Investigating student ability to draw conclusions from measurement data

Tong Wan and Joshua M. Mickelsen
Department of Physics, Westminster College, 1840 South 1300 East, Salt Lake City, Utah, 84105

In this study, we surveyed students from a calculus-based and an algebra-based introductory physics course at a liberal arts college about their ability to draw conclusions from measurement data. Both courses are taught in a studio mode and use the *Workshop Physics* curriculum. The survey questions were adapted from Kok *et al.* (2019), who found that an increase in the number of decimal places hinders students' ability to compare data sets. We administered the survey online before and after instruction on measurement uncertainty. On the survey, students considered two experiments that differ only by one setup. Students were first asked to make predictions about the experimental outcomes as to whether or not the outcomes agree, and then were given data to analyze and draw conclusions. The survey had two versions where the measurement data had either two or four decimal places. We used the framework of point and set paradigms to characterize student reasoning. The set paradigm emphasizes a measurement distribution rather than a single measurement; it is considered more expert-like. The results show that students tended to switch from a correct to an incorrect answer after analyzing the data. The number of decimal places did not seem to correlate with the switch in student answers. We also found that after instruction on measurement uncertainty, student reasoning tended to shift from the point paradigm toward set paradigm as many students included the standard deviation or standard deviation of the mean in the analysis. However, many students did not seem to recognize how the uncertainty could inform the conclusion as to whether or not the data sets agree; most students appeared to rely only on the comparison of the means. We discuss implications for instruction as well as future research areas.

2021 PERC Proceedings edited by Bennett, Frank, and Vieyra; Peer-reviewed, doi.org/10.1119/perc.2021.pr.Wan

I. INTRODUCTION

Proficiency with data evaluation has been identified in national reports as one of the key learning goals in K-12 as well as in undergraduate education [1–3]. Undergraduate students are expected to "analyze and display data using statistical methods and critically interpret the validity and limitations of these data and their uncertainties" and "present results and ideas with reasoned arguments supported by experimental evidence [3]." Indeed, measurement uncertainty is an important topic for students to master in order to draw data-driven conclusions. Many research-based curricula for introductory physics lab courses include measurement uncertainty as a learning goal [4–8].

Student reasoning with measurement uncertainty has been characterized by the point and set paradigms [9]. The point paradigm is characterized by the belief that a single measurement can in principle yield the true value. Individual measurements are considered independent of one another. If repeated measurements are made, a student with the point paradigm makes decisions based on individual data points, such as a recurring value in a set data. In contrast, a student with the set paradigm considers that individual measurements are only approximations to the true value. A set of measurements that form a distribution is required to obtain information regarding the true value. In other words, the point paradigm is characterized by the notion that the measurement uncertainty can in principle be zero while the set paradigm is characterized by the notion that the measurement uncertainty is inevitable, which is more expert-like. It is worthwhile to note that the point and set paradigms focus on statistical uncertainty. Systematic errors are beyond the scope.

In line with the framework of point and set paradigms, the Physics Measurement Questionnaire (PMQ) [9, 10] was developed to classify student reasoning about statistical uncertainty. Prior studies have shown that traditional labs improve student reasoning about measurement uncertainty as measured by PMQ [10, 11], and reformed labs have an even larger effect [11].

Other studies have explored factors that influence students' ability to draw conclusions from measurement data [12–14]. When given a data set, statistical uncertainty can be identified by a qualitative examination of variations within the data set. Data with more decimal places are more exact, but the uncertainty becomes more apparent. Kok *et al.* [13] found that an increase in the number of decimal places hinders German middle school students' ability to critically compare data sets. As the number of decimal places increased, more students switched from a correct to an incorrect hypothesis about the measurements from two experiments that differ only by one setup. They argued that this is because students lack the knowledge of measurement uncertainty.

Informed by the findings from Kok *et al.*, we have started to investigate college students' ability to draw conclusions from measurement data, and the extent to which instruction on measurement uncertainty supports student reasoning and decision-making. We are not aware of any studies that examine effective instruction that addresses the issue identified in Kok *et al.* at the K-12 level. Therefore, it is worthwhile to examine how, if at all, the number of decimal places influence college students' ability to compare data sets.

