Investigating changes in student views of measurement uncertainty in an introductory physics lab course using clustering algorithms

Alexandra Werth,[1,2,*] Benjamin Pollard,[1,2,3] Robert Hobbs,[4] and H. J. Lewandowski[1,2]

[1]*Department of Physics, University of Colorado, Boulder, Colorado 80309, USA*
[2]*JILA, National Institute of Standards and Technology and University of Colorado, Boulder, Colorado 80309, USA*
[3]*Department of Physics, Worcester Polytechnic Institute, Worcester, Massachusetts 01609, USA*
[4]*Department of Physics, Bellevue College, Bellevue, Washington 98007, USA*

(Received 6 August 2023; accepted 27 September 2023; published 18 October 2023)

Understanding measurement uncertainty is crucial in experimental physics, as it enables accurate and reliable measurements, facilitates comparison between measurements, and aids in designing experiments. Consequently, measurement uncertainty has emerged as a critical learning goal for many introductory physics labs. Here, we explore the impact of a recently transformed introductory physics lab at the University of Colorado Boulder on student understanding and interpretation of measurement uncertainty. The transformed course was explicitly designed to prioritize understanding measurement uncertainties as a learning goal and replaced verification labs with measurements where students could not predict the outcomes in advance. We used the physics measurement questionnaire to assess changes in student reasoning about measurement uncertainty at the beginning and end of the semester. Using a subparadigm coding scheme, we assessed different types of prevalent student reasoning and observed for trends in reasoning surrounding measurement uncertainty from the beginning to end of the lab course. Clustering algorithms were utilized to categorize student reasoning and compare these pre- and postsurvey responses. This analysis offers valuable insights into students' reasoning about measurement uncertainty, including the diversity of initial reasoning clusters and the narrowing of reasoning elements into primarily more expertlike responses after the transformed course. However, challenges were observed in transitioning students from certain clusters, especially those that exhibited brevity in their presurvey responses. Overall, the findings reveal the potential for targeted interventions to deepen these students' understanding of measurement uncertainty in experimental physics and underscore the significance of evidence-based instructional strategies in physics labs for improving student learning outcomes.

DOI: 10.1103/PhysRevPhysEducRes.19.020146

I. INTRODUCTION

Laboratory courses play a vital role in undergraduate physics curricula, offering students hands-on experience and opportunities to develop practical skills. Physics education research has increasingly focused on identifying effective teaching practices in introductory and advanced lab courses, encompassing various goals and objectives [1–7]. One prominent goal, as highlighted by the American Association of Physics Teachers Recommendations for the Undergraduate Physics Laboratory Curriculum, is for students to recognize and comprehend the limitations and uncertainties inherent in measurements and measurement devices [8].

Several institutions have undergone lab transformations with a specific emphasis on teaching the concept of measurement uncertainty [9–18]. For instance, Cornell University's introductory calculus-based physics lab series was revamped to emphasize conceptual introductions to measurement uncertainty, resulting in improved student views on the importance of uncertainty in evaluating the trustworthiness of results [9,10]. Similarly, the Scientific Community Laboratory (SCL) at the University of Maryland recognizes measurement uncertainty as a critical component of teaching students how to produce, analyze, and evaluate scientific evidence, and places it on par with physics concepts [11,12]. Despite the shared goal of addressing measurement uncertainty in physics labs, the depth, breadth, style, and instructional framing can vary significantly from one course to another [19].

The introductory physics lab course at the University of Colorado (CU) Boulder has recently undergone a

*alexandra.werth@colorado.edu

significant transformation with the explicit integration of a learning goal centered on the understanding of measurement uncertainties and the concept that repeated measurements yield a distribution with a mean and a standard deviation [20]. To emphasize this reasoning, the transformed course has eliminated verification labs [21] (e.g., measuring the gravitational constant, g) and instead focused on measurements where students could not predict the outcomes in advance [22]. Our objective in this study is to investigate the changes in student reasoning regarding measurement uncertainty within the context of this transformed course. For a comprehensive analysis of the impact of the course transformation on student reasoning about measurement uncertainty, as compared to previous iterations of the course, please refer to Ref. [22].

Several assessment tools have been utilized by physics educators and education researchers to evaluate student comprehension of measurement uncertainty [23–27]. In this study, we employ the Physics Measurement Questionnaire (PMQ) [25], a well-established tool developed at the University of Cape Town, South Africa, over a decade ago as part of a lab curriculum reform project [25]. The PMQ consists of nine probes, featuring both multiple-choice and open-ended responses [28]. Student responses to the open-ended questions are analyzed using the "point" and "set" paradigms [29]. The point paradigm reflects reasoning based solely on individual measurements, while the set paradigm, aligned with the learning goals of the CU Boulder lab course, signifies an understanding that all measurements possess uncertainties and do not represent the "true" value. Additionally, it recognizes that repeated measurements form a distribution with a mean and a standard deviation [29].

The PMQ was selected for this study due to its close alignment with the learning goals of the transformed course [30] and the absence of other research-based assessments explicitly focusing on these goals at the time of the study. However, one challenge in utilizing the PMQ is that student reasoning regarding measurement uncertainty is expressed through open-ended responses, necessitating intensive qualitative coding. Previous work by Wilson *et al.* [31] has explored the use of machine learning and natural language processing to address this coding challenge. While their work successfully demonstrated the application of the point and set paradigm-level codes, they did not delve into using machine learning to discern more nuanced distinctions in student reasoning (i.e., "subparadigm" reasoning).

Furthermore, the insights and challenges derived from using the PMQ to evaluate changes in student reasoning regarding measurement uncertainty in CU Boulder's introductory lab course inspired the development of a new assessment tool called SPRUCE [26,27]. The Survey of Physics Reasoning on Uncertainty Concepts in Experiments (SPRUCE) was recently created to measure similar, but also expanded, concepts relating to measurement uncertainty compared to the PMQ. Importantly, SPRUCE employs a format that eliminates the need for qualitative coding of responses.

Nonetheless, the qualitative coding of student responses to the PMQ using a subparadigm coding scheme [22] yields valuable insights into the evolution of student reasoning about measurement uncertainty within the introductory lab course at CU Boulder. This coding scheme allows us to match students' precourse and postcourse responses, cluster students based on their reasoning patterns, and investigate the changes that occur as a result of their engagement with the course. By employing this analysis scheme, we are able to construct a student-derived model that captures nuanced aspects of student reasoning, identifies common trends, and tracks shifts in their thinking. Specifically, we aim to address the following research questions:

RQ1: What nuances in student reasoning regarding measurement uncertainty are captured by a clustering algorithm using the subparadigm coding scheme?

RQ2: How does students' reasoning about measurement uncertainty change from precourse to postcourse?

To address these research questions, we employ hierarchical clustering analysis [32,33] using an agglomerative (i.e., "bottom-up") linkage criteria on the hand-coded open-ended student survey responses from the PMQ. This analysis uses a student-centered, bottom-up clustering approach to capture the wide range of student reasoning that might not neatly fit into predefined categories. Still, it enables us to identify common elements of reasoning within large clusters of students (*RQ1*) and gain insights into common changes in student reasoning resulting from the course (*RQ2*).

II. COURSE CONTEXT

CU Boulder Physics 1140: Experimental Physics 1 is a large enrollment, introductory physics lab course designed for engineering and physical science majors (Table I). Typically taken during the second semester of students' college education, it often serves as their first exposure to a college-level physics lab. The course is not directly tied to a full lecture course, but rather has six lectures on measurement and measurement uncertainty (described later in this section). The lab activities cover topics from mechanics, electricity and magnetism, and optics. Students meet weekly in two-hour lab sessions to work through a new lab activity each week for a total of 12 activities.

This version of the course was run for the first time in Fall 2017 [34] after being transformed to better align with the needs of students in the departments served by the course and emphasized the following learning goals:

1. Students' epistemology of experimental physics should align with expert views.
2. Students should have a positive attitude about the course.
3. Students should have a positive attitude about experimental physics.

TABLE I. Self-reported demographic data of CU Boulder PHYS 1140 Experimental Physics I students enrolled in the course in Spring 2018.

Gender	% of students
Woman	23.6
Man	75.1
Other gender	1.3
Major	**% of students**
Physics and eng. phys.	17.2
Other engineering	44.8
Math and other science	35.1
Other Disciplines	3.0
Race or ethnicity	**% of students**
American Indian or Alaskan Native	0.9
Asian	14.4
Black or African American	2.2
Hispanic/Latino	8.8
Native Hawaiian or Pacific Islander	0.7
White	69.0
Other race or ethnicity	4.0

4. Students should be able to make a presentation quality graph showing a model and data.
5. Students should demonstrate a setlike reasoning when evaluating measurements.

As part of the transformation, a new series of lab activities were developed with a focus on incorporating measurement uncertainty as a central element. More details on the course and the transformation process can be found in Refs. [20,22,30,34–37].

One key aspect of the transformed course is that each lab activity requires students to measure a quantity or outcome that they do not know beforehand [22]. This approach avoids "verification labs," [21] where students measure values that they could look up in textbooks or on the internet, and instead encourages students to actively engage in the process of measurement, prediction, comparison, and communication of results with their peers. This approach provides opportunities for students to consider and communicate both the value and the uncertainty of their measurements, and to discuss the context of their choices involving data collection and procedure.

In addition to the lab activities, the transformed course includes lectures that specifically focus on measurement uncertainty concepts [22]. Four out of the six lectures in the course are dedicated entirely to topics such as the importance of measurement uncertainty, estimating uncertainty from single and multiple measurements, standard deviation, and standard deviation of the mean (also called standard error), distributions and the normal distribution, making comparisons between measurements, and systematic errors in comparison to random uncertainties. These lectures aim to provide students with a solid foundation in understanding measurement uncertainty and its significance in experimental physics.

III. METHODOLOGY

A. The physics measurement questionnaire

Each of the nine probes in the PMQ [25] assesses a specific aspect of measurement, such as data collection, data processing, and data comparison. The probes are framed in the context of an experiment involving rolling a ball down a slope and measuring the distance it travels in free-fall. For example, one probe, for "different mean same spread" (DMSS), asks students to compare two sets of data and decide if the two groups had the same or different results. This probe is aptly called DMSS because the two groups' results gave different means, but had the same sample minimum and maximums. However, for the purposes of this study, only four out of the nine probes were considered for analysis, as the others were either incompatible with the electronic administration format, considered less useful by the researchers who developed the PMQ, or did not appear on both the pretest and post-test versions of the PMQ [30]. The four probes chosen for analysis are DMSS, along with "repeated distance" (RD), "using repeats" (UR), and "same mean different spread" (SMDS) [25].

The PMQ was administered in the course at CU at both the beginning (pre) and end (post) of the Spring 2018 semester. Participation in both the pre- and post-PMQ survey was a normal part of the course and students were awarded a small amount of credit for completing the survey [30]. The PMQ was completed by students outside of class using the Qualtics survey platform. The pre- and post-PMQ survey responses were matched for each student. Of the 722 students enrolled in the Spring 2018 course, 499 completed both the pretest and post-test, and were included in the dataset analyzed here.

