New perspectives on student reasoning about measurement uncertainty: More or better data

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

(Received 6 December 2022; accepted 8 June 2023; published 19 July 2023)

Uncertainty is an important and fundamental concept in physics education. Students are often first exposed to uncertainty in introductory labs, expand their knowledge across lab courses, and then are introduced to quantum mechanical uncertainty in upper-division courses. This study is part of a larger project evaluating student thinking about uncertainty across these contexts. In this research, we investigate advanced physics student thinking about uncertainty by asking them conceptual questions about how a hypothetical distribution of measurements would change if "more" or "better" data were collected in four different experimental scenarios. The scenarios include both classical and quantum experiments, as well as experiments that theoretically result in an expected single value or an expected distribution. This investigation is motivated by our goal of finding insights into students' potential point- and setlike thinking about uncertainty and of shining light on the limitations of those binary paradigms.

DOI: 10.1103/PhysRevPhysEducRes.19.020105

I. INTRODUCTION

The concept of uncertainty is a fundamental aspect of physics [1], particularly in undergraduate instructional laboratories (labs) [2]. While many physics lab instructors cite uncertainty-related goals for their courses, the format of these goals ranges from procedural (e.g., carrying out procedures to propagate uncertainties or reporting measurements with uncertainties), conceptual (e.g., describing how standard deviation captures the variability between trials or rounding limits inform uncertainty in single measurements), to more agentic (e.g., deciding what are the major sources of uncertainty in an experiment and designing experiments to minimize those sources). The physics education research community has also captured students' understanding of and proficiency with uncertainty through a range of perspectives with a range of intended goals [[3], references therein].

The most prominent physics education research thread has been through the classification of students' reasoning about uncertainty as either part of a point paradigm or set paradigm [4]. Reasoning with the point paradigm includes ideas such as that any individual measurement could be exactly the "true" value, repeated measurements are not necessary and do not need to be combined, and measurements do not need to be listed with their uncertainties. Reasoning with the set paradigm includes ideas such as that any measurement is just an approximation of the phenomenon being measured, a deviation between measurements is to be expected, combining repeated measurements helps establish the best estimate and its uncertainty, and all measurements should be reported with their uncertainties.

Research characterizing students according to these two paradigms has generally found that students may exhibit either style of thinking depending on the question [4–7]. For example, in these studies, many students studied exhibited pointlike thinking on questions about whether repeated measurements were necessary. When comparing datasets, however, many students used mixed reasoning, which includes both pointlike and setlike ideas. More recent work has found evidence that purely pointlike reasoning is quite rare among introductory college physics students [8–11]. Altogether, these findings suggest that student thinking about uncertainty is not unidimensional, may be context dependent [12], and may not neatly fit into one of the two paradigms.

Evaluation of students as setlike or pointlike thinkers have evaluated students' *procedural* knowledge about uncertainty, defined as being "concerned with 'doing science' … rather than with the scientific concepts themselves. Thus, procedural knowledge (in the context of experimental work) will inform decisions, for example, when planning experimental investigations, processing data and using data to support conclusions" [4] (p. 1137). For example, the Physics Measurement Questionnaire [13]

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

presents students with experimental scenarios and sample measurement data and asks students about possible next procedures, including whether to perform repeated measurements, how to report the best estimate of the measurand, or how to compare pairs of measurements. Underlying these procedural decisions is a conceptual understanding of what uncertainty is and where it is coming from. In our companion work, we studied students' conceptions of where measurement uncertainty may be coming from in a range of experiments [14]. We find that students hold a wide range of ideas about the sources of uncertainty that depend on the experimental scenario. We also find that there is much more nuance in student thinking about measurement beyond the point and set paradigms.

Altogether, these studies motivate a broader examination of student thinking about measurement uncertainty from new and distinct perspectives. This study is part of a broader project to evaluate these new perspectives. While previous assessments have asked students about what they think should be done next in an investigation, here we probe what students think will happen to experimental measurement data under two experimental interventions: (i) Adding additional data (under the same conditions) and (ii) obtaining new data by experts with the best possible equipment. We refer to these as the "more" and "better" data questions, respectively. We can contrast these questions to the procedural assessments of students' thinking about uncertainty, which ask students about *whether* to take more data or improve measurement equipment. Those assessments infer pointlike thinking through responses that indicate, for example, not needing additional data [5]. Conversely, our questions can assess students' pointlike or setlike thinking through their perspectives of what *will* happen under these experimental settings. For example, pointlike thinking may be exhibited by students indicating that either setting will result in a single value. Setlike thinking, however, may be represented by any number of options (the distribution of data points may not change, may become more narrow but still have some variability, or may become more wide). Compared with the procedural assessments used to characterize pointlike and setlike thinking, students' justifications for their predictions will provide new and unique insights into their understanding of what is causing the variability and whether variability is a relevant and necessary construct for the experiments.