We adapted the questions from Kok *et al.* and administered in both algebra-based and calculus-based introductory mechanics courses at a liberal arts college in the western United States. Both courses are taught in a studio mode and use the *Workshop Physics* [8] curriculum, in which measurement uncertainty is discussed extensively in one of the units. The adapted questions are in line with items in PMQ, which was developed for college level. It is worth noting that all probes in PMQ regard one single experiment whereas the questions in Kok *et al.* and ours concern two experiments that differ by one setup. To evaluate the extent to which instruction on measurement uncertainty supports students to critically compare data sets and draw data-driven conclusions, we administered the questions before and after instruction on measurement uncertainty. We used the framework of point and set paradigms to characterize student reasoning.

This study is intended to answer two research questions (RQs): (1) Does the number of decimal places influence college students' ability to draw conclusions from measurement data? and (2) To what extent does instruction on measurement uncertainty impact students' ability to analyze data and draw data-driven conclusions?

II. METHODS

The algebra-based and calculus-based introductory mechanics courses cover the first 13 units of the *Workshop Physics* [8] curriculum, which utilizes a collaborative, activity-based approach in place of traditional lectures and labs. The first two units cover measurements and uncertainty. The rest of the units cover typical topics in a mechanics course (e.g., motions and forces). Unit 1 introduces measurement, data collection, and graphical display of data. Unit 2 introduces statistical uncertainty, systematic errors, as well as formalism and interpretations of standard deviation (SD) and standard deviation of the mean (SDM). Students are also instructed to use spreadsheets to record data and conduct data analysis. It is worth noting that the curriculum involves cases where experimental data are compared to the theoretical values, but does not involve cases where two data sets are compared against each other.

The courses were taught by two faculty members, each of whom taught two sections. Each section meets for 1 h and 50 mins, three times per week. In light of COVID-19, the calculus-based course had half of the class periods meet in person, and the other half were held online synchronously over the course of instruction of the first two units. The algebra-based course adopted a hybrid mode with about two-thirds of the students meeting in person while the rest joining online using instructor pre-made videos. Every student at-

tended two in-person classes and one online class each week. Another faculty (other than the instructor of record) joined online to engage students. Since the first two units only involved time measurements, students who joined online made measurements using the videos with their own devices. We acknowledge that the teaching modalities in light of the pandemic is a limitation of the study. However, the study only spanned over two weeks and a few online classes may not have had a measurable impact on student learning.

A. Survey questions

We administered the survey (adapted from Kok *et al.*) online before (pretest) and after instruction of unit 2 (post-test). It is worth noting that both pre- and post-test were given before instruction on motions and forces so that students must reason based on the experimental data. Table I shows the participants in each section. The class enrollments are shown in the parentheses. Only students who completed both pre- and post-test were included.

On the survey, students considered the elapsed times of a free falling object from two experiments. The two experiments are identical except one object has a zero and the other has a non-zero initial horizontal velocity. Students were first asked to make a prediction about the elapsed times as to whether or not they are the same. It is worth noting that students were not able to change their predictions after they submitted their answers. They were then given two sets of data and asked about the strategy they would use to analyze the data. They were also asked to conduct an analysis and draw a conclusion. Lastly, students were asked to explain their reasoning.

TABLE I. Participant in each section of the courses. The enrollments are shown in parentheses. The total number of participants is 49.

	Algebra-based	Calculus-based
Section 1	14 (16)	9 (11)
Section 2	17 (23)	9 (10)

TABLE II. Data sets of the two experiments given to students in each section of both courses. The elapsed times are in seconds.

Version 1 (Section 1)		Version 2 (Section 2)	
Ball A	Ball B	Ball A	Ball B
0.59	0.67	0.5949	0.6734
0.71	0.79	0.7127	0.7923
0.67	0.72	0.6754	0.7202
0.64	0.62	0.6452	0.6210
0.66	0.73	0.6623	0.7374
0.66	0.62	0.6687	0.6252

As shown in Table II, students enrolled in different sections of the courses received different versions of the data sets (either in 2 or 4 decimal places). On both pre- and post-test, students enrolled in section 1 of both courses completed version 1, and students in section 2 completed version 2.

The prediction and conclusion were in multiple-choice format, and the data analysis strategy and explanation to conclusion were in free-response format. The prediction question was excluded in the post-test because we believed that most students would probably predict the same as their conclusion on the pretest, which would not provide new insights.