B. PMQ coding scheme

The PMQ was initially developed by researchers in York, UK, for primary school students aged 9–16 [38]. It involved categorizing students' ideas about experimental data into eight levels representing a progression in their understanding of measurement concepts [38]. Researchers at the University of Cape Town later adapted the materials from York for their first-year university physics classes, creating the PMQ specifically for their context [29]. They extended the framework from York, introducing the point and set paradigms to analyze the open-ended responses of the PMQ [29].

Then in 2016, researchers at CU Boulder developed a modified coding scheme using responses collected from the pretransformed course. They initially started with the codebook from the Cape Town as a guide. However, this codebook did not align with many of the students' ideas in their responses and most students at CU Boulder did not change paradigms (Fig. 1) making it challenging to evaluate the PMQ at the paradigm level. Since the researchers aimed to capture the full range of student ideas, they created an emergent coding scheme that initially had around 100 codes per PMQ probe. This extensive codebook was able to

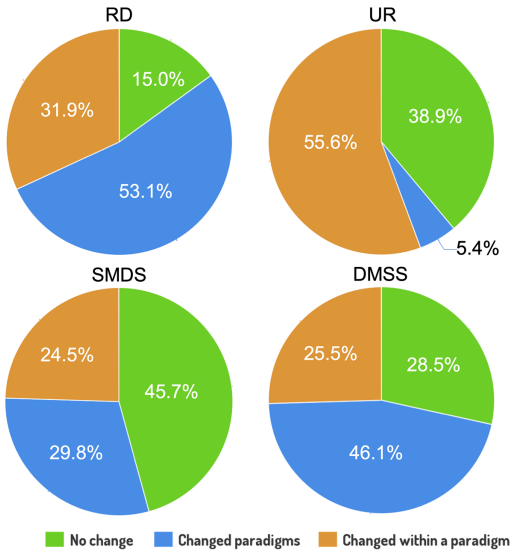

FIG. 1. Percentage of students from Spring 2018 for each probe who, from pre- to postassessment, did not change at all (green), changed paradigms (blue), and changed within a paradigm (orange). Most students did not change paradigms; however, when including changes within paradigms, in all cases, most students did change their reasoning from pre- to postassessment.

represent the many varied ways students discussed their reasoning around measurement uncertainty, many of which were not predicted by previous research or the research team. This emergent coding scheme was crucial to ensure that student ideas were not overlooked, which could have happened if only *a priori* codes were used. However, this nuanced codebook was too large to be of practical use and many codes were similar enough to be combined without losing the essence of student reasoning. The research team collectively assigned a paradigm (point, set, or neither or undefined) to each code based on the reasoning. They then consolidated and refined the codes, grouping them thematically to create a more manageable coding scheme.

The consolidated codes were then applied and refined using PMQ responses from Fall 2017. Interrater reliability was assessed using the Cohen's kappa statistic [39], with discussions and refinements made as necessary. This process was repeated for each of the four PMQ probes. The resulting codebook contained 12 to 16 codes in total per probe. The codes were given a two-character identifier, with the first character explicitly tying the code with a paradigm (set: S, point: P, undefined: U). Multiple codes could be assigned to a single response. A subset of the codebook is shown in Table II.

TABLE II. Selected codes from the new PMQ coding scheme. Reproduced from Ref. [22]. The full codebook can be found in Appendix.

Probe	Identifier	Name	Definition: "Argument is that…"
RD	S2	Measure an average	…multiple measurements will allow the experimenter to calculate an average or mean
RD	S4	Reduce uncertainty of mean	…multiple measurements will be used to reduce the error/uncertainty of the mean or average.
RD	P1	Measure the true value	…the experimenter could measure the correct value in a single measurement.
RD	U2	More data cancel out error	…experimenter needs to take more data to cancel or out-weigh the effect of error.
RD	U3	More data are better	…more data is better / more accurate / more precise / etc. Includes if reasoning other than 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	S4	Report average and spread	…experimenter should report the average and the uncertainty, range, or spread.
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, or range is better, more accurate more precise, etc. The response does not mention external factors, outliers, human error, etc.
SMDS	P1	The means are the same	…the groups agree because the means are the same.
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	P3	Means close enough, treats average as point	…the groups agree because the means are close enough

TABLE III. Percentage of students responses with setlike, pointlike, and undefined reasoning. Undefined responses refer to student responses that neither fit into set- nor pointlike reasoning.

Presurvey			
Probe	Set	Point	Undefined[a]
RD	42.7%	21.8%	35.5%
UR	91.0%	0.4%	8.0%
SMDS	51.5%	13.4%	26.7%
DMSS	33.5%	32.9%	33.7%
Postsurvey			
Probe	Set	Point	Undefined
RD	84.8%	1.0%	14.2%
UR	96.0%	0.0%	3.4%
SMDS	57.5%	9.2%	24.8%
DMSS	69.1%	7.4%	23.4%

[a]Some responses may be double or triple coded. In these cases, undefined refers to responses that are only undefined as well as those mixed with point- and setlike reasoning.

The final code definitions were then used to code responses from Spring 2017 and Spring 2018. The Spring 2018 data were used in the analysis for this study. A further description of the coding scheme development can be found in Refs. [22,30].

C. Clustering analysis of student responses

With a large number of student responses reflecting setlike reasoning in both the pre- and postsurvey, particularly in the UR probe (Table III), it is challenging to assert the influence of the course on student understanding of measurement uncertainty [30]. To better understand the nuances of student reasoning, we look at the subparadigm level; however, this presents a new challenge due to the high dimensionality of the data, with 12–16 codes per probe [22]. To address this issue, we employ a partitioning (or clustering) approach to group students with similar pre- or postsurvey reasoning on the PMQ probes. By analyzing student movement within these subgroups (i.e., from their pre- and postsurvey clusters), we can gain insights into student learning and reasoning surrounding measurement uncertainty while still maintaining the integrity of capturing as many individual reasoning elements and shifts in reasoning as possible.

We use hierarchical clustering, a popular and simple clustering method [32], which builds a hierarchy of groups. One advantage of hierarchical clustering is that it can handle categorical or mixed-type data effectively, since any valid measure of distance can be used for the algorithm [32,33]. To calculate a dissimilarity matrix of the student responses, we use Gower's distance [40], which measures the dissimilarity between two observations on a scale of 0 to 1 and does not require the observations to be numeric [40]. We calculated Gower's distance by computing all the pairwise dissimilarities between each student response for each probe in the dataset. Each student response was transformed into a vector, with each element of the vector corresponding to a potential code for that probe. If a student received a code, the element would be marked with a "Yes"; otherwise, it would be marked with a "No." The hierarchical clustering analysis then uses a "linkage criterion" to determine the distance between sets of observations based on pairwise distances between observations. In our analysis, we use Ward's method [33], which is designed to minimize the total within-cluster variance. Ward's method is an agglomerative or "bottom-up" approach, where each observation starts as a singleton cluster, and at each step merges the clusters such that there is a minimum increase in total within-cluster variance [33].

It is important to note that the choice of distance measure and linkage criteria can significantly impact the results of the clustering. Therefore, clustering should be viewed as an exploratory tool to visualize and interpret data, rather than considering the clusters as results in themselves. Additionally, the decision to use an agglomerative linkage criteria is a purposeful methodological choice to highlight emergent patterns from the student reasoning rather than prescribe an *a priori* framework. While this choice may risk missing a higher-level expert-like interpretation of the data, our fine-grained coding allows for a comprehensive examination of how students' reasoning evolves across a broad spectrum of ideas. And this approach may have uncovered more unexpected trends or patterns that a "top-down" clustering might overlook.

To both highlight and emphasize this, we describe our process for interpreting the output of the clustering algorithm via a dendrogram, our approach for determining the clusters used in our analysis, and how we match the pre- and postreasoning based on the clustering algorithm for the RD probe as an example.

1. Interpreting dendrograms

Figure 2 depicts a dendrogram consisting of clades, which are vertical lines connecting in a hierarchical treelike structure. The height of each clade reflects the "distance" between the clusters being connected. The leaves at the bottom of the dendrogram, shown in boxes A and B, represent individual data points, in this case, student responses to the PMQ survey.

Zooming in on the leaves in boxes A and B in Fig. 2, we can examine the responses of each student in the postsurvey. In Fig. 3, the largest cluster of students (33.3% of the respondents), stemming from the first clade, all

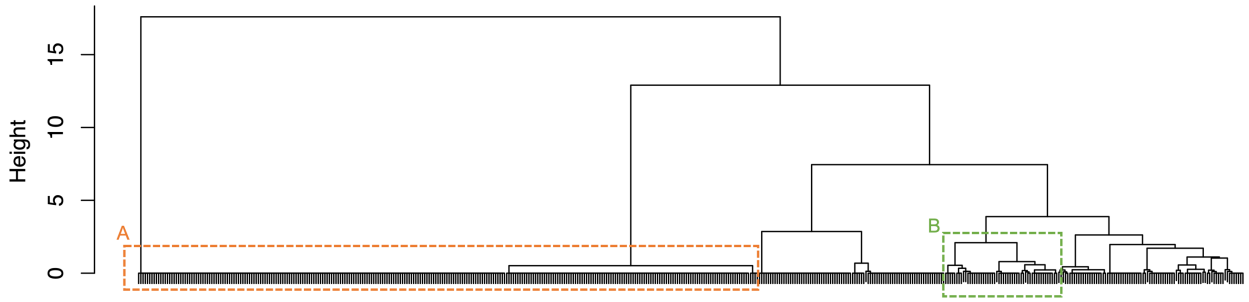

FIG. 2. The dendrogram from the hierarchically clustering of the RD probe of the PMQ postsurvey student responses. Figures 3 and 4 are magnified versions of boxes A (orange) and B (green).

FIG. 3. The dendrogram from Fig. 2 zoomed in on box A. We see a large cluster of students who responded with only S4 reasoning on the PQM and a smaller cluster that is comprised of students who responded with U3 reseasoning only and those who responded with U3 and S4 reasoning.

responded with S4 reasoning on the RD probe in their postsurveys. The second clade represents two distinct groups of students: 110 students who responded with U3 reasoning and an additional 4 students who responded with both U3 and S4 reasoning. The utilization of clustering algorithms is particularly advantageous in this context, as it allows us to effectively reduce the dimensionality of the dataset while still capturing the presence of the small population exhibiting a combination of U3 and S4 reasoning. These students are absorbed into a single cluster denoted as the "U3 cluster," which primarily exhibits U3 reasoning. Employing this approach enables meaningful comparisons and analyses of reasoning patterns within and between clusters.

