Investigating Students' Understanding of Measurement Uncertainty in a Two-Course Physics Laboratory Sequence

Authors & Affiliations

River Ward[1], Trevor Franklin[1], Marcos D. Caballero[1-4], and Rachel Henderson[1,3]
[1]Department of Physics & Astronomy, Michigan State University, 567 Wilson Rd. East Lansing, MI 48824
[2]Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, 48824
[3]CREATE for STEM Institute, Michigan State University, East Lansing, MI 48824
[4]Department of Physics and Center for Computing in Science Education, University of Oslo, Oslo, Norway

I. Introduction

An important aspect of teaching is the ability to measure the progress of your own students and make improvements to instruction. With the recent development of the American Association of Physics Teachers (AAPT) Recommendations for the Undergraduate Physics Laboratory Curriculum [1], a significant portion of the Physics Education Research (PER) community has turned much of their attention toward the undergraduate laboratory environment. The recommendations emphasize the learning of laboratory practice and research skills such as data analysis, scientific communication, and, more specifically, the understanding of measurement uncertainty.

In general, the Physics Measurement Questionnaire (PMQ) has historically been used to assess student understanding of measurement uncertainty; however, little is known about the PMQ's ability to measure learning over multiple measurements throughout a laboratory course sequence. Throughout this study, we investigated the longitudinal properties of responses from the PMQ in the context of Michigan State University's introductory physics laboratories. Over the two semesters, we tracked students' evolution in the types of responses given and determined two statistically significant shifts in the categories of responses.

II. Background

A. DATA Lab Background

In response to the national call from the PER community to focus on student engagement within the physics laboratory context [1], Michigan State University (MSU) physics department has recently transformed its algebra-based, introductory physics laboratory curriculum. Design, Analysis, Tools, and Apprenticeship (DATA) Lab is a two-course laboratory sequence specifically intended for non-physics majors [2]. Much like traditional laboratory courses, DATA Lab consists of one mechanics-based course (DATA Lab I) and another involving electromagnetism and optics (DATA Lab II). However, unlike more traditional laboratory courses, DATA Lab emphasizes the development of experimental skills and laboratory practices and provides students with an authentic physics laboratory experience instead of validating physical phenomena traditionally taught within physics lectures. In this course, students engage in the exploration of physical systems to increase their understanding of data analysis, model development, measurement uncertainty, and scientific communication. In the first-semester mechanics course, students spend several weeks, through guided workshops, focusing on specific laboratory practices such as measurement uncertainty, data analysis and modeling, and experimental design; this results in delaying technical laboratory work until the latter half of the semester. This is done so that by the time students are ready to conduct an experiment, they are familiar with common lab practices, uncertainty calculations, and working collaboratively in a laboratory environment. In the electromagnetism lab, students begin experimenting much sooner with periodic reminders to refresh laboratory norms and practices. For more details about the overall design and structure of the newly transformed laboratory sequence, see Funkhouser, *et al.* [2].

B. PMQ Background

To formally assess the student's understanding of measurement uncertainty, we administered the Physics Measurement Questionnaire (PMQ). The PMQ is an open-ended assessment tool used to measure student understanding surrounding measurement uncertainty [3]. The assessment centers around the context of an experiment of a ball rolling down a slope and off of a table. Multiple measurements of the distance that the ball travels from the edge of the table are then given and multiple questions (probes) are asked in relation to the uncertainty of these measurements [4]. Specifically, the work presented below focused on four probes: Repeating Measurements (RD), Using Repeated Measurements (UR), Same Mean Different Spread (SMDS), and Different Mean with Same Spread (DMSS).