The original survey is written in German. We created an English version based on the description of the questions in Ref. [13] and made a few modifications. First, we changed the multiple-choice format of the data analysis strategy question to a free response format. Since the populations were different, the answer options from the original study may not apply to our study. Second, the data points presented to students were different. On the original survey, means are the same (when rounded to the same decimal place as the data points), and the SDMs are very small. In our study, the two data sets have different means but are overlapped within the SDMs, which is similar to items in PMQ (where the means are different but are overlapped within the SDs). Ball A had a mean elapsed time of $0.66s \pm 0.02s$, and Ball B had a mean elapsed time of $0.69s \pm 0.03s$. The confidence intervals defined by $mean \pm SDM$ are overlapped, and therefore the elapsed times are essentially the same. Additionally, the original study used three versions of measurement data, and we used two due to a small sample. Lastly, the original study had additional questions that were beyond the scope of this study, and therefore were excluded. The validity of the modified questions was checked by consultations with physics faculty and individual student interviews ($N = 5$).

B. Data Analysis

We used the point and set paradigms to code students' data analysis strategy and conclusion [9]. The coding scheme (as shown in Table III) was a modified version of the DMSS probe in PMQ developed by Pollard *et al.* [11, 15]. The DMSS probe concerns comparing two data sets collected by two experimenters from the same experimental setup. The modifications were intended to account for the differences in the questions, as well as institutional and course context. Specifically, we added codes P5, S4, and T, and we split one code in Ref. [11] into U1 and U2 to capture new themes emergent from the responses.

Multiple codes can be assigned to a single response. A response would be considered a point response if point codes were assigned but no set codes assigned. It would be a set response if set codes were assigned but no point codes assigned. Lastly, it would be mixed if both point and set codes were assigned. There were a few instances (4%) when neither point or set codes was assigned. These were also counted as

TABLE III. Modified definitions of codes based on Ref. [11].

Paradigm	Identifier	Definition
Point	P1	compare means; one mean is greater
	P2	compare means; means are close enough
	P3	data are compared point-by-point; one set is greater
	P4	data are compared point-by-point; the two sets are about the same
	P5	compare max, min, range, and/or median
	P6	misc. point
Set	S1	calculate mean, SD and/or SDM; mentions overlap
	S2	calculate mean, SD and/or SDM; no mention of overlap
	S3	discuss significance of the difference in means in general
	S4	compare difference in means to the range of data points
	S5	misc. point
Neither	U1	compare difference in means to the uncertainty due to other factors
	U2	compare difference in means to the uncertainty due to instrument precision
	T	theoretical argument, not based on data
	O	misc. point or blank

mixed paradigm since students' responses showed no clear inclination toward either point or set paradigm.

There were some differences between coding the strategy and conclusion. For strategy, the coding was simply based on the free-response question. However, for conclusion, codes were assigned based on both the multiple-choice question and the explanation. Another difference was that a few pairs of codes P1 & P2, P3 & P4, and S1 & S2 were combined when coding strategy question because these codes provided information about the conclusion. In general, the codes were not mutually exclusive with a couple exceptions. If a student mentioned they would calculate SD and/or SDM in addition to the mean, the combined code S1 & S2 would be assigned without P1 & P2. For conclusion, either P1 or P2 can be coded together with S1 or S2 if students calculated SD/SDM but draw a conclusion based on the means only.

To examine the interrater reliability (IRR), both authors coded the responses. First, each author familiarized themselves with the responses as well as the code definitions before they coded independently. They then selected responses from 10 participants randomly, and compared the codes. After inconsistencies were discussed and resolved, each author then reviewed the codes for the rest of the 39 participants. The IRR was calculated based on the responses from the 39 participants. The Cohen's κ was 0.79, indicating a substantial agreement [18]. Lastly, the authors discussed the rest of the responses and resolved inconsistencies.

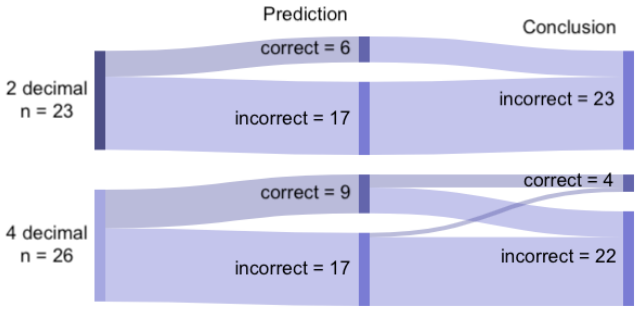

FIG. 1. Number of students in each group who gave a correct or incorrect answer for the prediction and conclusion on pretest.

Due to a small sample, we used Fisher's exact test [16] (with unmatched data) and McNemar's test [17] (with matched data) for multiple-choice questions. We used the same tests for free-response questions after the frequencies of the student paradigms (point, set, or mixed) were obtained. Cramer's V and odds ratio were calculated for effect sizes.