In addition to the new perspective provided by the types of questions, we also probe student thinking in several different experimental scenarios (Fig. 1) that span both physics paradigms (classical and quantum mechanics) as well as theoretical expected outcome (single value and a distribution). In contrast, previous assessments have only probed student thinking within a single-value classical

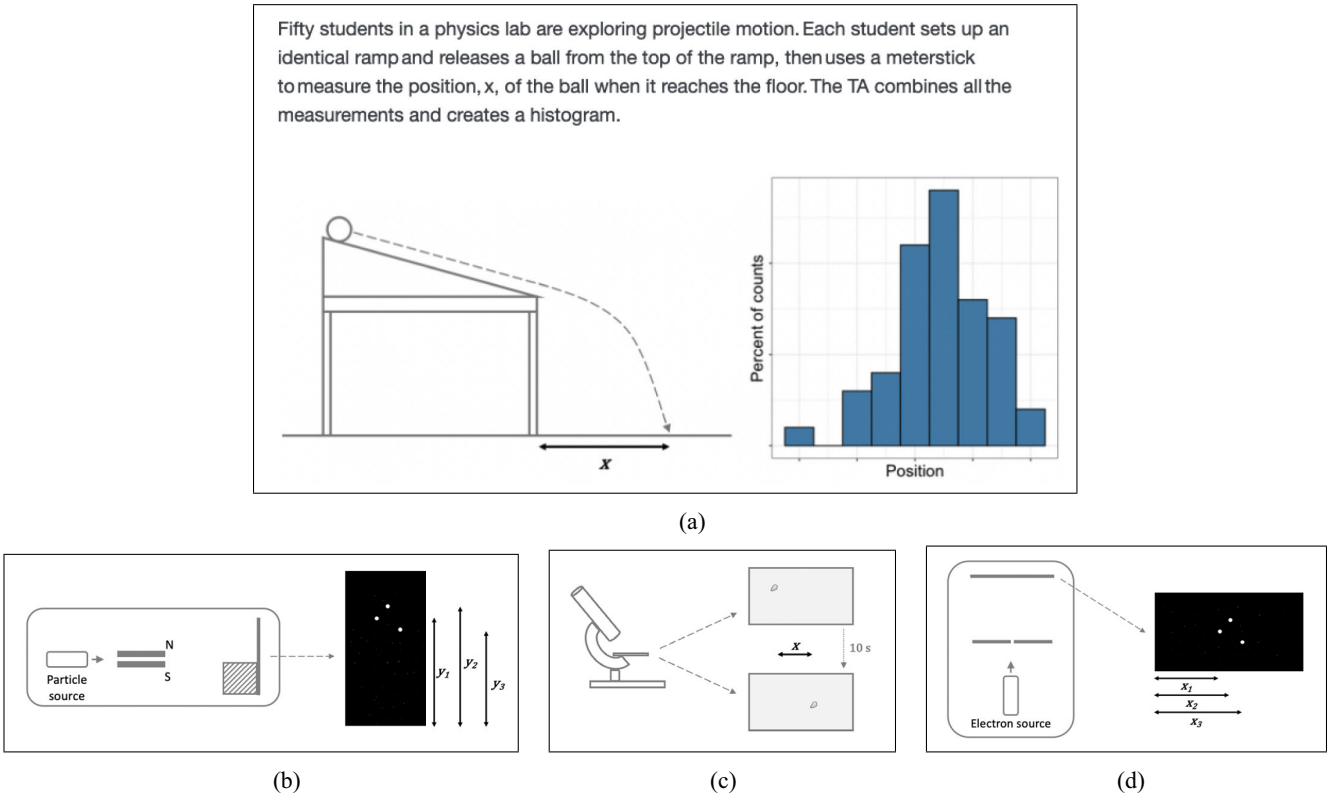

FIG. 1. Diagrams for the four experimental scenarios included in the survey. The descriptive text and histogram for the projectile motion experiment are included as an example. (a) Projectile motion scenario, (b) Stern-Gerlach scenario, (c) Brownian motion scenario, and (d) single-slit scenario.

mechanics scenario, which we include as one of our four scenarios. We also survey advanced physics students, as opposed to those in introductory laboratory courses. Our preliminary research has identified that students may consider classical and quantum systems as distinct experimental scenarios with different rules regarding uncertainty and variability [15–17]. This distinction is not surprising given the documented examples of conflicts and tensions in student thinking between classical and quantum mechanics more generally [18–21].

Ultimately, our research question asks how do students evaluate the impact of more and "better data" on experimental measurements across a range of experimental scenarios? In what follows, we find four main results. First, very few students exhibit definitional pointlike thinking, where the measurements would result in a single value. Second, many students appropriately understand the ways in which more data do not impact the width of a distribution of measurements and the ways in which better data cause a distribution to narrow. Third, we find two key alternative concepts: A misattributed "more data is better" heuristic (or conflating standard deviation and standard error); and that fundamental physical principles (such as the Heisenberg Uncertainty principle) eclipse measurement limitations for some experimental scenarios. Finally, we find that student thinking about the impacts of more and better data varies marginally with experimental physics scenario.

II. METHODS

This research is part of a larger project investigating student thinking about uncertainty and measurement. Data for this work come from a survey we developed for broad dissemination to probe student thinking across multiple perspectives. For analysis of student thinking about uncertainty from a different perspective, please see our companion paper [14].

A. Survey development

The survey centers on four experimental scenarios in which students in a laboratory course perform an experiment and take data. Respondents are provided with a description of each scenario, a schematic of the experimental setup, and a histogram of fictitious data, as seen in Fig. 1. After the description of the experimental setup, respondents are asked a series of questions, after which they are presented with a second experimental scenario and asked the same set of questions again in this new scenario. The first scenario shown to all students in this study was the projectile motion scenario [shown in Fig. 1(a)]. The second scenario was randomly chosen between a Stern-Gerlach experiment [Fig. 1(b)], a Brownian motion experiment [Fig. 1(c)], and a single-particle, single-slit experiment [Fig. 1(d)]. These scenarios provide a set of conditions in a 2×2 of physics paradigm (classical: projectile motion and Brownian motion; and quantum: Stern-Gerlach and single-slit) and theoretical expected outcome (single value: projectile motion and Stern-Gerlach; and distribution: Brownian motion and single-slit). The Stern-Gerlach experiment has one of the two output channels blocked so that the theoretical experiment outcome would be a single value. The histogram of fictitious data is identical across scenarios and represents distance measured with a ruler.