Figure 4 shows a more complex clustering, combining multiple small groups of students. The clade on the left splits into five groups of student reasoning: 5 students with S2 and S3 reasoning, 3 students with S2 and P2 reasoning, 1 student with S2 and U1 reasoning, 1 student with S2 and U3 reasoning, and 13 students with S2 reasoning. Notably, all of these students stemming from this clade demonstrated S2-type reasoning. The clade on the right splits into seven groups of student reasonings: 1 student with S1 and P2 reasoning, 10 students with S1 reasoning, 2 students with

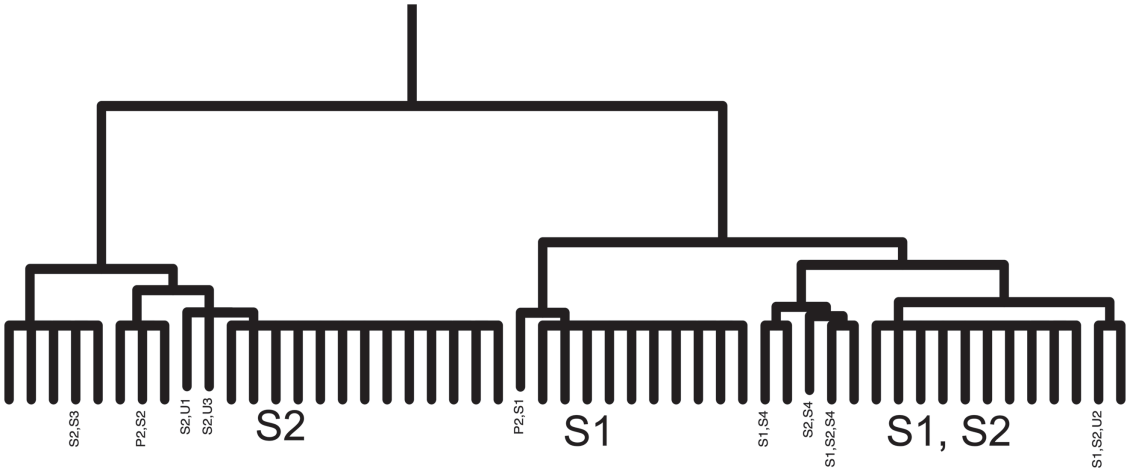

FIG. 4. The dendrogram from Fig. 2 enlarged on box B. We see an example of more complex clustering where clusters are combinations of many small groups of students with different reasoning.

TABLE IV. RD presurvey student reasoning divided into 10 clusters.

Cluster label	Number of students ($n = 499$)	Percentage of students	Cluster kept for analysis
U3	138	27.7%	Yes
S4	55	11.0%	Yes
P1	49	9.8%	Yes
P2	48	9.6%	Yes
S2 and U2	47	9.4%	Yes
Other	47	9.4%	Yes
S2	35	7.0%	Yes
U2	34	6.8%	Yes
S1	24	4.8%	No
S3	22	4.4%	No

TABLE V. RD postsurvey student reasoning divided into 10 clusters.

Cluster label	Number of students ($n = 499$)	Percentage of students	Cluster kept for analysis
S4	166	33.3%	Yes
U3	114	22.8%	Yes
U2	42	8.4%	Yes
S2 and U2	42	8.4%	Yes
Other	31	6.2%	Yes
S1	28	5.6%	No
S2	23	4.6%	Yes
S3	22	4.4%	No
U1	16	3.2%	No
P2	15	3.0%	Yes

S1 and S4 reasoning, 1 student with S2 and S4 reasoning, 2 students with S1 and S2 and S4 reasoning, 10 students with S1 and S2 reasoning, and 2 students with S1 and S2 and U2 reasoning. While each unique reasoning element on its own represents only a small portion of the class (2.6% of the class used S2-only reasoning), together, these two clades form a cluster representing students with "primarily S1 and/or S2 type reasoning," which constitutes 10% of the class. To access a comprehensive list of how each individual student code aligns with the selected clusters for every probe, please refer to the Supplemental Material [41].

2. Determining the clusters

In our study, we encountered the common challenge of determining the optimal number of clusters in a dataset when using clustering algorithms. However, there is no definitive way to partition clusters, regardless of the clustering algorithm used [42]. Often, researchers rely on subjective methods, such as inspecting elbow plots or dendrograms and determining what "looks" best. For instance, when examining Fig. 2, one may naturally see that the data would be best split into four clusters while another researcher might believe that is it best split into three. It is important to acknowledge that these methods are inherently subjective, and it may be more advisable to embrace the subjectivity and choose the number of clusters based on their interpretability, utility, interest, and underlying theory surrounding measurement uncertainty.

In our case, for the RD probe, we started by arbitrarily creating 10 clusters for both the pre- and postresponses, which is over double the number that was likely necessary. We then examined the reasoning used by each student in these clusters and labeled them based on the dominant reasoning elements (Tables IV and V). It is important to note that the clusters labeled as S1 and S2 in Table V represent the two clades shown in Fig. 4, and consist of many reasoning elements beyond just the S1 or S2 only codes.

Next, we chose which clusters to include in our analysis and which to combine into an "other" category based on their ability to help us answer our research questions. For example, when we examine general trends in student reasoning between the pre- and postcourse surveys, we focused on clusters that represented consistent reasoning from a large fraction of the class (approximately $> 10\%$), or those that were close to 10% of the class, but had mostly consistent reasoning amongst the students in the cluster (e.g., the U2 cluster representing only 8.4% of students in the presurvey and 6.8% in the postsurvey, but not combined with students with lots of mixed reasoning elements). Additionally, we chose to keep clusters that were "large" in the presurvey, but small in the postsurvey (e.g., P1 and P2 type reasoning in the RD probe) for comparison of pre and postresponses. However, we also aim to explore the subtleties of student reasoning around measurement uncertainty captured by the clustering algorithm and subparadigm coding scheme, so we also considered medium-sized clusters (5%–10% of students) that had interesting interpretability or theoretical implications (e.g., S2&U2 cluster representing only 9.4% of students in the presurvey and 8.4% in the postsurvey, but interestingly, these two codes were often double coded together).

After careful analysis and interpretation of the clusters, we decided to retain seven clusters for our analysis of the RD probe: the S4 cluster, the U3 cluster, the P1 cluster, the P2 cluster, the S2 and U2 cluster, the U2 cluster, and an *other* cluster.

3. Matching pre- and postsurvey reasoning

Once the clusters have been determined and labeled for the probe, we use a chord diagram to visually compare the pre- and postsurvey reasoning and observe course trends. A chord diagram is a graphical method of displaying flows of data within a circle. Nodes representing each of the clusters are placed around the circumference of the circle. Arcs are drawn into or out of each node, with the thickness of the arcs representing the number of students who changed into

TABLE VI. Ten student reasonings to the RD probe on the pre- and postsurvey and their assigned cluster.

Student	Precode	Precluster	Postcode	Postcluster
Student 1	S1	Other	U2	U2
Student 2	U3	U3	U3	U3
Student 3	S2, U2	S2 and U2	U2	U2
Student 4	P1	P1	U3	U3
Student 5	P2, P3	P2	S4	S4
Student 6	S2, U2	S2 and U2	U3, S4	U3
Student 7	P2, S4	P2	U3	U3
Student 8	S1, S2, S4	Other	S1, S2	Other
Student 9	P1	P1	S2, U2	S2 and U2
Student 10	P2, S2	P2	S2, U2	S2 and U2

or out of that type of reasoning. The "mounds" in the diagram represent the number of students who remained in that particular cluster from the pre- to postsurvey.

To illustrate, Table VI provides an example of how ten actual student codes from the pre- and post-RD probe correspond to cluster names and are matched. Additionally, Table VII presents the total number of students who moved between or stayed in each of the clusters from the pre- to postsurvey, which serves as the raw data used to create the chord diagrams.

IV. THE RD PROBE

The RD probe asks students whether they should measure the distance that the ball lands a few more times, one more time, or no more times after they had already measured the distance once before. The students are then prompted to, "Explain your choice" (Fig. 5).

A variety of reasoning elements emerged among students as they explained their choices (see Appendix, Table XII). Students using pointlike reasoning presented diverse arguments, such as the potential for the experimenter to capture the accurate value in a single measurement (P1). Some students, recognizing the necessity to identify outliers,

TABLE VII. Matched clusters from pre to postsurvey with student counts. Presurvey reasoning is represented vertically in the first column, while postsurvey reasoning is represented horizontally in the first row.

Clusters	Other	P1	P2	S2	S2 and U2	S4	U2	U3
Other	27	3	2	2	6	29	6	18
P1	7	1	1	4	5	14	4	13
P2	12	2	1	1	4	13	5	10
S2	6	0	0	5	5	8	4	7
S2 and U2	7	0	0	3	6	13	8	10
S4	5	0	2	2	6	29	3	8
U2	3	0	1	1	3	17	2	7
U3	20	3	8	5	8	43	10	41

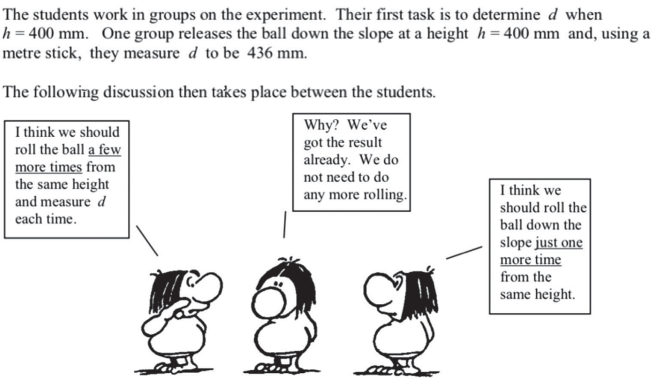

FIG. 5. The RD probe of the PMQ. Students are asked to choose "With whom do you most closely agree?" then they are then prompted to "Explain your choice." Reproduced from [28].

emphasized the requirement for repeated measurements to discern mistakes or outliers (P2). This perspective involves the condition that the experimenter must obtain consistent results at least twice for that number to be deemed correct. In consideration of available time or resources, some students highlighted that taking a single measurement was a better course of action based on such practical constraints (P3). Furthermore, students indicated that practice was essential during measurements to address errors and external factors (P4).

Students employing setlike reasoning all acknowledged the need for multiple measurements. Certain students advocated for this without delving into statistical terminology, stating that utilizing all measurements collectively would improve the accuracy and precision (S3). Others focused solely on measuring spread (S1) or average (S2), but did not discuss both. The most sophisticated set-based reasoning (S4) came from students who asserted that using multiple measurements would reduce the uncertainty surrounding the mean value.

Additionally, a subset of students offered reasoning that did not align within either a setlike or pointlike paradigm. These students emphasized acquiring more data without explicit statistical justification (U1), posited that additional data would counteract errors (U2), or simply asserted the benefits of more data (U3).

It is important to note that students occasionally integrated multiple lines of reasoning in their explanations. For instance, as will be discussed later in this section, students frequently combined perspectives like S2 and U2, stating that "… multiple measurements would enable the experimenter to compute an average, thereby mitigating the impact of errors."

We begin the analysis of the RD probe by interpreting the chord diagram shown in Fig. 6, discussing dominating clusters of these subparadigm codes (*RQ1*) and trends in student movement in reasoning from the presurvey to the postsurvey (*RQ2*). Last, we look at a particularly interesting cluster of double-coded responses that

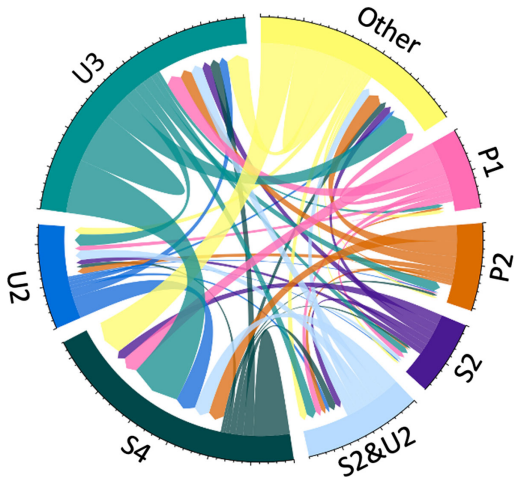

FIG. 6. Chord diagram representing changes in student reasoning between the pre- and postsurvey on the RD probe.

represented the reasoning of almost 10% of the students in the course (*RQ1*).