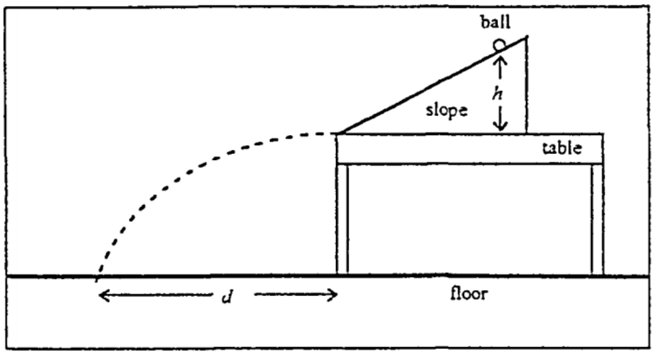


Fig. 1. The setup for the experiment done by each of the four probes. A ball is released from height *h* and rolls off the edge to a distance *d* away from the table. Students are then tasked with measuring varying concepts surrounding *d*. This figure is reproduced from Ref. [3].

The RD probe presents a single measurement to students and asks if additional measurements are needed to be made in order for the data to be valid. The UR probe’s format is slightly different from the other probes: it presents one table of five measurements, and asks students to record a single value as the final result. Students must then give a free response explaining why they chose their value.

The SMDS probe has students presented with two tables of five measurements and asked to select a table and explain if either set of data is ‘better’ than the other. Each of the tables has the same mean but differing spreads.

The DMSS probe presents two data sets as well, but with differing means and the same spread instead. Students must explain which of the two tables is better or if they are in agreement with each other.

As each of the four probes are designed to allow students to explain their reasoning, the open-ended nature of responses are intended to be categorized into either point-like or set-like reasoning (as defined by Allie *et al.* [3]). Point-like reasoning focuses on individual measurements within the probe. This type of reasoning would typically ignore the presence of statistical error or spread. Set-like reasoning takes into account the data as a whole, looking for trends and noting spread and statistical error [3]. This type of reasoning includes the proper use of terminology such as mean or standard deviation. Below, we describe the procedure we conducted to categorize the student responses into point-like and set-like reasoning.

III. Methods

A. Coding Procedure

As open-ended responses vary considerably from person to person, we need a method to classify them into broad paradigms (i.e. point-like and set-like). Based upon the definitions of point-like and set-like reasoning from Allie *et al.* [3], researchers at the University of Colorado Boulder developed a codebook for the PMQ by grouping similar responses together into smaller categories called subcodes [4]. Following this process, about fifteen to twenty subcodes were generated per each of the PMQ probes. Although students tend to align with only one subcode, some responses are labeled with multiple subcodes. Moreover, students’ responses sometimes do not fit perfectly into point-like or set-like. For example, the beginning of a student’s explanation might be strongly set-like while the latter half is strongly point-like. In this case, we think of their response as ‘mixed’. Most probes have a low amount of students as mixed, typically less than 10%, with the exception of RD which has about 50%. Responses that are generally unrelated or miscellaneous are also included in the mixed paradigm.

Despite having a succinct method of classification, there will always be differences in how language is interpreted. To help eliminate individual biases in interpretation, two researchers used the CU Boulder codebook [4] to independently code small subsets of about 100 to 200 responses. Once complete, an Inter-Rater Reliability (IRR) score was calculated. This process was repeated until a comfortable IRR score, based on Cohen’s Kappa, for each probe was achieved and the remaining responses were coded separately. The highest Cohen’s Kappa scores achieved before independent coding for RD, UR, and SMDS indicated “substantial agreement” at 0.60, 0.76, and 0.69 respectively. Likewise, Cohen’s Kappa for DMSS was 0.56, indicating moderate agreement. Ultimately, 9,504 student responses were coded.

B. Student Population

From the students who took a DATA Lab course in the 2018-2019 academic year, 986 students responded to the PMQ. Of these students, 785 were enrolled in only one course in the DATA Lab sequence (either DATA Lab I or II) while 201 students enrolled in both. The PMQ was administered at the beginning and end of each course, meaning the 201 students enrolled in both courses responded a total of four times. While we present data from each of these moments of assessment, we will pay special attention to this group of students for a longitudinal PMQ assessment.