The final two questions asked in each scenario are the focus of this paper. These questions asked respondents what might happen to the distribution if *more* data or *better* data were taken. The full language of these questions can be seen in Fig. 2. Respondents could select one option from

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

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

Multiple choice options for each question:

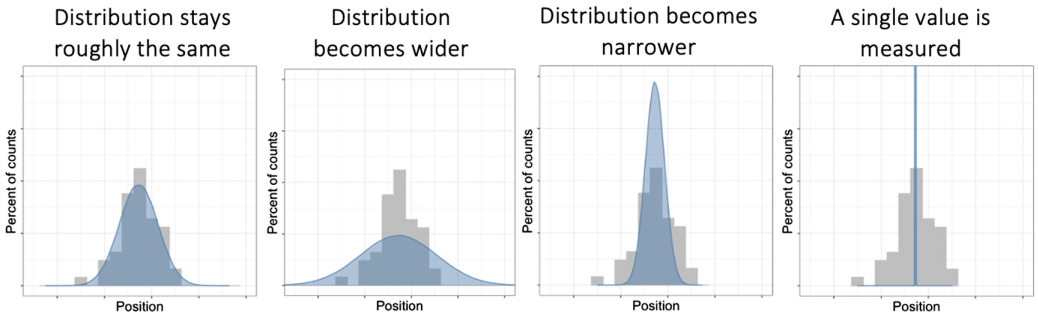

FIG. 2. More and better data questions with the multiple-choice options. After each question, students were provided with a text box to explain their reasoning.

the four presented in Fig. 2 and were given an open text box to explain their reasoning. These questions were designed to probe the ideas of setlike and pointlike thinking in more detail and through new perspectives. While the PMQ probes setlike and pointlike thinking through procedural questions, such as "what should you do next?" our questions are more conceptual in nature and instead provide students with the opportunity to describe how (if at all) they believe the results of an experiment would change if more data or better data were taken. For example, we would expect that pointlike thinkers would select that the distribution would become a single value with either *more* or *better* data.

B. Survey distribution

The survey was distributed to students in quantum mechanics courses at five institutions: Cornell University, University of St. Andrews, Michigan State University, University of Colorado Boulder, and California State University, Fullerton. The surveys were administered during the second half of both the Fall 2020 and Spring 2021 semesters electronically using Qualtrics. Students were invited to participate by their course instructors and could opt to enter a draw to win a $25 gift card if they responded to the survey. In total, we received 150 completed student responses to the survey with the majority of students in the third year or fourth of their degree. The demographic information on the participating students is provided in Table II. We do not have any information about the students' relevant prior knowledge, the ways their instruction had discussed measurement and uncertainty, nor whether they had completed experiments similar to the ones in this survey. We did ask students to indicate their level of comfort with each of the scenarios they saw in the survey. Over 90% of student respondents reported being comfortable with the projectile motion and single-slit experiments and approximately 75% of student respondents reported being comfortable with the Stern-Gerlach and Brownian motion experiments. Future work will seek to evaluate the ways prior knowledge and instruction interact with students' reasoning on these and other questions about measurement and uncertainty.

C. Data analysis

The questions analyzed in this work are multiple choice with a prompt to provide an explanation. Student explanations were coded using a coding scheme developed during the analysis of other questions on the survey and using the Modeling Framework for Experimental Physics [22]. The coding scheme is described below and detailed information on its development and modifications for other uses can be found in Ref. [14].

Each student response was given one or more of the following codes developed *a priori*: principles, limitations, statistics, or other. The first two codes were based on components of the Modeling Framework for Experimental Physics [22]. Examples of student explanations that fit each code and code definitions are provided in Table I. The *principles* code identifies reasoning that indicates the shape of the distribution is caused by variability inherent to a theoretical abstraction of the experiment. Student reasoning about principles can be either classical or quantum mechanical in nature and can relate to the physical or measurement systems of the experiment, as distinguished in the Modeling Framework for Experimental Physics [22]. The *limitations* code identifies reasoning that describes some practical limitations with the experiment, including instrumental imperfections and human error, or anything that is not inherent to the principles of the system being measured [22]. The *statistics* code is unlike the other two in that it is not based on the modeling framework and is unique to the more and better data questions. This code is used when the student explains the shape of the new distribution through a statistical or data-driven lens, such as considering the statistical effects of adding more measurements to a dataset. These responses do not reference any element of the experimental scenario or physical or measurement systems, instead considering the statistical effects independently from what might be physically occurring. As in the examples in Table I, responses coded for statistics could be applied to any of the experiments. Students may have been considering how the probabilistic nature of a particular system is the cause of the distributions they mention, but, as written, we do not have evidence to this effect. Finally, the *other* code was used when the explanation did not fit into any of the three previous categories, often because it was too vague or simply descriptive of their answer choice without explaining why they chose it.

It is important to note that there is not necessarily a single correct code for any of the scenarios or questions. Student responses assigned to each code demonstrate a wide range of expertlike and novicelike thinking. Moreover, a student response that is coded as "limitations" (for example) does not indicate any lack of understanding or awareness of reasoning about the principles of the experiment. Rather, the response only tells us what they chose to write to justify their multiple-choice answer to this particular question. This is particularly true in the case of the statistics code, which is distinctly different from the other two in that the student is responding to the question by only discussing the data and not the experimental parameters (be it the physical or measurement system principles or limitations [22]). Both the answers students selected (single value, more narrow, same, and more wide) and the code assigned to their reasoning provide insights into student thinking. We find that the most information is learned when we look at the answer and reasoning codes together.