A. Interpretation of the chord diagram

Analysis of the chord diagram shown in Fig. 6 reveals that the two primary reasoning elements students used were U3 and S4. Combined, S4 and U3 clusters account for 38.7% of the student reasoning on the presurvey and 56.1% on the postsurvey. This indicates that students shifted from a more diverse range of responses in the presurvey to primarily S4 and U3 reasoning in the postsurvey, suggesting changes in reasoning around measurement uncertainty. A summary of the key takeaways from the analysis of the RD probe is shown in Table VIII.

One of the most prominent trends observed in student reasoning was the movement towards S4 clusters from all other clusters, with students who initially exhibited S4 reasoning continuing to demonstrate S4 reasoning in the postcourse assessment [Fig. 7(a)]. Similarly, there was very little movement away from S4 reasoning [Fig. 7(b)]. The code S4 represents students who explained their choice by stating that they need to "reduce the uncertainty of the mean." This type of reasoning aligns closely with the learning goals of our course, as it recognizes that all measurements have associated uncertainties and that reducing uncertainty requires considering a set of repeated measurements.

Another interesting trend is that there were comparable numbers of students who moved into, out of, and stayed in the U3 cluster (Fig. 8), which is not surprising considering that U3 represents students who responded with the belief that "more data is better." This reasoning does not fit into either the set- or pointlike paradigms, as students do not provide additional information on *why* they believe more data is better. The brevity of this response could be attributed to a lack of deeper understanding of measurement uncertainty, lack of interest in filling out the PMQ survey, or limited time to complete the survey.

Furthermore, it was observed that many students moved "out" of the P1 and P2 clusters (Fig. 9) and into many other reasoning clusters.

B. A notable cluster

One cluster in the RD probe (Fig. 6) showed students double coded with S2 and U2, representing approximately

TABLE VIII. Key takeaways from the analysis of the RD probe.

Movement towards expertlike reasoning: Students in the postsurvey were more likely to use reasoning that discussed multiple measurements being needed to reduce the error or uncertainty of the mean or average (S4). This reasoning aligns with the course goals and was the most expertlike reasoning demonstrated by the students because, unlike the other setlike responses, students discussed the importance of both the mean and the uncertainty.
Comparable numbers of students responded by saying "more data is better" in the pre- and postsurveys: While many students who began in the U3 cluster, stating that "…more data is better, more accurate, more precise, etc.," moved out of this cluster, approximately equal number of students moved into this cluster. This speaks to the importance of the cluster analysis on individual student matched responses rather than the cumulative analysis of the class-wide data which may have missed this movement. This cluster is neither in the set- nor pointlike paradigms because the students do not use statistical reasoning to explain their answer and it is particularly marked by brevity of the response.
Movement out of pointlike reasoning: Many no longer used reasoning such as "…the experimenter could measure the correct value in a single measurement" (P1) nor "…repeated measurements are needed in order to know which measurements were mistakes or outliers, after all measurements are taken" (P2). These pointlike reasonings were used by almost 20% of the students in the presurvey, but less than 5% in the postsurvey. Furthermore, there was no clear differences as the movement from students who started in the P1 or P2 clusters and these students moved into setlike, unknown, or other types of reasoning about evenly.
Students using setlike language without expertlike set reasoning: One cluster in the RD probe showed students double coded with S2 and U2, representing approximately 9% of the students in the postsurvey. Students in this cluster stated reasoning such as, "multiple measurements will allow the experimenter to calculate an average or mean which will cancel or outweigh the effect of error." The double-coded S2 and U2 response indicates that, perhaps, students are using setlike language but do not exhibit expertlike set reasoning.

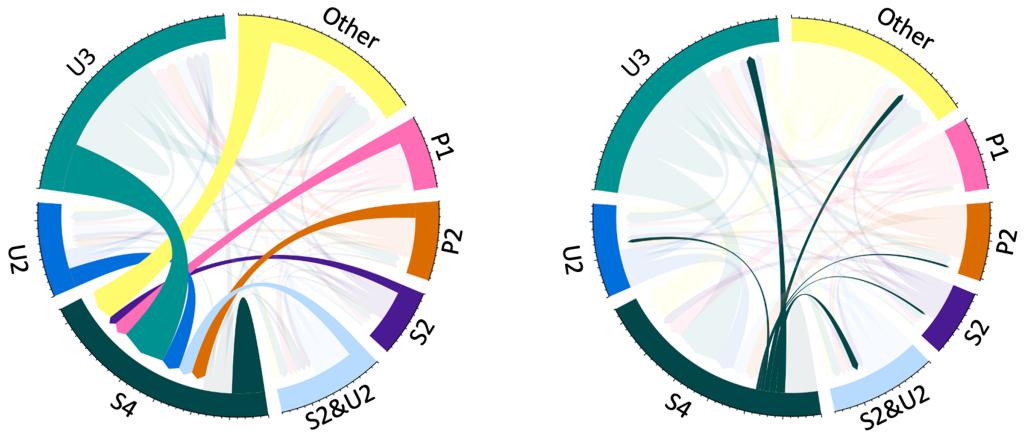

FIG. 7. Student reasoning movement (a) into and (b) out of the S4 cluster for the RD probe.

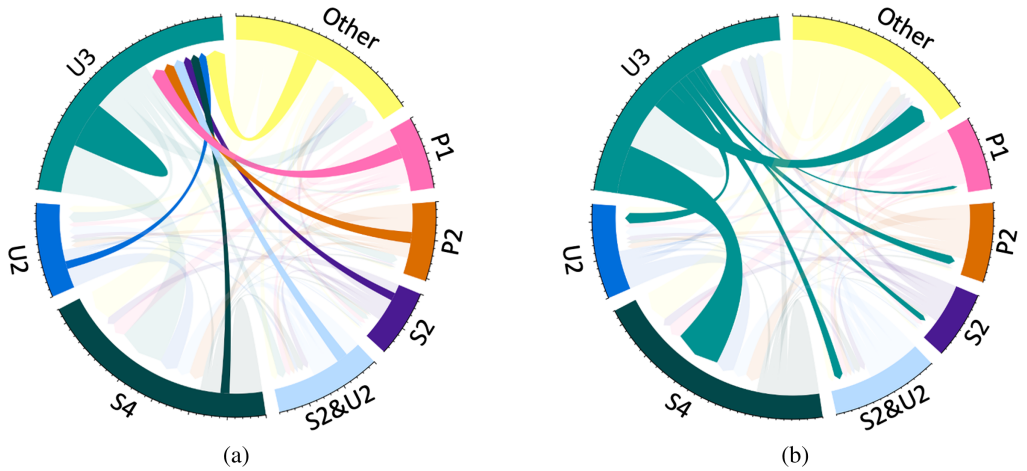

FIG. 8. Student reasoning movement (a) into and (b) out of the U3 cluster for the RD probe.

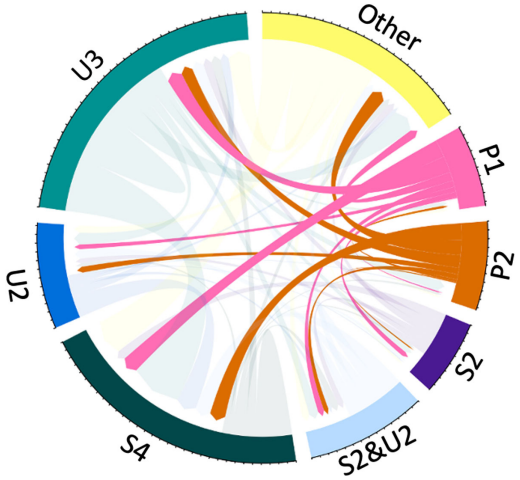

FIG. 9. Student reasoning movement out of the P1 and P2 clusters for the RD probe.

9% of the students in the postsurvey. This finding highlights an interesting limitation of the point or set paradigm. The combined S2 and U2 codes suggest that "multiple measurements will allow the experimenter to calculate an average or mean which will cancel or outweigh the effect of error." Examples of actual student responses that were double coded include:

> In order to get a more accurate result, it is important to take multiple data points and average them out.
> Increase sample size, then average values for a more consistent observation.

The double-coded S2 and U2 response indicates that students are using setlike language (e.g., "multiple data points," "sample size," "average") but do not exhibit expertlike set reasoning (e.g., stating that taking an average "cancels out" error). One possible explanation for this observation is that students might be repeating setlike terms

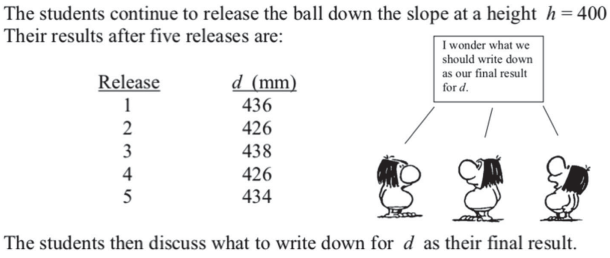

FIG. 10. The UR probe of the PMQ. Students are asked to "Write down what you think the students should record as their final result for d." Then students are prompted to "Explain your choice." Reproduced from [28].

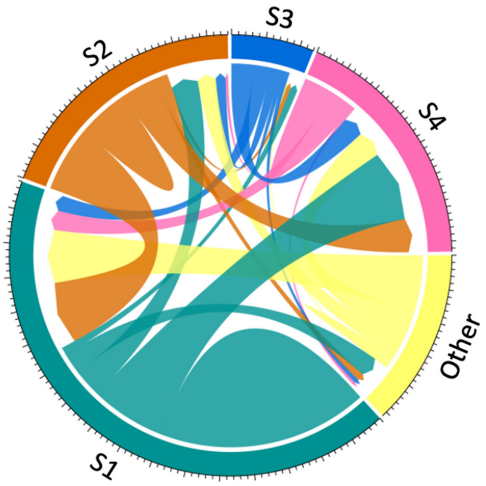

FIG. 11. Chord diagram representing changes in student reasoning between the pre- and postsurvey on the UR probe.

used by the instructional team without deeply grasping their meaning or not adopting a genuine setlike perspective. They may simply *remember* [43] that taking an average or mean is important but may not fully *understand* or *evaluate* [43] why it is necessary within this particular context.

V. THE UR PROBE

Figure 10 shows the UR probe, which asks students to report a value of the ball distance after five trials were taken, followed by an explanation of their choice.

Students explained their choices for the UR probe using a range of reasoning elements (see Appendix, Table XIII). Students using pointlike reasoning responded with various ideas, such as choosing a single value to report (P1) or using the average as a last resort (P2).