IV. Results

A. Longitudinal Graphs

Figure 2 describes the percentage of students in each paradigm at each assessment in their two-course journey. The red squares represent the fraction of students who responded point-like, blue circles are set-like, and purple diamonds are 'mixed-like'. One can notice these fractions and evolutions vary dramatically by probe. It is important to note the time scale on which assessments were given. The time between DATA Lab I pretest and posttest was a full semester as well as between DATA Lab II pretest and posttest. DATA Lab I posttest and DATA Lab II pretest are instead separated by a few weeks, between fall semester 2018 and spring semester 2019.

For the RD and DMSS probes, there is a 8% and 18% increase, respectively, in the fraction of students in the set-like reasoning paradigm from the DATA Lab I pretest to the DATA Lab I post-test; however, for the UR and the SMDS probes there is very litter shift in the fraction of students in each of the paradigms. An additional observation from Figure 2 is that the UR probe has more that 90% of students responding with set-like reasoning and retains that level throughout the two-course laboratory sequence.

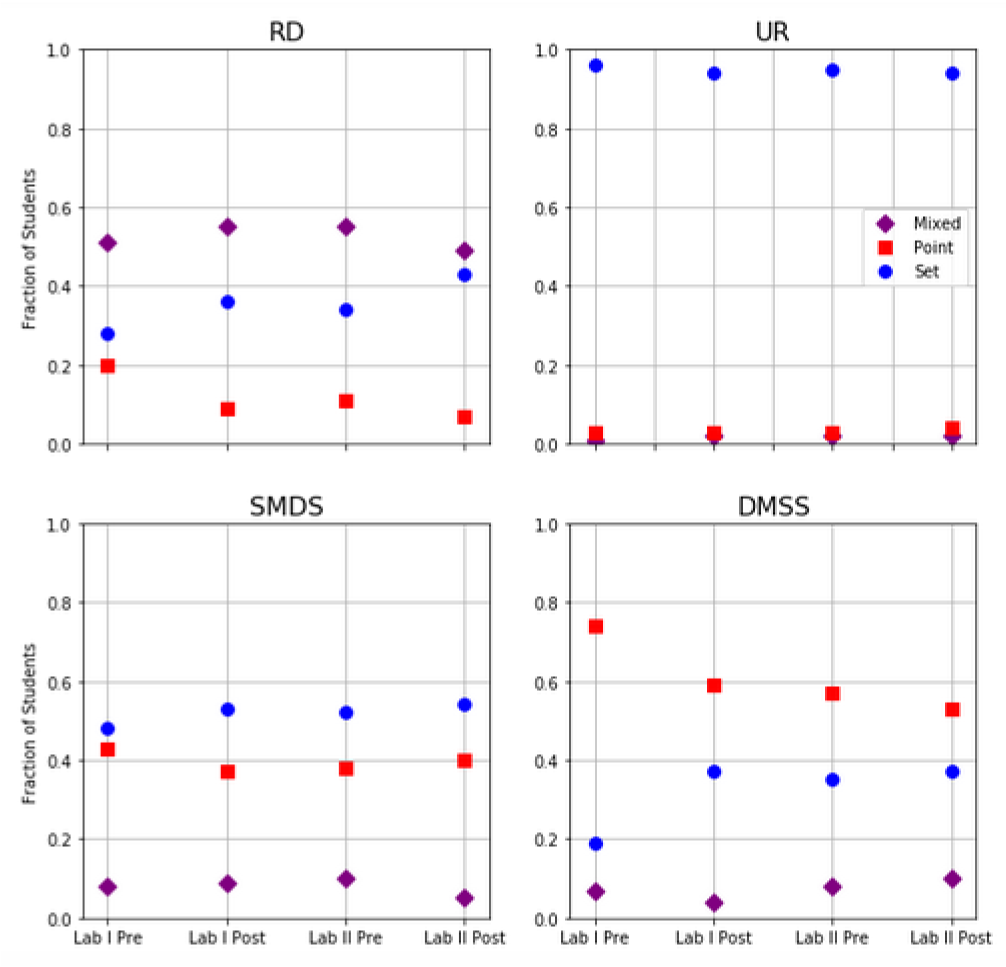


Fig. 2. Fractions of student responses in each paradigm separated by probe. Paradigm fractions are taken at each of four assessments. Labels are abbreviated from DATA Lab I & II pretest and DATA Lab I & II post test.