A subset of the data (95 responses) was coded by three of the authors. Cohen's kappa was calculated to be greater than

TABLE I. Coding scheme for more and better data questions. Responses that did not fit into these codes or were too vague to code were coded as "other."

Code	Definition	Examples
Principles	Reasoning that indicates the shape of the distribution is caused by variability inherent to a theoretical abstraction of the experiment. It may be classical or quantum in nature.	"As you increase the number of iterations your system should converge to the proper solution which should be one solveable value for a kinematic problem." (Projectile motion) "The distribution of values obtained comes from the fundamental randomness of the superposition breaking into one of the spin eigenstates. It doesn't have to do with equipment and expertise." (Stern-Gerlach) "I think the randomness of Brownian motion is probably contributing more to the variation than experimental errors, so the distribution might stay the same regardless." (Brownian motion) "I think the dominant defining characteristic of the distribution remains quantum mechanical and inherent to the physics of the experiment, and so the general of the shape of the distribution would stay the same, but become more perfect to a particular functional form." (single-slit)
Limitations	Reasoning that describes some practical limitations with the experiment, including instrumental imperfections to human error, that is not inherent to the principles of the physical or measurement system.	"Using the best possible equipment means reduction in human error (like giving initial push and initial position on ramp)." (projectile motion) "The expert will eliminate the difference in the initial velocity, so the position uncertainty will be reduced to zero, that is, all the upwardly deflected particles will reach the same position." (Stern-Gerlach) "I know some labs are set on material that absorbs vibrations and you can probably compensate somewhat for the lingering motion from moving the plate." (Brownian motion) "The better equipment will allow more accuracy in electron production and precision in measurements. Also any irregularities in the slit can be reduced." (single-slit)
Statistics	Reasoning that explains the shape of the new distribution through a statistical or data-driven lens, such as considering the statistical effects of adding more measurements to a dataset.	"By increasing the sample size, were likely to get more measurements around the peak of the original histogram and less outlying measurements. This is because an increase in sample size decreases the standard deviation for measurement." (projectile motion) "The expectation value (or peak) should remain the same under multiple trials; however, the probability to measure each thing should also stay the same." (Stern-Gerlach) "The underlying distribution is Gaussian so you expect it to remain about the same." (Brownian motion) "The positions will be different for the 3 electrons for any 3 particular measurements but the more measurements that are taken, the more the statistical shape of the position probability density will begin to take shape." (single-slit)

0.8 across each pair, indicating fair interrater reliability. All disagreements were then discussed and resolved, with clarifications then made to the codebook to reduce the occurrence of further disagreements. The full dataset was then coded by the first author and a fourth author read through all responses and associated codes to double-check the consistency with which codes were applied to responses. Fewer than a handful of codes were changed during this final review.

Our goal here was to uncover overarching trends in the student responses. We chose not to perform statistical tests because several of the proportions were too small to perform χ^2 tests of distinguishability and relying on p values for interpretation is strongly misleading [see, for example, [23–25]]. Any statements about "differences" or "similarities" in our results section, therefore, should be interpreted as "qualitatively distinguishable from the graph" or "not qualitatively distinguishable from the graph," respectively.

III. RESULTS

We break down our discussion of the results by first looking at which multiple-choice options students selected for each of the more and better data questions for each experiment (Sec. III A). We then look more closely at the reasoning students provided to justify their answers (Sec. III B).

A. Student responses to how the distribution might change

The distributions of student responses for each question and each experimental scenario are shown in Fig. 3. In response to both questions and all scenarios, we see that most students selected that the distribution would get more narrow or stay the same, with a small number of students selecting that it would become more wide or result in a single value.

Looking more closely at the "more data" responses, we note almost no students indicated that more data would result in a single value. This is perhaps unsurprising, as the only change in the experiment is taking additional data and the initial measurements resulted in a distribution. We also notice a few students indicating that the distribution will become more wide, with more students selecting this option for the three experiments that are either quantum mechanical and/or have a theoretically expected distribution (that is, not the projectile motion experiment).

The most common response for how the distribution might change when more data are taken is that the distribution will stay the same. We found this to be the case for all four experimental scenarios. We identify this response as the most expertlike response, as more data collected with the same procedures and equipment will not affect the overall variability in the measurements, though the estimate of the mean of those data would have smaller uncertainty. The next most common response is that the distribution would become more narrow. The fewest number of students selected this option in the Brownian motion scenario.

In response to the better data question, we notice a larger difference between experiments. We see that the projectile motion scenario receives a much larger proportion of more narrow responses than the other experiments. We identify this as an expertlike response (across all experiments) because experts with the best equipment will perform measurements with fewer experimental limitations than students with basic equipment.

We do not notice much difference between the other three experiments with students expecting the distributions to become more narrow (most frequent response) or stay the same (second most frequent response) at similar rates. Additionally, the projectile motion experiment received the largest number of responses indicating that there would be a single value, and students were overall more likely to predict that a single value would result from better data than from more data.

B. Student explanations about what happens to the distributions

We now take a look at the explanations students provided to justify their choices about the distributions. Because most students indicated that the distributions would either become more narrow or stay the same, we only investigate explanation codes for students who selected these two options. The distributions of student responses for these two answers are provided in the scaled pie charts in Figs. 4 and 5.