However, most students responded with setlike reasoning where they specified reporting the average would be the best option. However, some did this tersely saying simply "average" (S1), others discussed why the average was useful (S2) or appropriate (S3). A few students talked about how to mathematically compute the average (S5). However, the most expertlike response came from students who specified that the experimenter should report both the mean and the spread (S4).

Here, we discuss the dominant clusters for the UR probe shown in Fig. 11 (*RQ1*). We then look at trends in student movement in reasoning from the presurvey to the postsurvey (*RQ2*), particularly the transition from S3 to S4 reasoning. A summary of the key takeaways from the analysis of the UR probe can be found in Table IX.

TABLE IX. Key takeaways from the analysis of the UR probe.

The vast majority of students stated simply "average" or named reported value as average: The dominant cluster in both the presurvey and the postsurvey was students who only state things like "I averaged," "do the average," "average is best," or "it is the average," but does not elaborate along the lines as described in the other setlike codes (S1). This cluster also includes students who reported numerical value of the average (S1). This represents the most rudimentary of the setlike reasonings, suggesting that many students were not yet familiar with more advanced concepts of measurement uncertainty or did not feel it was necessary to report the uncertainty. Furthermore, students in the S1 reasoning cluster tended to stay with S1 reasoning (61.6%). One possible explanation is that the terseness of S1 responses may have made it difficult for students to provide more detailed explanations of their reasoning.

Other set-like reasoning: All of the major clusters from the UR probe were variations of set-like reasoning, indicating the utility of the subparadigm coding scheme. The second largest cluster in the presurvey and the third largest cluster in the postsurvey were students who discussed why the average is useful (S2), which is a more advanced type of setlike reasoning that considers the spread of the data. However, students in this category still only discuss the mean. S4 reasoning, where students reported both the mean and the spread, is considered the most expertlike response and was the second largest cluster in the postsurvey. This suggests that some students were beginning to understand the importance of reporting uncertainty along with the mean.

Movement toward reporting both the mean and spread: The movement toward the S4 reasoning cluster where students explained that "…experimenter should report the average and the uncertainty, range, or spread" showed a positive outcome for the transformed course. We see that only 7.6% of students began in this cluster initially, yet 23.6% of student moved into this cluster in the postsurvey. We saw that students from the S3 cluster, who said that reporting the average is best because all data matters or because the spread of the data is small were more likely to move into the into the S4 cluster (38.6%). Whereas about 20% of students from the S1 and S2 clusters and about 24% of students from the other reasoning cluster moved into the S4 cluster. Overall, this trend suggests that the course was effective in helping students develop a more sophisticated understanding of measurement uncertainty. However, more work is needed to ensure that all students are able to reach this level of understanding since students who responded with simply "average" was still the dominant cluster.

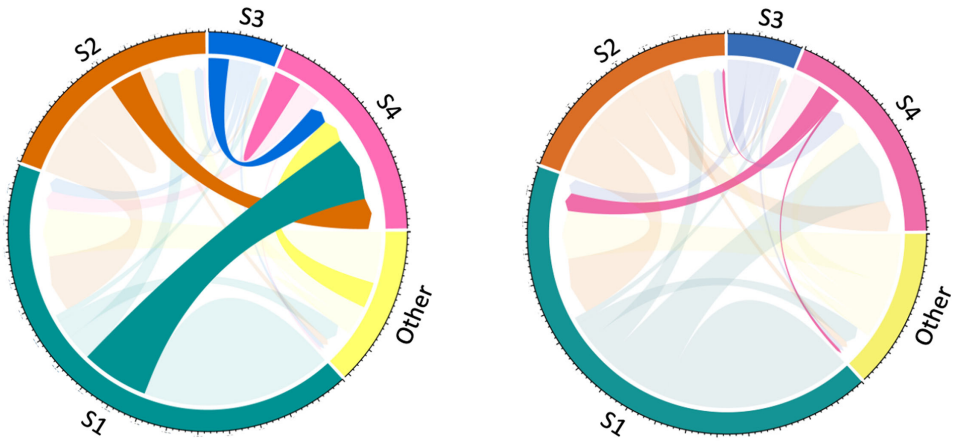

FIG. 12. Student reasoning movement (a) into and (b) out of the S4 cluster of the UR probe.

A. Interpretation of the chord diagram

The dominant cluster in both the pre and postsurvey was students who responded with S1 only reasoning. The S1 code was assigned to student responses that stated "I averaged," "do the average," or "average is best," but did not provide any further explanation (i.e., not co-coded with S2, S3, S4, or S6, see code definitions in Appendix). The S1 code may have been double coded with S5 (when the method of averaging is explained as additional information) or with S7 when students wrote that taking the average was the "logical" or "correct" thing to do, as these also do not provide further insight into set paradigm reasoning. However, such double coding with S5 or S7 was rare and categorized into the "other" cluster. The S1-only cluster represented 45.9% of students in the presurvey and 49.7% in the postsurvey.

The second largest cluster in the presurvey (22.2% of students) and the third largest cluster in the postsurvey (15.2% of students) were students who responded with S2 reasoning, describing "why the average is useful." For example, students may have written that reporting the average is best because it accounts for fluctuations or errors, or because it predicts future measurements.

S4 reasoning, where students reported both the average and the spread, was the second largest cluster in the postsurvey, consisting of 22.4% of the students. However, it only represented 7.6% of the students in the presurvey.

B. Trends in student reasoning

The analysis of the survey responses revealed interesting trends in students' reasoning about measurement uncertainty. Together, the four set paradigm clusters shown in the chord diagram represent a large majority of the class—83.6% of the students in the presurvey and 89.8% in the postsurvey. Among these clusters, the S4 reasoning cluster, which is considered most aligned with expert reasoning and characterized by the use of standard deviation and reporting uncertainty, showed a positive outcome for the transformed course, as many students moved into this cluster from the presurvey to the postsurvey, and very few students changed out of this type of reasoning (Fig. 12).

In contrast, many students moved out of the S3 reasoning cluster on the postsurvey. Specifically, only 2% of students ended up in the S3 cluster in the postsurvey, compared to 10% of students who started in the S3 reasoning cluster in the presurvey. Moreover, among the students who started in the S3 cluster in the presurvey, 29% moved to the S1 cluster, 18% moved to the S2 cluster, and 38% moved to the S4 cluster in the postsurvey. These findings suggest that starting with S3 reasoning in the presurvey is more likely to contribute to the development of expertlike reasoning (S4) in the postsurvey. However, the underlying reasons for this trend require further investigation. To explore this idea, we examine specific quotes from student responses.

C. Transition from S3 reasoning to S4 reasoning

The S3 code represents students who said that reporting the average is best because all data matters, or because the spread of the data is small enough. This code includes reporting all data as well as the average, but does not include students who wrote "it is the correct thing to do."

The transition from S3 to S4 reasoning is demonstrated by various student responses. For example, a student initially classified as S3 in the presurvey expressed support for using the mean in their response while also acknowledging the potential use of the median:

> I chose to use the mean of the results. We could also use the median, which would be 434 mm in this situation, but I feel that doesn't accurately represent the weight of the multiple 426 mm results. One could argue that using the median could reduce the impact of a significant

> outlier, but in these results nothing is over 3% from the mean.

However, in the postsurvey, this student shifted to S4 type reasoning and emphasized the importance of reporting the mean along with the standard deviation of the mean as a measure of uncertainty:

> They should report the mean, which would be 432. They should also display their uncertainty, which they could do by reporting the standard deviation of the mean.

Likewise, another student who transitioned from S3 reasoning in the presurvey to S4 reasoning in the postsurvey argued in the presurvey to use the mean because there were no outliers in the data:

> There are no clear outliers therefore all measurements must be assumed to be equally valid. Thus an average must be taken.

In the postsurvey, this student reiterated the use of the mean and suggested reporting the standard deviation of the mean to represent uncertainty:

> The mean is 432, and the uncertainty is the standard deviation of the mean.

These responses illustrate the similarly thoughtful justifications provided by students in both the S3 and S4 reasoning clusters. In contrast, we see that many students in the S1 reasoning tended to stay with S1 reasoning. Unlike S3 or S4 reasoning, S1 reasoning provides terse responses such as "take the average" or "average," which lack detailed insight into their reasoning about measurement uncertainty. These brief responses from the S1 cluster offer limited information regarding student reasoning about measurement uncertainty, leaving questions about whether students are rushing through the questionnaire without much thought or if they have not considered reporting the standard deviation.

VI. THE SMDS PROBE

The SMDS probe, as shown in Fig. 13, was used to ask students to compare two groups of students who had both collected five data points with the same averages.

In response to the SMDS probe, students commonly used both point- and setlike reasoning (see Appendix, Table XIV). Students using pointlike reasoning presented diverse arguments, such as the treating the mean as a single value instead of a statistical tool and arguing that the means were the same for the two groups (P1). Additionally, some pointlike reasoning specified that the fact that the spreads or individual trials are different does not matter, including responses that focus on agreement of the averages while providing a reason for why the sets are different (P2).

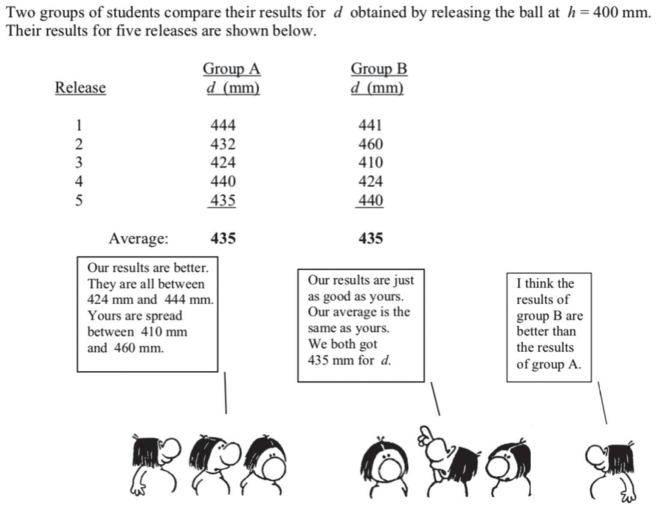

FIG. 13. The SMDS probe of the PMQ. Students are asked to choose "With which group do you most closely agree?" then they are then prompted to "Explain your choice".

Others said that one dataset was better because that group had fewer outliers with no mention of the spread (P3) or that the experimenters were more careful during the experiment (P4).

Students using setlike reasoning all said that the dataset with a smaller spread was better (group A). Some did this by simply saying that group A was better, more accurate, or more precise (S1). Others said that group A was better because the spread was smaller, but mentioned human error might be a factor in this outcome (S3). The most expertlike response came from students who indicated that smaller spread is better, but did not mention external factors (S2).

Here, we interpret the chord diagram for the SMDS probe shown in Fig. 14 and discuss the two dominant clusters (*RQ1*) as well as the movement between them from the presurvey to the postsurvey (*RQ2*). A summary of the key takeaways is shown in Table X.