Probe	DATA Lab I (Pretest → Post-test)	Transition (Post-test → Pretest)	DATA Lab II (Pretest → Post-test)
RD	0.0650*	0.0269	0.0735
UR	0.0404	0.0119	0.00928
SMDS	0.0431	0.0136	0.0650
DMSS	0.138*	0.0510	0.0301

Table 1. Cramer's V effect size separated by probe and by assessments. Results with " * " represent p-values less than 0.017 to indicate statistical significance.

B. Contingency Tables

While the visualization of the plots above are important to understanding the longitudinal properties of the PMQ probes, we also determined which, if any, of the overall paradigm shifts between assessments were statistically significant. To do so we constructed a 2x3 contingency table of assessment number (pretest and post test) and paradigm (set, point, or mixed). Table 1 presents the effect sizes in addition to the statistical significance which has been Bonferroni corrected for the inflation of Type I error. Cramer's *V* effect sizes are quantified as 0.1 as a small effect, 0.3 as a medium effect, and 0.5 as a large effect. Only one assessment transition, DMSS DATA Lab I pretest to DATA Lab II posttest, can be described as having a small effect size of 0.138. All remaining transitions and probes have effect sizes less than 0.08. Therefore from Table 1, DATA Lab I has the largest impact on an increase in set-like thinking.

In general, the paradigm proportions in all probes (with the possible exception of RD) either saturate after subsequent assessments or remain mostly unchanged throughout all assessments. This can be both observed visually in Figure 2 but is seen analytically in Table 1. Only two instances occurred where a p-value was less than 0.0167 to indicate statistical distinguishability: RD and DMSS between the pretest and post-test of the first course. The remaining combinations of comparisons are insignificant. In particular, paradigm shifts after the DATA Lab I post test are always insignificant, regardless of the probe. This suggests that students in DATA Lab II do not experience overall significant changes in knowledge.

V. Discussion

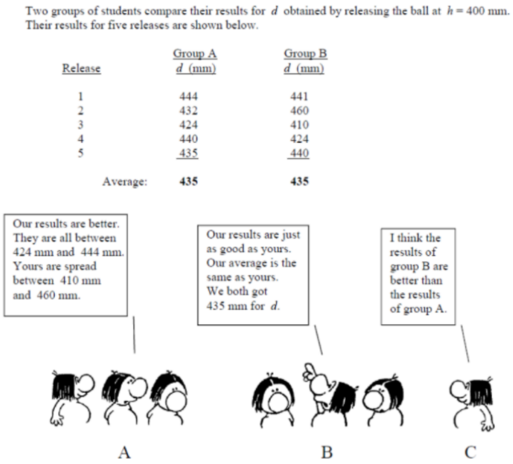
For the most part, throughout the two-course laboratory sequence, we find that the fraction of students with set-like reasoning changes very little over time. However, one possible explanation for the small change in student knowledge within the RD and DMSS probes at the beginning of the course sequence is the difference in course structure from DATA Lab I to DATA Lab II. As mentioned in the DATA Lab background above, the first semester course focuses on developing an understanding of data analysis and laboratory techniques followed by students conducting their own experiments given a certain physical system. In contrast, the second semester course layers these experiments with several sessions to refresh knowledge learned in the previous semester. Therefore, the initial "jump" in set-like reasoning seen in the RD and DMSS probes may be a result of specific reinforcement of measurement uncertainty knowledge present in the first semester course through a guided workshop at the very beginning of the course sequence. Knowledge that students gain in DATA Lab II is not detectable by the PMQ and may require a different assessment tool.