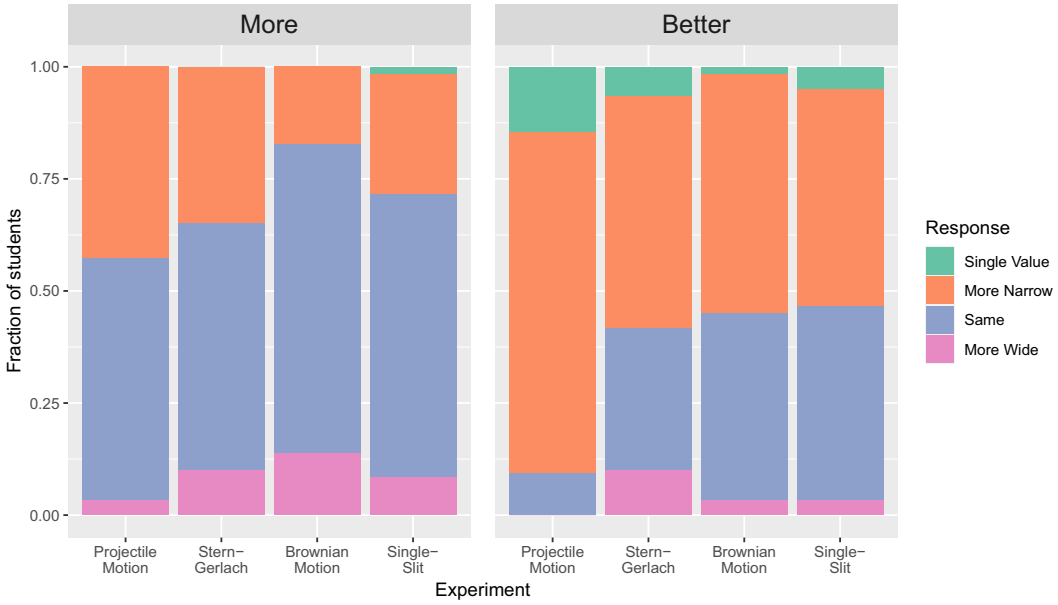

FIG. 3. Stacked bar plot of student responses to the more and better data questions for each experimental scenario.

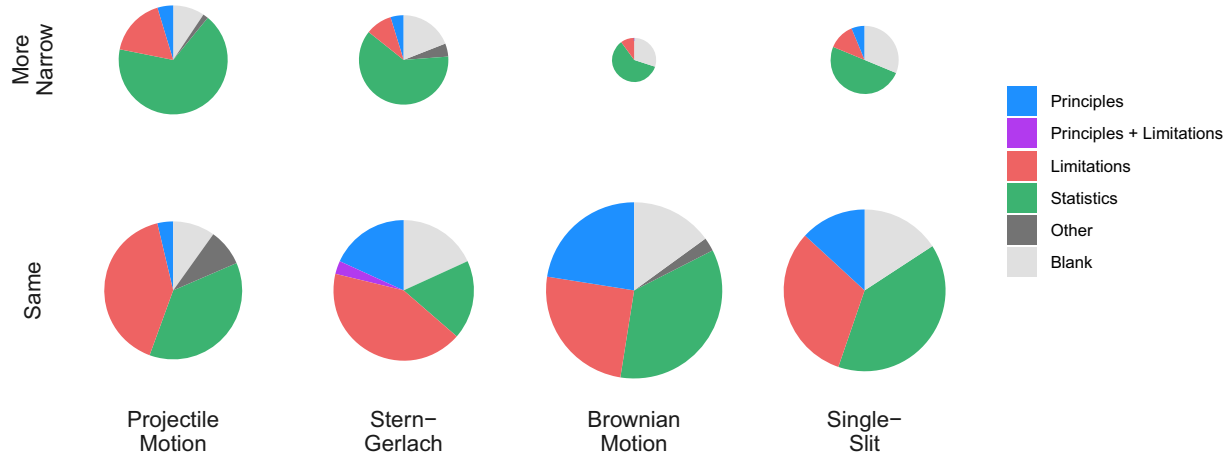

FIG. 4. Scaled pie charts of the more data explanation codes for more narrow and same distribution responses. The pie charts are normalized vertically according to the proportion of students who selected that distribution and the slices show the proportional breakdown of explanation codes.

1. More data explanations

In Fig. 4, we see that students who indicated the distribution would become more narrow used different reasoning than those who indicated the distribution would stay the same. When students answered more narrow, their explanations were primarily coded for statistics. These responses referred to purely statistical effects that do not consider the physical system, such as "the law of large numbers" or saying that more data reduce the standard deviation. For example, one student considering the projectile motion scenario said, "Distribution will indeed be a Gaussian, as you increase the number of trials, the random uncertainty in the outcome will decrease, standard deviation will decrease and hence standard error decrease to create narrower distribution." Another student, considering the Brownian motion scenario, said, "If there is a true value of x, then distribution would become narrower because of how most data points should trend to the true value of x. There will still be less points on the sides." A minority of students used limitations to explain the distribution becoming more narrow, typically inferring that more data reduce student-driven variability. For example, a student considering the single-slit scenario said, "If more students conduct the experiment, then the distribution will become more narrow because of the uncertainty caused by human-error is reduced." Another student considering the projectile motion scenario said, "I want to say the distribution becomes narrower, because the more that are set up the same, there are less factors to change the distribution and therefore the percentage of counts increases at the average."

Most students, however, indicated the distribution would stay the same, with a range of explanations across scenarios.