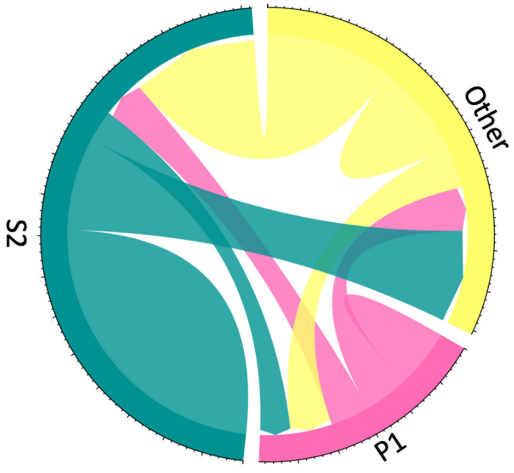

FIG. 14. Chord diagram representing changes in student reasoning between the pre and postsurvey on the SMDS probe.

TABLE X. Key takeaways from the analysis of the SMDS probe.

Two dominant clusters: Over half of the students responded with to the SMDS probe by saying the "smaller spread is better" without mentioning of external factors (S2) and approximately one-sixth of the class responded by saying that the "the groups agree because the means are the same" (P1) in both the pre and postsurvey. While these two clusters represented over half the class, there were many other diverse types of reasoning used by students (29.2% in the presurvey and 25.6% in the postsurvey) that did not merge into distinct reasoning clusters.
Despite the stability of the number of students in the two clusters pre and post, many students changed their reasoning: While the percentages of students in the two dominant clusters remained relatively stable from the pre to postsurvey, we observed significant shifts in student reasoning within these clusters.
Transition from point to set reasoning: Students who began with pointlike reasoning in the P1 cluster were more likely to transition to setlike reasoning in the S2 cluster after the course (27.8%) than vice versa (0.07%). Perhaps students began to recognize the importance of data consistency, spread, and uncertainty in evaluating the quality of experimental results—one of the learning goals for the course. However, 41.8% of the students remained in the P1 cluster indicating that more work is needed to ensure that all students are able to reach this level of understanding.
Students who stated that the "smaller spread is better" in the presurvey tended to also simply say that the smaller spread is better in the postsurvey: Students who initially started in the S2 cluster, stating that group A is better because smaller spread is better, predominately remained in the S2 cluster (70.3%). However, approximately one-third of students did move out of the S2 cluster, with 7.3% moving into P1 reasoning and 22.4% moving into other types of reasoning.

A. Interpretation of the chord diagram

Two dominant clusters, S2 and P1 reasoning clusters, were identified in both the pre- and postsurvey responses.

Over half of the students (50.8% in the presurvey and 56.0% in the postsurvey) responded with S2-like reasoning, stating that a "smaller spread is better" without mentioning any external factors, outliers, or human error. For example, one student wrote:

> While they did end up with the same final answer, group A's data has a smaller standard deviation and gives us more confidence that their answer is correct.

S2 reasoning aligns closely with the learning goals for the course.

Approximately one-sixth of the class (17.3% in the presurvey and 16.0% in the postsurvey) responded with P1-like reasoning, stating that "the groups agree because the means are the same" without mentioning the spread. Examples of students using P1 reasoning include

> All distances will be different so I agree with student B becasue [sic] they both have the same averages.

and

> Both datasets are just as meaningful.

B. Trends in student reasoning

While the percentages of students in the S2 and P1 clusters remained relatively stable from the pre to postsurvey, we observed prominent shifts in student reasoning (see Fig. 14). This finding highlights a strength of the clustering method employed in this study, as it allowed us to identify specific changes in individual students' reasoning that may not have been apparent from examining course-wide trends alone. Interestingly, we found that only 41.8% of the students who initially belonged to the P1 cluster remained in that cluster after the course. Approximately 30% of students who initially used P1 reasoning transitioned to S2 reasoning, indicating a shift towards recognizing the importance of considering spread of the data. Additionally, another 30% of students moved to other types of reasoning, indicating a diverse range of perspectives that emerged throughout the course. Within the other category, 15.2% of students adopted P2 reasoning, indicating that they explicitly believed the differences in spreads for individual trials did not matter.

To illustrate the transition from P1 to S2 reasoning, we provide two examples of student responses.

In the presurvey, the first student wrote

> Results can't be better. It is merely based on what their tests show.

This statement suggests that the student believed the quality of the results depended solely on the outcomes of the tests conducted by each group. The student did not consider the spread or consistency of the data as relevant factors in determining the quality of the results.

In the postsurvey, the same student stated

> I think their data is better because it is less spread out and more consistent.

Here, we see a clear shift in the student's reasoning. The student now recognizes the importance of data consistency

and reduced spread. They acknowledge that data with less variability and a smaller spread is preferable, indicating an understanding of the significance of reliable and precise measurements.

Another student expressed in the presurvey

> Ultimately, both groups of students reached the same conclusion. The idea of "better" results is quite subjective in this instance.

In this response, the student suggests that, within the context of measurement uncertainty, the notion of better results is subjective and all that matters is the data average.

However, in the postsurvey, this same student's perspective has become more sophisticated:

> The group with the smaller spread likely has the smaller standard deviation, which tends to indicate better or clearer results.

They now emphasized that the group with a smaller spread, which suggests a smaller standard deviation, likely possesses better or clearer results. This shift in reasoning suggests a growing recognition of the relationship between variability and the quality of results, reflecting a more refined understanding of measurement uncertainty and its impact on data interpretation.

These examples demonstrate the transition from P1 reasoning, where the focus is primarily on the test outcomes, to S2 reasoning, which emphasizes the significance of data consistency, spread, and uncertainty in evaluating the quality of experimental results. It showcases how students' perspectives can evolve as they engage with the course material and develop a deeper understanding of measurement concepts.

In contrast to the students who began with P1 reasoning, students who initially started in the S2 cluster predominately remained in the S2 cluster (70.3%). However, approximately 30% of students did move out of the S2 cluster, with 7.3% moving into P1 reasoning and 22.4% moving into other types of reasoning.

VII. THE DMSS PROBE

The DMSS probe, shown in Fig. 15, required students to compare two groups of students who had collected five data points with different averages, but with the same spread.

Various aspects of reasoning came to light as students described their decisions to the DMSS probe (see Appendix, Table XV). Students using pointlike reasoning often treated the mean and spreads as single values rather than a statistical tools saying that the means and spreads must both match (P1), the means must match (P2), or that the means are "close enough" (P3). Some specified that the two groups do not agree (P4) or agree (P5) after doing a point-by-point comparison.

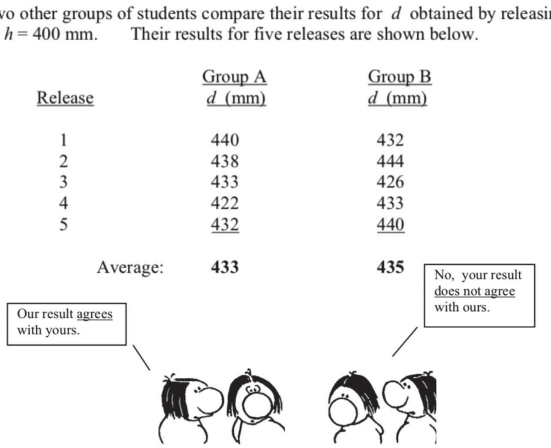

FIG. 15. The DMSS probe of the PMQ. Students are asked to choose "With which group do you most closely agree?" then they are then prompted to "Explain your choice".

Students employing setlike reasoning all also discussed the means, but treated them as a statistical tool and often talked about statistical variance or spread. For example, some students said that the means are close enough and talked about statistical variation in general (S1). Others said that the two groups had similar means and spreads, but did not mention of overlap (S2). The most sophisticated set-based reasoning (S3) came from students who discussed the means, spreads, and the overlap.

Additionally, a subset of students offered reasoning that did not align within either a setlike or pointlike paradigms. These students gave nonstatistical reasoning such as systematics (U1) or stated that they could not calculate the uncertainty or the spread (U2).

We interpret the chord diagram for the DMSS probe shown in Fig. 16 and discuss the dominant clusters (*RQ1*).

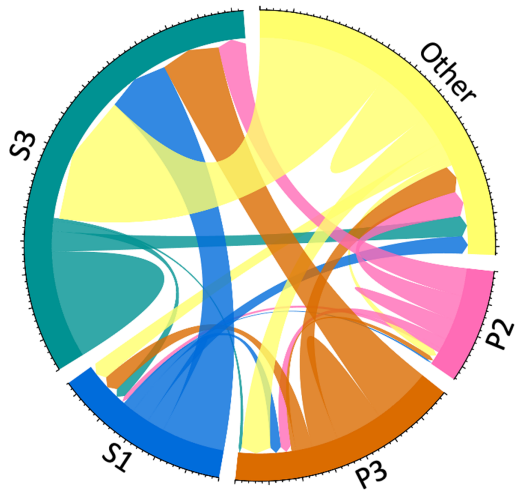

FIG. 16. Chord diagram representing changes in student reasoning between the pre and postsurvey on the DMSS probe.

TABLE XI. Key takeaways from the analysis of the DMSS probe.

Dominant cluster where students mention the similar means and spreads and the overlap of the means and spreads: The postsurvey results of the DMSS probe were dominated by a single cluster, representing 52.9% of the class, where students stated "…the groups agree because the means and spreads are similar" and their argument considers the overlap between the means and/or spreads of the two datasets (S3). This cluster represents the most expertlike reasoning.

Movement into the most expertlike cluster: Notably, 57.3% of students who initially reasoned that means were close enough and discussed statistical variation in general (S1) transitioned to the more expertlike S3 cluster. In contrast, only 38.5% of students who started with pointlike reasoning stating that the "..the groups agree because the means are close enough," in other words, treating the mean as a point, (P3) moved to the most expertlike S3 reasoning cluster. While this is still a positive shift that suggests that increased exposure to measurement uncertainty concepts in the course influenced students to adopt more expertlike reasoning strategies in their postsurvey responses, more work is still needed to make this an equitable shift for all the students.

Use of statistical reasoning: Students who used overlaps and statistical reasoning to support their choices in the presurvey, instead of generically referencing measurement error to justify their reasoning, adopted more expertlike reasoning strategies in their postsurvey responses.

Additionally, we discuss the trends in movement in reasoning about measurement uncertainty from the presurvey to the postsurvey (*RQ2*). A summary of the key takeaways is shown in Table XI.

A. Interpretation of chord diagram

The postsurvey results of the DMSS probe were dominated by a single cluster, representing 52.9% of the class, which exhibited S3 reasoning. Students with S3 reasoning argued that the groups agree because the means *and* spreads are similar, considering the overlap between the means and spreads of the two datasets. The second largest postsurvey cluster, comprising only 13.6% of the class, consisted of students with P3 reasoning. These students argued that the "means are close enough" and treated the average as a single point.

These results contrasted with the clusters found in the presurvey, where only 19.2% of the class fell into the S3 cluster and 20.8% were in the P3 cluster. Additionally, 15.0% of students exhibited S1 reasoning in the presurvey, while 12.0% exhibited P2 reasoning. P2 reasoning represents students who argued that the groups do not agree because the "means are not the same" without mentioning spread. S1 reasoning is similar to P3 reasoning in that students argued that the groups agree because the averages are close enough. However, S1 reasoning also incorporated a discussion of statistical variation in general. For example, a student with P3 reasoning wrote

> The results are close enough to agree with each other.