III. RESULTS

Data were aggregated across the algebra-based and calculus-based courses due to a small sample. As as a result, we draw conclusions for all students in the introductory mechanics courses rather than by course, which we acknowledge as a limitation of the study.

We compared students' predictions and conclusions on the pretest. Of all students, 12 switched from a correct to an incorrect answer after analyzing data, while only one switched from an incorrect to a correct answer (McNemar's test, $p = 0.006$, odds ratio $= 12$). This suggests that students tended to switch from a correct to an incorrect answer after analyzing the data. The most common (65%) reasoning in the explanation to conclusion was P1, the mean of one data set is greater than that of the other.

To answer RQ1, we first compared student answers (correct vs incorrect) on the prediction between the groups (2 vs 4 decimal places). The difference was not statistically significant (Fisher's exact test, $p = 0.552$), which suggests that the two groups of students were comparable. This allowed us to isolate the number of decimal places as a factor for conclusion. We then examined student answer patterns on prediction and conclusion (see Fig. 1). In both groups, six students switched from a correct to an incorrect answer. On the 2 decimal places version, no student switched from an incorrect to a correct answer; on the 4 decimal places version, one student switched from an incorrect to a correct answer. We found no significant correlation between the number of decimal places and student answer patterns (Fisher's exact test, $p = 0.354$).

To answer RQ2, we first compared students' conclusions on pretest and post-test (see Table IV). Most students (40 out of 49) chose an incorrect conclusion on both pretest and post-

TABLE IV. Students' conclusions on pretest and post-test.

		Post-test	
		Correct	Incorrect
Pretest	Correct	2	2
	Incorrect	5	40

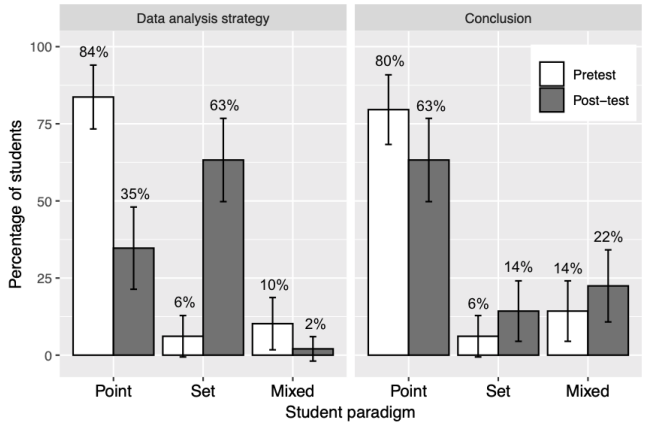

FIG. 2. Percentage of students in different paradigms (point, set, or mixed) on the questions of data analysis strategy and conclusion on pretest and post-test. Error bar represents 95% confidence interval.

test. Only a few students switched conclusions, two switched to an incorrect conclusion and five switched to a correct conclusion on post-test, but the difference is not statistically significant (McNemar's test, $p = 0.450$).

We also evaluated student paradigms on the question of data analysis strategy, and explanation to conclusion, respectively (see Fig. 2). On the data analysis question, most students (84%) were categorized in the point paradigm on the pretest. Of all students, about 86% stated that they would compare the means, and about 10% included statements that suggest they would compare data point-by-point. On the post-test, the proportions of students in all three paradigms are significantly different from the pretest (Fisher's exact test, $p < 0.001$, Cramer's $V = 0.603$, large effect). Pairwise comparisons with the Holm-Bonferroni correction suggests that students shifted from the point paradigm toward the set paradigm ($p_{adj} < 0.001$), and also from the mixed paradigm toward the set paradigm ($p_{adj} < 0.001$). On the post-test, about 61% stated they would calculate SDs and/or SDMs in addition to the means.

We observed no shift in student paradigms on the explanation to conclusion from the pretest to post-test (Fisher's exact test, $p = 0.200$). On the post-test, about 63% of students were in the point paradigm, reasoning based on the appeared difference in the means. About 14% of students reported SD and/or SDM, but concluded that the elapsed times are different. It is likely that these students also drew conclusions based on the means only as the statements did not explicitly discuss how SD and/or SDM informed the conclusion. Interestingly, another 6% stated that SD/SDM is irrelevant to the comparison of the elapsed times since SD/SDM concerns precision.