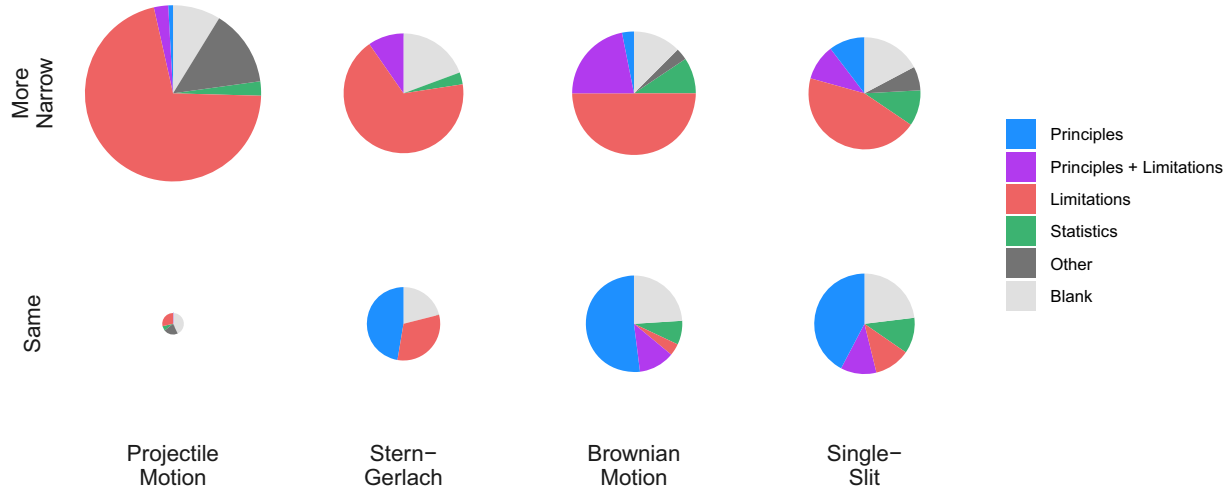

FIG. 5. Scaled pie charts of the better data explanation codes for more narrow and same distribution responses. The pie charts are normalized vertically according to the proportion of students who selected that distribution and the slices show the proportional breakdown of explanation codes.

Many students also used purely statistical ideas to explain why the distribution would stay the same, typically describing that more data would not change the standard deviation. For example, one student considering the Stern-Gerlach scenario said, "The distribution may become smoother, but will not change in characteristic shape as more data is collected with the same." Another student considering the Brownian motion experiment said, "Increasing n should not change the distribution since n is already large."

Similar numbers of students, however, used limitations to explain the distribution staying the same, such as that the additional students were still using the same equipment and procedures and so the same sources of uncertainty exist. For example, one student considering the projectile motion scenario said, "A hundred students will make as many random errors as fifty—the standard deviation of the measurement shouldn't change." Another student considering the Brownian motion scenario said: "Since the equipment is the same, the systematic errors should be similar. Presumably the other variables are the same (Temp), so statistical errors are similar as well. Thus the distribution is similar."

A large minority of students also argued for physical and measurement principles, particularly for the Stern-Gerlach, Brownian motion, and single-slit experiments. In these cases, students primarily indicated that the phenomena themselves were random and so additional data would not remove that inherent randomness. For example, a student considering the Stern-Gerlach scenario said, "The limitation is not the students' accuracy or the sample size but the quantum mechanical system." Another student considering the Brownian motion experiment said, "The motion is random, so it does not matter how many trials are taken."

Interestingly, we see relatively little variation between experimental scenarios on the explanation codes. Students used similar reasoning to explain why the distribution would become more narrow regardless of the physics paradigm or theoretical expected outcome. For the distribution staying the same, the relative proportions differ in only minor ways (such as fewer principles explanations for projectile motion and more limitations codes and fewer statistics codes for Stern-Gerlach). In addition, the role of statistics and limitations appear to be context independent, in that many of the quoted responses above could have been associated with any of the experimental scenarios.

2. Better data explanations

In Fig. 5, we see that when students indicated the distribution would become more narrow, their explanations were primarily coded for limitations. Most often, students' responses related to the fact that experts with expert equipment would remove much of the variability, uncertainty, and error present in the student measurements. For example, a student considering the projectile motion scenario said, "The distribution would become narrower because, in my opinion, it is safe to assume that the experts, in this case, will use significantly better equipment that has much lower systematic error associated with it. This corresponds to a narrower distribution!" Another student considering the single-slit scenario said, "considering that I think the variability results more so from the quality of the equipment, perhaps the best possible equipment that the experts are using may result in a narrower and more precise distribution." Many students also used limitations to justify why the distribution would not be a single value. For example, a student considering the projectile motion scenario said, "The best possible equipment will still have error associated with it, not possible to reduce the error to zero, will always have random fluctuations. Hence, the distribution will still be a Gaussian and with more measurements taken, would become narrower."

Similarly, an interesting minority of students used principles when explaining the distribution becoming more narrow (most commonly alongside limitations). These principles responses often were to justify the choice of the distribution becoming more narrow rather than a single value. For example, one student considering the projectile motion scenario said, "This minimized a few sources of uncertainty, but there are still some that can't be erased completely. Physics is never completely deterministic or perfect." Another student considering the Brownian motion scenario said, "Experts can reduce some of the sources of the distribution but not others so that would make the distribution narrower. However, I think the primary source of the distribution (random Brownian motion) cannot be reduced so the distribution would not be narrowed as much as above." We also consider such responses to be expertlike because the students appropriately identify that these systems have an inherently limiting uncertainty.