In contrast, a student with S1 reasoning wrote:

> There is probably some form of uncertainty in the measurements causing them to be more or less the same.

B. Trends in student reasoning

When examining the movement from the presurvey clusters to the postsurvey clusters (Fig. 16), a significant trend emerges, with students from all other reasoning clusters shifting towards S3 reasoning. This is an encouraging finding, as S3 reasoning is more closely aligned with expertlike views.

Notably, 57.3% of students who initially belonged to the S1 cluster transitioned to the S3 cluster [Fig. 17(b)]. In contrast, among students who started in the P3 cluster, only 38.5% moved to the S3 cluster, while 30.8% remained in the P3 cluster (Fig. 17). Although the logic of students in the P3 and S1 clusters is similar, with both clusters positing that the averages are close enough, it seems that the consideration of statistical variation, in general, led students to adopt S3-type reasoning. It is possible that students who adhered to the simplicity of P3 reasoning lack interest in completing the survey or had limited time to provide further explanation of their choices. On the other hand, students who took the time to discuss the statistical variation in the presurvey were, perhaps, more invested in the PMQ, but did not yet possess the content knowledge pertaining to measurement uncertainty to respond with S3-type reasoning.

For instance, one student responded with S1 reasoning in the presurvey:

> Since the two group's values are close this deviance could have occurred from error.

And the same student used S4 reasoning in the postsurvey:

> There is a fair amount of overlap between the two groups, based upon their averages and standard deviations, so I think that the groups data agree.

Clearly, this student was engaged with both the pre- and postsurvey. However, in the postsurvey, they employed more expertlike language to describe the pattern they had identified in the presurvey.

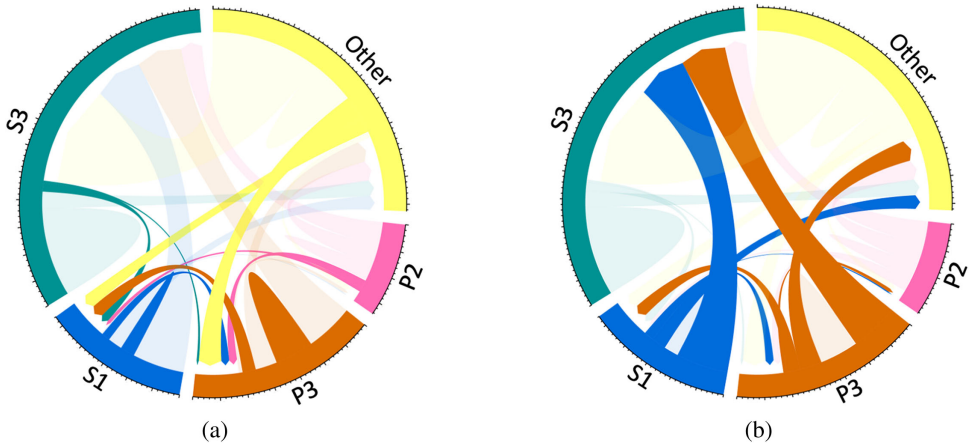

FIG. 17. Student reasoning movement (a) into and (b) out of the S1 and P3 clusters of the DMSS probe.

Another student who moved from the S1 to S3 cluster expressed a similar shift in their survey responses. In the presurvey, they stated

> There is some inconsistency but there could have been small measurement errors making them different.

And in the postsurvey, they said

> The upper bounds overlap with the lower bounds of the two measurements so the data seams [sic] to agree.

Both of these students used overlaps and statistical reasoning to support their choices, instead of generically referencing measurement error to justify their reasoning. This suggests that increased exposure to measurement uncertainty concepts in the course influenced students to adopt more expertlike reasoning strategies in their postsurvey responses.

VIII. DISCUSSION

The analysis of the probes provides valuable insights into students' reasoning about measurement uncertainty and highlights areas for further discussion and exploration. Below, we highlight some takeaways from the analysis of all four probes.

A. Development of expertlike reasoning

The analysis of the pre- and postsurvey responses revealed trends in the development of expertlike reasoning about measurement uncertainty. Across the four probes, there was evidence of students transitioning from less sophisticated reasoning clusters to more expertlike ones.

In the RD probe, students from many clusters demonstrated a movement towards S4 reasoning, which reflects a deeper understanding of measurement uncertainty and the need to reduce uncertainty through repeated measurements.

Similarly, in the UR probe, students showed shifts from other setlike reasoning to S4 reasoning, which recognized the importance of reducing uncertainty through repeated measurements and considering the spread of data. This was particularly true for students who began in the S3 cluster, where they discussed in detail why the average is appropriate.

In the SMDS probe, students moved towards S2 reasoning, which considered both the mean and spread of data when comparing datasets. While students moved into this reasoning from many other types a reasoning, there was a large group that moved from P1, where they only considered the average, into S2 reasoning.

Lastly, in the DMSS probe, students demonstrated a transition from S1 and P3 reasoning to S3 reasoning, indicating an increased recognition of the importance of statistical variation and spread when comparing datasets.

Additionally, the presurvey results of all four probes showed a diverse range of reasoning clusters, indicating that students enter the course with varying levels of understanding and perspectives on measurement uncertainty. This suggests that students entered the class with with varying levels of prior knowledge about measurement uncertainty. However, in the postsurvey, there was a notable convergence of reasoning elements towards more expertlike clusters. This suggests that the transformed course had a unifying effect on students' reasoning about measurement uncertainty.

The consistent patterns of transition from less advanced reasoning clusters to more advanced ones across the different probes provide strong evidence for the development of expertlike reasoning about measurement uncertainty. The findings echo prior work [22] and indicate that the course had a significant impact on students' understanding of measurement uncertainty.

B. Challenges in transitioning to expertlike reasoning

One of the goals of the clustering analysis was to determine whether students who started with a particular type of reasoning all tended to move (or not move) together into another type of reasoning. This understanding could help create targeted interventions specifically tailored to students within certain clusters. While there were many positive shifts towards expertlike reasoning from some clusters, such as the transition from S3 to S4 in the UR probe or from S1 to S3 in the DMSS probe, challenges were observed in transitioning students from other clusters. This highlights the complexity and difficulty associated with developing a deep and nuanced understanding of measurement uncertainty.

One of the challenges observed was with students who responded with brevity to the open-ended questions in the PMQ, particularly those exhibiting U3 reasoning in the RD probe, stating "more data is better" without further explanation, or S1 reasoning in the UR probe, simply saying average. These students provided limited depth and insight in these responses, indicating a potential lack in understanding or engagement with concepts of measurement uncertainty.

These challenges in transitioning students from certain clusters may be attributed to a variety of factors. For example, students' prior knowledge and experiences may have influenced their initial reasoning patterns. Students who exhibit limited understanding or engagement with the concept of measurement uncertainty may have lacked prior exposure to the topic or may have encountered misconceptions that hindered their progression towards expert-like reasoning. Or, perhaps, students who responded with brevity may not be as comfortable with metacognitive practices of explaining their reasoning.

This finding underscores the opportunity to develop targeted interventions specifically tailored to students within these clusters. To support their learning and growth in reasoning about measurement uncertainty, interventions can take various forms. For example, for students who responded with brief responses on the PMQ, one may provide these students with increased opportunities to practice reflecting on their reasoning, encouraging them to critically analyze and evaluate their understanding of measurement uncertainty.

Furthermore, designing differentiated tasks that cater to the diverse reasoning observed within the class can be an effective strategy. By providing students with tasks that align with their current level of understanding and challenge them to progress further, we can foster the development of expertlike reasoning. These tasks can be carefully crafted to address the unique needs and difficulties associated with each cluster, promoting engagement, deep thinking, and application of statistical principles.

Moreover, scaffolded peer discussions can play a valuable role in enhancing students' understanding of measurement uncertainty. Students could be grouped to have a diversity of incoming reasoning responses. By engaging in collaborative conversations, students can actively explore and analyze uncertainty, interpret measurement data, and make informed judgments based on statistical principles. Peer discussions provide an opportunity for students to share their perspectives, challenge each other's reasoning, and collectively construct knowledge. Scaffolding these discussions, through prompts, guiding questions, or facilitation techniques, can support students in developing more sophisticated and nuanced understanding of measurement uncertainty.

C. Benefits of clustering analysis

Analyzing clusters of reasoning, rather than relying solely on class-wide trends, offers valuable benefits in understanding and addressing students' learning needs more effectively. For example, while previous class-wide analysis of the SMDS probe indicated little change in students' reasoning about measurement uncertainty [22], a closer examination through clustering analysis revealed a more nuanced picture. In fact, when individual student matched transitions were considered, it was found that over 50% of the class had actually changed their reasoning (Fig. 1). This highlights the importance of utilizing clustering analysis to interpret complex, high-dimensional data in a meaningful way.

By employing clustering analysis, we gain insights into the progression of student learning. This understanding allows us to delve deeper into the diversity of student reasoning and identify specific groups that may require targeted interventions. Unlike classwide trends that may mask individual variations, clustering analysis uncovers distinct clusters of reasoning patterns, enabling us to tailor instructional strategies and support to meet the specific needs of each group.

IX. CONCLUSIONS AND FUTURE DIRECTIONS

In conclusion, our study uses the PMQ and subparadigm coding scheme to investigate changes in students' reasoning about measurement uncertainty before and after an introductory physics lab course. Our findings revealed the complexity of student reasoning in this area, both in terms of the subtleties in understanding reflected in student responses (*RQ1*) and the changes in reasoning from pre to postsurvey (*RQ2*).

The use of clustering algorithms as an exploratory tool allowed us to identify distinct clusters of students based on their pre- and postsurvey responses, and revealed significant changes in reasoning clusters after the course transformation. We were able to successfully differentiate subparadigm level responses in order to analyze class-wide trends and quantify the effect of the lab on student alignment within the setlike paradigm.

The analysis of the probes provides valuable insights into students' reasoning about measurement uncertainty,

highlighting both the positive shifts and persistent challenges. The findings underscore the importance of targeted instructional interventions and comprehensive education about measurement uncertainty for physics students. Continued research and refinement of instructional strategies can contribute to improving students' scientific thinking and reasoning skills in the context of experimental measurements. By fostering a more expertlike understanding of measurement uncertainty, students can develop a stronger foundation for engaging with scientific inquiry and making informed decisions based on experimental data.

In conclusion, our study contributes to the understanding of measurement uncertainty reasoning in physics education and highlights the importance of developing undergraduate laboratories with explicit learning goals in this area. We also provide insights into tools and methodologies, such as the PMQ and hierarchical clustering algorithms, that can be used to evaluate the effectiveness of lab courses for promoting student learning. Further research in this area, including the use of computational solutions for qualitative coding and the use of new assessment tools, will continue to advance our understanding of how to effectively teach and assess measurement uncertainty in physics education.

ACKNOWLEDGMENTS

We thank the students in the course who took the time to share their experiences with us. This work is supported by STROBE National Science Foundation Science Technology Center, Grant No. DMR-1548924, and National Science Foundation Grant No. PHY-2317149.

APPENDIX: SUBPARADIGM CODING SCHEME

Full codebooks of sub-paradigm coding scheme. The development of the codebooks is described in Ref. [30].