IV. DISCUSSION

This study examined the correlation between the number of decimal places and introductory students' ability to draw data-driven conclusions, and the impact of instruction on measurement uncertainty on students' ability to analyze data and draw conclusions. The results showed that students tended to switch from a correct to an incorrect answer after analyzing data sets from two experiments, in which the means appear different. It seemed students drew conclusions based on the appeared difference in means only, even if many students realized it is necessary to examine SD and/or SDM in data analysis. We found no correlation between the number of decimal places and student answer patterns, which contrasts with the finding from Kok *et al.* We note two major differences between their study and ours. The study of Kok *et al.* involved German middle school students in grades 8−10, while our study involved college students in the U.S. Additionally, the means of the two data sets given in Kok *et al.* appear the same, but they appear different in ours. To account for the differences in the study design and results, we came up with two hypotheses. First, college students have already gained sufficient experience with data collection and analysis from high school and/or other science courses in college before taking introductory physics, and therefore their ability to compare data sets are not influenced by the number of decimal places. Second, the number of decimal places has an influence on college students, but the effect of the appeared difference in the means dominates, which makes the influence of number of decimal places less significant. Future research should test these hypotheses. The results would provide insights into instruction around measurement uncertainty.

We also found that after instruction on measurement uncertainty, students shifted toward the set paradigm on the data analysis strategy. About two-thirds of all students mentioned that they would calculate SD and/or SDM. Some of them provided interpretations of SD/SDM. One student, for example said that "I would calculate the standard deviation to determine how spread out each data set is and calculate the SDM to evaluate uncertainty." However, we found no significant change in either student answer choice or paradigm on the conclusion. Many students seemed to be able to interpret SD/SDM, but did not recognize how SD/SDM could inform a conclusion about two data sets. As mentioned previously, the *Workshop Physics* curriculum does not introduce methods for comparing data sets, but only covers comparisons of one data set to a theoretical value. The results suggest that students do not spontaneously transfer their knowledge of uncertainty between scenarios. Instruction should explicitly address cases for comparing data sets.

[1] NGSS Lead States, Next Generation Science Standards: For States, By States (The National Academies Press, Washington DC, 2013).
[2] AAMC-HHMI Committee, Scientific foundations for future physicians. Report of the AAMC-HHMI committee (2014).
[3] AAPT Committee on Laboratories, AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum (American Association of Physics Teachers, College Park, MD, 2014).
[4] R. F. Lippmann, Students' Understanding of Measurement and Uncertainty in the Physics Laboratory: Social Construction, Underlying Concepts, and Quantitative Analysis, University of Maryland, 2003.
[5] R. J. Beichner, J. M. Saul, D. S. Abbott, J. J. Morse, D. L. Deardorff, R. J. Allain, S. W. Bonham, M. H. Dancy, and J. S. Risley, The Student-Centered Activities For Large Enrollment Undergraduate Programs (SCALE-UP) project, Research-Based Reform of University Physics edited by E. F. Redish and P. J. Cooney (American Association of Physics Teachers, College Park, MD, 2007), Vol. 1.
[6] E. Etkina and A. V. 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. J. Cooney (American Association of Physics Teachers, College Park, MD, 2007), Vol. 1.
[7] N. G. Holmes, Structured Quantitative Inquiry Labs: Developing Critical Thinking in the Introductory Physics Laboratory, The University of British Columbia, 2014.
[8] P. W. Laws, The Workshop Physics Activity Guide (Wiley, 1998 & 2004), Modules 1-4.
[9] A. Buffler, S. Allie, F. Lubben, and B. Campbell, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).
[10] 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).
[11] 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).
[12] A. Buffler, F. Lubben, B. Ibrahim, The relationship between students' views of the nature of science and their views of the nature of scientific measurement, Int. J. Sci. Educ. **31**, 1137 (2009).
[13] K. Kok, B. Priemer, W. Musold, and A. Masnick, Students' conclusions from measurement data: The more decimal places, the better?, Phys. Rev. Phys. Educ. Res. **15**, 010103 (2019).
[14] B. Priemer, S. Pfeiler, and T. Ludwig, Firsthand or secondhand data in school labs: It does not make a difference, Phys. Rev. Phys. Educ. Res. **16**, 013102 (2020).
[15] B. Pollard, R. Hobbs, D. R. Dounas-Frazer, and H. J. Lewandowski, Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses, in Proceedings of the 2019 Physics Education Research Conference, Provo, UT, edited by Y. Cao, S. Wolf, and M.B. Bennett (AIP, New York, 2020).
[16] R. A. Fisher, On the interpretation of χ^2 from contingency tables, and the calculation of P, J. R. Stat. Soc. **85**, 87 (1922).
[17] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika **12**, 153 (1947).
[18] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. (Routledge, London, 2013), p. 567.