For the smaller fraction who indicated the distribution would stay the same, many argued using principles. As with the more data explanations, students argued that, for the Stern-Gerlach, Brownian motion, and single-slit distributions, the variability in the system was inherent to the physical system and could not be removed, even by experts with expert equipment. For example, a student considering the Stern-Gerlach scenario said, "The imprecise measurement doesn't play a role here. The distribution of values obtained comes from the fundamental randomness of the superposition breaking into one of the spin eigenstates. It doesn't have to do with equipment and expertise." Another student considering the single-slit scenario said, "The thing here is, that because of quantum randomness—it should stay the same!"

We again see relatively little variation between experiments on the explanation codes, though projectile motion appears to be distinct from the other three scenarios. For students who indicated the distribution would become more narrow, a smaller proportion of students used principles explanations for projectile motion as compared to the

Stern-Gerlach, Brownian motion, and single-slit experiments (which are otherwise similar in proportions). For students who indicated the distribution would stay the same, we see minor differences in the relative proportions of statistics and limitations codes among the Stern-Gerlach, Brownian motion, and single-slit experiments, with the major reasoning being tied to inherent principles of the physical systems. Because so few students indicated that the distribution for the Projectile motion experiment would stay the same, we can say little about their explanations.

IV. DISCUSSION

In this study, we evaluated students' perceptions of what would happen to hypothetical distributions of measurements from four different physics experiments under two settings: collecting more data (100 more students with the same equipment) or collecting better data (experts with the best possible equipment). We identified whether students thought the distributions would become wider, stay the same, become more narrow, or result in a single value and evaluated their reasoning for their choice.

A primary motivation for this analysis was to shed new light on previous understandings of student thinking about measurement; namely, the point and set paradigms [4] that classify students in terms of either thinking about experimental results as individual measurements or distributions of measurements. Our expectation was that students exhibiting setlike thinking would select that any of the distributions would either become more narrow, stay the same, or become wider with "more" or "better" data. We similarly expected that students exhibiting pointlike thinking would select that the distributions would result in a single value (represented by a delta function) in response to the better data question. In our data, very few students indicated the distribution would result in a single value: around 10% in the better data question for the projectile motion experiment, fewer than 5% in the better data question for the other experiments, and effectively 0% in the more data questions across experiments.

This result inspires multiple plausible explanations. First, perhaps, very few students in our sample were exclusively pointike thinkers. This explanation is supported by work across several institutions that claim that exclusive pointlike thinking, as measured by the Physics Measurement Questionnaire, is quite rare [8,10,11,26]. Second, pointlike thinking may be more nuanced than expecting that one can measure exactly the "true value" under ideal conditions. Many students who thought the distributions would become more narrow explicitly commented on the distribution centering on the true value but qualified their answer with the idea that uncertainty and errors could never be completely eliminated. Given that pointlike thinking focuses on true values and setlike thinking focuses on distributions [4], one might argue that this line of reasoning could be consistent with either type of thinking. We argue here, therefore, that our survey questions (and their associated responses) provide a new perspective on what it means to be a pointlike or a setlike thinker.

More in line with setlike thinking, the majority of students across questions and experiments indicated the distributions would either stay the same or become more narrow, but the relative proportions differed for the more and better data questions. The explanations for their choices also differed between the more and better data questions, as well as between experimental scenarios.

On the more data question, most students appropriately expected the distribution would stay the same, with the primary justification being that the additional data did not affect the limitations of the experiment or that additional data would not change the standard deviation. A minority of students justified this response with the inherent principles of the physical or measurement processes, although this explanation was much less frequent in the projectile motion scenario than in the other experiments.

In contrast, students who indicated the distribution would become more narrow primarily argued through purely statistical reasoning. We suggest that students were applying a "more data is better" heuristic or a form of a phenomenological primitive [27]. Previous work in statistics education has found that students often use analysis procedures as rote algorithms [28]. Plausibly, the rote algorithm associated with collecting many data points may have been internalized through the heuristic that more data is better. Under this heuristic, students appropriately drew on one of the key ideas behind setlike thinking: one needs multiple data points to estimate any phenomenon [4] and, indeed, we become much more confident in our estimate of a parameter (e.g., the mean of the distribution) with many more data points. This idea may have become overgeneralized, however, such that students inferred that the distribution itself must become more narrow. This explanation is particularly compelling given that the explanations coded for statistics used purely statistical reasoning; they were not considering the physical situation. In many cases, students also stated explicitly that the standard deviation would become smaller. Future work should evaluate whether students are making this claim through the heuristic described above or through short-circuited mathematical reasoning. For example, the division by N in the equation for standard deviation may lead students to infer that a larger N makes the standard deviation smaller (ignoring the added terms in the summation in the numerator). Alternatively, students may simply be confusing the standard deviation with the standard uncertainty of the mean (with its extra division by the square root of N).

On the better data question, most students appropriately expected the distribution would become more narrow, with almost all of the codeable responses reflecting an improvement in the limitations of the experiment. A small minority of students commented on the principles of the physical systems to justify the distribution becoming more narrow

but not reaching a single value. For the smaller fraction of students who suggested the distributions would stay the same, the primary reasoning was the inherent principles of the physical or measurement processes in the Stern-Gerlach, Brownian motion, and single-slit scenarios.

In contrast to our previous work [17], these data do not show a clear split between student thinking across classical (projectile motion and Brownian motion) versus quantum (Stern-Gerlach and single-slit) mechanical scenarios. Nor do the data show a clear split between single-value deterministic (projectile motion and Stern-Gerlach) and probabilistic distribution (Brownian motion and single-slit) experiments. Instead, we see that students responded differently to the projectile motion experiment than to the other three experiments, such that reasoning among the other three experiments was quite similar.