TABLE XII. Codes for the RD probe, reprinted from Ref. [22].

Probe	Number	Paradigm	Name	Definition: "Argument is that…"
RD	P1	P	Measure the true value	…the experimenter could measure the correct value in a single measurement.
RD	P2	P	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	P3	P	Available time or resources	…a course of action is better due to considerations about how much time or resources it would require
RD	P4	P	Need to practice as you go	…practice is needed to account for errors or outside factors as measurements are being made
RD	P5	P	Misc. point	Pointlike argument that doesn't fit the other pointlike codes
RD	S1	S	Measure a spread	…multiple measurements will allow the experimenter to calculate or estimate a spread, variation, or uncertainty
RD	S2	S	Measure an average	…multiple measurements will allow the experimenter to calculate an average/mean
RD	S3	S	Use all the data together	…multiple measurements will all be used together to improve accuracy, precision, or goodness. Doesn't talk about average or spread specifically.
RD	S4	S	Reduce uncertainty of mean	…multiple measurements will be used to reduce the error or uncertainty of the mean or average.
RD	S5	S	Misc. set	Setlike argument that doesn't fit the other setlike codes
RD	U1	U	Just take more data	…experimenter needs to take more data. No statistical reasoning apparent.
RD	U2	U	More data cancels out error	…experimenter needs to take more data to cancel or outweigh the effect of error.
RD	U3	U	More data is better	…more data is better, more accurate, more precise, etc. Includes if reasoning other than statistical reasoning is apparent.
RD	U4	U	Misc.	Argument that doesn't fit into any of the other codes.
RD	U5	U	Unintelligible	Unintelligible, blank, or logically incoherent

TABLE XIII. Codes for the UR probe, reprinted from Ref. [22].

Probe	Number	Paradigm	Name	Definition: "Argument is that…"
UR	P1	P	Choose single value	…experimenter should choose a single value to report (for any reason).
UR	P2	P	Average as last resort	…experimenter should report the average because no better option exists.
UR	P3	P	Misc. point	Pointlike argument that doesn't fit the other pointlike codes.
UR	S1	S	Simply "average," or names reported value as average	States things like "I averaged," "do the average," "average is best," or "it is the average," but does not elaborate along the lines of the other codes. Includes statements that simply say what the reported value is.
UR	S2	S	Why average is useful	…reporting the average is best, because (in general) it accounts for fluctuations or errors, or because it predicts future measurements.
UR	S3	S	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. Does not include "it is the correct thing to do" (see S7).
UR	S4	S	Report average and spread	…experimenter should report the average and the uncertainty, range, or spread.
UR	S5	S	How to compute	Response explains how to compute the average. May be double coded when a separate explanation appears.
UR	S6	S	Discard outliers, then average	…experimenter should discard outliers or extreme data points, and then compute an average from the data that remains.
UR	S7	S	Misc. set	Setlike argument that doesn't fit the other setlike codes. Rule based reasons are coded here (e.g., "logical thing to do" or "the correct thing to do").
UR	U1	U	Misc.	Argument that doesn't fit into any of the other codes.
UR	U2	U	Unintelligible	Unintelligible, blank, or logically incoherent

TABLE XIV. Codes for the SMDS probe, reprinted from Ref. [22].

Probe	Number	Paradigm	Name	Definition: "Argument is that…"
SMDS	P1	P	The means are the same	…the groups agree because the means are the same.
SMDS	P2	P	Spreads don't matter	…the fact that the spreads or individual trials are different does not matter, including responses that focus on agreement of the averages while providing a reason for why the sets are different.
SMDS	P3	P	A has fewer outliers	…A is better because that group has fewer outliers, or A's individual measurements are more precise. Contains no reasoning about spread.
SDMS	P4	P	Differences in carefulness	…differences in the spread are due to differences in how carefully the measurements were performed.
SMDS	P5	P	Chose B, no explanation	Student chose "B" but left the explanation blank.
SMDS	P6	P	Misc. point	Pointlike argument that doesn't fit the other pointlike codes.
SMDS	S1	S	A is better	…group A is better, more accurate, more precise, etc. No further explanation.
SMDS	S2	S	Smaller spread is better, no mention of external factors	…a smaller spread, uncertainty, or range is better, more accurate, more precise, etc. The response does not mention external factors, outliers, human error, etc.
SMDS	S3	S	Smaller spread is better, due to external factors	…a smaller spread, uncertainty, or range is better, more accurate, more precise, etc. The response mentions external factors, outliers, human error, etc.
SMDS	S4	S	Chose A, no explanation	Student chose "A" but left the explanation blank.
SMDS	S5	S	Misc. set	Setlike argument that doesn't fit the other setlike codes.
SMDS	U1	U	Misc.	Argument that doesn't fit into any of the other codes.
SMDS	U2	U	Unintelligible	Unintelligible, blank, orlogically incoherent

TABLE XV. Codes for the DMSS probe, reprinted from Ref. [22].

Probe	Number	Paradigm	Name	Definition: "Argument is that…"
DMSS	P1	P	Means and spreads must both match	…the groups do not agree because in order to agree, the means and the spreads must both match.
DMSS	P2	P	Means must match	…the groups do not agree because the means are not the same (no mention of spread).
DMSS	P3	P	Means close enough, treats average as point	…the groups agree because the means are close enough.
DMSS	P4	P	Compare point-by-point, don't agree	…the groups do not agree. Data are compared point by point.
DMSS	P5	P	Compare point-by-point, do agree	…the groups agree. Data are compared point by point.
DMSS	P6	P	Chose B, blank explanation	Student chose "B" but left the explanation blank.
DMSS	P7	P	Misc. point	Pointlike argument that doesn't fit the other pointlike codes.
DMSS	S1	S	Means are close enough, talks about statistical variation in general	…the groups agree because the averages are close enough. Argument contains no reference to spreads, but does discuss statistical variation in general.
DMSS	S2	S	Similar means and spreads, no mention of overlap	…the groups agree because the means and spreads are similar. Argument does not consider the overlap between the means and/or spreads of the two datasets.
DMSS	S3	S	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	S	Chose A, blank explanation	Student chose "A" but left the explanation blank.
DMSS	S5	S	Misc. set	Setlike argument that doesn't fit the other setlike codes.
DMSS	U1	U	Not about statistics	…only nonstatistical things, such as systematics, are mentioned.
DMSS	U2	U	Cannot calculate uncertainty or spread	…the student states that they could not calculate the uncertainty or the spread, or that such values were not provided, with no further reasoning.
DMSS	U3	U	Misc.	Argument that doesn't fit into any of the other codes.
DMSS	U4	U	Unintelligible	Unintelligible, blank, or logically incoherent

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

[2] V. Otero and D. Meltzer, The past and future of physics education reform, Phys. Today **70**, No. 5, 50 (2017).

[3] B. R. Wilcox and H. J. Lewandowski, Developing skills versus reinforcing concepts in physics labs: Insight from a survey of students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 010108 (2017).

[4] B. R. Wilcox and H. J. Lewandowski, Open-ended versus guided laboratory activities: Impact on students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 020132 (2016).

[5] E. Etkina, Millikan award lecture: Students of physics—listeners, observers, or collaborative participants in physics scientific practices?, Am. J. Phys. **83**, 669 (2015).

[6] R. Khaparde, What are the objectives and goals of physics laboratory courses? A survey of college teachers, J. Phys. Conf. Ser. **1286**, 012037 (2019).

[7] 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).

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

[9] N. G. Holmes and E. M. Smith, Operationalizing the AAPT Learning Goals for the Lab, Phys. Teach. **57**, 296 (2019).
[10] 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).
[11] 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.
[12] R. L. Kung, Teaching the concepts of measurement: An example of a concept-based laboratory course, Am. J. Phys. **73**, 771 (2005).
[13] R. Beichner, The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project, in *Research-Based Reform of University Physics*, edited by E. F. Redish and P. Cooney (American Association of Physics Teachers, College Park, 2007), Vol. 1, https://www.per-central.org/items/detail.cfm?ID=4517.
[14] D. S. Abbot, Assessing student understanding of measurement and uncertainty, Ph.D. thesis, North Carolina State University, 2003.
[15] E. Etkina and A. Van Heuvelen, Investigative science learning environment - A science process approach to learning physics, *Research-Based Reform of University Physics*, edited by E. F. Redish and P. Cooney (American Association of Physics Teachers, College Park, 2007), Vol. 1, https://www.per-central.org/items/detail.cfm?ID=4988.
[16] 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).
[17] N. G. Holmes, Structured quantitative inquiry labs: Developing critical thinking in the introductory physics laboratory, Ph.D. thesis, The University of British Columbia, 2014.
[18] L. E. Strubbe, J. Ives, N. G. Holmes, D. A. Bonn, and N. K. Sumah, *Developing Student Attitudes in the First-Year Physics Laboratory* (American Association of Physics Teachers, College Park, MD, 2016), pp. 340–343.
[19] 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.
[20] H. J. Lewandowski, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and B. Pollard, Student reasoning about measurement uncertainty in an introductory lab course, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.056.
[21] E. M. Smith and N. G. Holmes, Best practice for instructional labs, Nat. Phys. **17**, 662 (2021).
[22] 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).
[23] J. Day and D. Bonn, Development of the concise data processing assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).
[24] 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).
[25] 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).
[26] B. Pollard, R. Hobbs, R. Henderson, M. D. Caballero, and H. J. Lewandowski, Introductory physics lab instructors' perspectives on measurement uncertainty, Phys. Rev. Phys. Educ. Res. **17**, 010133 (2021).
[27] M. Vignal, K. Rainey, B. Wilcox, M. D. Caballero, and H. J. Lewandowski, Affordances of articulating assessment objectives in research-based assessment development, presented at PER Conf. 2020, Grand Rapids, MI, 10.1119/perc.2022.pr.Vignal.
[28] 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).
[29] 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).
[30] B. Pollard, R. Hobbbs, D. R. Dounas-Frazer, and H. Lewandowski, Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Pollard.
[31] 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).
[32] S. C. Johnson, Hierarchical clustering schemes, Psychometrika **32**, 241 (1967).
[33] J. H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. **58**, 236 (1963).
[34] 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, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Lewandowski.
[35] H. J. Lewandowski, B. Pollard, and C. G. West, Using custom interactive video prelab activities in a large introductory lab course, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Lewandowski.
[36] B. Pollard and H. J. Lewandowski, Transforming a large introductory lab course: Impacts on views about experimental physics, presented at PER Conf. 2018, Washington, DC, 10.1119/perc.2018.pr.Pollard.
[37] B. Pollard, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab course on students' understanding of measurement uncertainty, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.073.
[38] 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).

[39] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20**, 37 (1960).

[40] J. C. Gower, A general coefficient of similarity and some of its properties, Biometrics **27**, 857 (1971).

[41] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.020146 for a comprehensive list of raw student code, their associated clusters, and number of student responses corresponding to the raw code for every probe pre and postsurvey.

[42] G. W. Milligan and M. C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika **50**, 159 (1985).

[43] *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives*, edited by L. W. Anderson and D. R. Krathwohl (Allyn & Bacon, New York, 2001), 2nd ed.