Additionally, there are small changes in paradigm proportions between DATA Lab I post test and DATA Lab II pretest for all probes. This is likely due to the presence of winter break (a few weeks) between courses as opposed to full semesters between other assessments. This is further corroborated by small decreases in set-like reasoning consistent with a regression in knowledge after instruction. This phenomenon is observed in other studies of student learning at the university level [5].

We should also acknowledge that although the overall proportion of paradigms remains roughly unchanged in DATA Lab II, this does not mean students are unchanged individually. For example, we traced the paradigm shift of each student individually and determined that shifts in paradigms can sometimes cancel each other out. In SMDS for example, roughly 10% of students shift from point to set and another roughly 10% of students shift vice versa. So although the net proportions remain mostly constant, some students have simply swapped each other's paradigms. More work should be done to further investigate this phenomenon.

Set and Point: An Opposite Relationship

An interesting observation is that a change in set-like responses is almost always accompanied by an equal and opposite change in point-like responses with little change in mixed. This is most easily observed in the first assessment comparison for RD and DMSS but can be generally seen in all probes except UR. Moreover, this effect seems to be independent of the proportion of students in each paradigm. Consider RD and DMSS in Figure 2. In RD, a majority of students respond as mixed, followed by set, and followed then by point. In DMSS, the opposite is observed - a majority respond as a point, followed by set, and followed then by mixed. Yet, despite the differing proportions, the relationship is still observed in the first assessment comparison.



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

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

Our results are better. They are all between 424 mm and 444 mm. Yours are spread between 410 mm and 460 mm.
Our results are just as good as yours. Our average is the same as yours. We both got 435 mm for d.
I think the results of group B are better than the results of group A.
A
B
C

FIG. 3. The prompt of the SMDS probe. Students are asked "With which group do you most closely agree?" and given the option to choose either A, B, or C. Students must then give a written response to explain their choice. This figure is reproduced from Ref. [3].

A possible explanation for this observation in RD and DMSS is that these probes offer the student potential actions rather than actions with detailed explanations as in SMDS. As seen in Figure 3 for SMDS; group A states, 'Our results are better' while group B states, 'Our results are just as good as yours' with both groups providing an explanation. Although group C does not provide an explanation for their thought, less than 2% of students choose C. In contrast, both answer choices in DMSS simply state if their results agree or disagree. A similar observation can be made for the RD probe. It is therefore possible that students have significant shifts in paradigm when provided with actions as in RD and DMSS but insignificant shifts when provided detailed explanations as in SMDS or no assistance whatsoever as in UR.

A. Limitations of the UR Probe

As seen in Figure 2, even prior to instruction, students respond as set-like at a rate of over 90% throughout their experience in DATA Lab. Although not pictured, this observation is equally seen in students who only enrolled in one section of the course sequence. In fact, we have not seen any significant increase or decrease in set-like reasoning in any subset of students thus far. There are a few potential explanations for this effect.

Students entering university may already be familiar with reporting an average. Since the PMQ was developed nearly twenty years ago for use in uncertainty assessment in South African schools, the education background for the intended population is potentially quite different compared to students enrolled at an American research-intensive university like MSU. It is possible that students entering MSU have already extensively used averages in school work or even in everyday life. Regardless of how this knowledge was acquired, this probe is not useful for measuring changes in student paradigms at the university level.

The UR probe might be too vague and does not promote students to answer in a way that demonstrates a change in knowledge. Of all the students who reported set-like responses, 44% responded by simply reporting an average without further elaboration. The remaining students expanded on this answer by explaining the importance of an average, or reporting a standard deviation, or considering the effect of an outlier on the average, etc. Since the UR probe simply asks, “what number should be reported for the data set”, students in general may be reluctant to report additional information aside from one number. It is also worth noting that of the four probes on the PMQ, UR is the only one that does not ask the students to expand on one of several provided answers. Therefore, UR might be assessing how students respond when completely unassisted. Even if this makes UR a more objective question, again, we did not observe any significant change in student understanding under subsequent assessments.

