tables, and the calculation of P, J. R. Stat. Soc. **85,** 87 (1922).

[53] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. **6,** 65 (1979), https://www.jstor .org/stable/4615733.

[54] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria (2019).

[55] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Routledge, London, 2013), p. 567.

[56] A. N. Parks and M. Schmeichel, Obstacles to addressing race and ethnicity in the mathematics education literature, J. Res. Math. Educ. **43,** 238 (2012).

[57] S. Kanim and X. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16,** 020106 (2020).

[58] B. R. Wilcox and H. Lewandowski, Students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for experimental physics, Phys. Rev. Phys. Educ. Res. **12,** 010123 (2016).

[59] J. Nissen, R. Donatello, and B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. **15,** 020106 (2019).