B. Implications for instructors

In this study, the only significant shifts in student reasoning around measurement uncertainty were during the first course of the two-course sequence. As discussed above, the curriculum for DATA Lab I course was designed to facilitate a workshop specifically scaffolding students to begin to think about measurement uncertainty. Keeping this in mind, it may be useful for laboratory instructors to implement these types of scaffolds to explicitly engage students in this scientific practice more thoroughly throughout the semester.

Additionally, it would be interesting to see the evolution of response paradigms at other institutions or even at the pre-university level. With the exception of RD, all probes tend to saturate in paradigm distribution by at most the fourth assessment. It is possible that RD offers the most insight into paradigm shifts at the university level; this is because RD had the overall lowest levels of statistical indistinguishability between assessments. However, due to having the only noteworthy effect size of 0.138, DMSS may be useful at measuring paradigm shifts at the university level. Other probes like SMDS could be useful when applied at the pre-university level since changes in knowledge for this probe may occur earlier in a student’s career.

Furthermore, the UR probe might be more suited for assessing students before entering a research-intensive university. Again, no significant change in student understanding was measured in any subset of students for this probe. Since UR is primarily focused on reporting an average, this probe may be a more appropriate measuring tool for high school science classes that cover basic data analysis. It would be interesting to see what the initial distribution of students looks like and how that evolves over a semester or possibly an academic year.

VI. Concluding Remarks

We have observed several effects present in the overall paradigm shifts for students enrolled in MSU's DATA Lab course sequence. Although statistically significant paradigm shifts occur from pretest to post-post in DATA Lab I for the RD and DMSS probes, SMDS and UR do not measure any paradigm shifts in student reasoning. All four probes on the PMQ measure a noticeable saturation in student knowledge after the first semester in DATA Lab. This could be due to the difference in course structure in DATA Lab I and DATA Lab II. Namely, DATA Lab I spends more time developing understanding of data analysis and less time performing traditional lab work than DATA Lab II. Differences in the question construction between these two groups of probes could explain why this occurs at the university level as well as the opposite relationship between point and set. Much like UR, SMDS could serve as an insightful probe to administer to students at the pre-university level as well. From what has been observed in this study, we hope to provide a meaningful step forward toward determining not what students know, but how students learn about the laboratory practice of understanding measurement uncertainty.

VII. References

[1] Kozminski, Joseph, et al. "AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum." *AAPT Physics Education*, American Association of Physics Teachers, 10 Nov. 2014, www.aapt.org/resources/upload/labguidlinesdocument_ebendorsed_nov10.pdf.

[2] Funkhouser, Kelsey, et al. " Design, Analysis, Tools, and Apprenticeship (DATA) Lab." *IOPscience*, European Journal of Physics, 11 Sept. 2019, iopscience.iop.org/article/10.1088/1361-6404/ab2f0d/pdf?casa_token=RK0FBxc-4IkAAAAA:1nmTxdtOcInn9-a_Xb4CMOPe8-Y4O6x5QqRcaYGxjqTLMxZ-6xVhWYnU710QkDewXIYwFA5uXQ.

[3] Allie, Saalih, et al. "First−Year Physics Students' Perceptions of the Quality of Experimental Measurements." *Taylor & Francis*, Journal International Journal of Science Education, 23 Feb. 2007, www.tandfonline.com/doi/abs/10.1080/0950069980200405.

[4] Pollard, Benjamin, et al. "Impact of an Introductory Lab on Students' Understanding of Measurement Uncertainty." *ArXiv.org*, Physics Education Research Conference, 14 Aug. 2020, arxiv.org/pdf/1707.01979.pdf.

[5] Kohlmyer, Matthew A., et al. “Tale of Two Curricula: The Performance of 2000 Students in Introductory Electromagnetism.” *Physical Review Physics Education Research*, American Physical Society, 5 Oct. 2009, journals.aps.org/prper/abstract/10.1103/PhysRevSTPER.5.020105.