This result is nontrivial, particularly when considering the Stern-Gerlach scenario. Theoretically, one could reasonably evaluate the two single-value deterministic experiments (projectile motion and Stern-Gerlach) in similar ways. That is, based on physical principles alone, experts with expert equipment would approach a single-value distribution for the Stern-Gerlach experiment in much the same way as in the projectile motion experiment. The reduction in limitations should make the distribution more narrow and, potentially, single valued. The existence of quantum mechanical principles does not justify the distribution staying the same for the Stern-Gerlach experiment and yet, a higher proportion of students expected the distribution would stay the same for the Stern-Gerlach experiment than for the projectile motion experiment, with either principles or limitations justifying that choice.

We infer, therefore, that many students are seeing a fundamental distinction between the principles of classical and quantum mechanics for single-valued experiments, but not for distribution experiments. Compared to the two single-value deterministic experiments, students considered the effect of more and better data on the two distribution experiments (Brownian motion and single-slit) in similar ways. While previous research has suggested that students carry over much of their thinking about classical mechanics to quantum mechanics [18–21], our results show that this carryover may be context dependent.

A. Implications for instruction

We argue that these results motivate two important lessons for instruction. First, from the more data results, the prevalence of the statistics code as justification for the distribution becoming more narrow suggests that lab instruction should more explicitly address the applicability of the more data is better heuristic in multiple experimental scenarios. One might address the heuristic through a more careful consideration of the difference between standard deviation and standard uncertainty of the mean, with explicit learning goals and activities focused on this distinction. Such instruction should attend to the productiveness of the more data is better heuristic (building on students' existing and productive resources [29,30]), while distinguishing the uncertainty in individual measurements from the uncertainty in the mean of multiple measurements. Given that the heuristic was equally present across experimental scenarios, a purely statistical treatment may be sufficient. Our companion work evaluating students' characterizations of the sources of uncertainty [14], however, motivates the need for instruction that links this heuristic to the physical and measurement properties of the experiment as well, such as by encouraging students to consider the sources of uncertainty and whether more data will reduce the effect of those sources. This treatment is supported by the large proportion of students already attending to the limitations of the experiment in arguing for how more data do not change the distribution.

Second, from the better data results, the prevalence of the principles code as justification for the distribution staying the same suggests that theoretical quantum mechanics instruction should more explicitly address the experimental limitations in quantum mechanical measurement as distinct from the uncertainty resulting from the principles of quantum mechanics. This recommendation is further supported by our companion work evaluating students' characterizations of the sources of uncertainty [14], which similarly finds that students strongly attend to the quantum mechanical effects of the Stern-Gerlach experiment, even when they do not explain the observed variability from the experiment. For the students who expected the distributions to stay the same, they largely did so based on physical principles with little to no consideration for the limitations of the experiments. While an experiment may reach a "quantum limit," students should be provided with experience in analyzing when that limit might be reached.

B. Implications for future work

This study includes multiple limitations and open questions that inspire further study. As with any physics education research study, our finite sample size motivates replication studies with larger datasets, particularly recruiting from primarily undergraduate institutions and minority-serving institutions. Data should also be collected from introductory and advanced students for the projectile motion scenario to understand the ways in which student thinking about that scenario may evolve over time. Given the high reliability of the coding scheme, we are intrigued by the possibility of natural language processing to parse the student explanations to facilitate much larger datasets, which has been recently carried out using the Physics Measurement Questionnaire [31]. Future work should also consider additional experimental scenarios to evaluate the generalizability of our claims regarding the role of physics paradigm and theoretical expected outcome. For example, how might students

evaluate the role of more or better data in various biological systems?

Future work should also seek to disentangle why student reasoning about the projectile motion scenario is distinct from reasoning about the other scenarios. To that effect, one direction would be to simply collect more data (pun intended) from the upper-level scenarios (Brownian motion, single-slit, and Stern-Gerlach). Another direction would be to provide upper-division students with two of the upper-level scenarios and compare their reasoning directly. This analysis would test whether the differences observed for projectile motion are due to our having students complete the projectile motion scenario and one of the other scenarios, such that their responses to the second scenario are all comparative to the projectile motion scenario (and education research makes clear the power of contrasting cases [e.g., [32–39]]. What would the results look like if students had not seen the projectile motion scenario? Future work could also test the robustness of this reasoning to other introductory-level classical experiments, such as pendulum motion or masses on springs.

Given our interpreted contrasts to student thinking about sources of uncertainty and the point and set paradigms, our future work will look at the relationships between student thinking on different types of uncertainty questions. For example, does a student who draws on limitations when considering the role of more data primarily list limitations as the sources of variability? Or does one survey question prompt different reasoning than another? Similarly, how do students' explanations regarding the role of more or better data compare to their reasoning on the Physics Measurement Questionnaire? With a larger dataset, we would be better able to break down student reasoning across these various categories.

Finally, research should evaluate how instruction in both lab and theory courses impacts students' reasoning across these items and contexts.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Grants No. DUE-1808945 and No. DUE-1809178 and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139899. We are grateful to Courtney White for her initial work on this project, our project evaluator, Ben Zwickl, for fruitful discussions of this work and to Peter Lepage and the Cornell Physics Education Research Lab for comments and feedback on this work over the last four years.

APPENDIX: RAW DATA TABLES

Participants' demographic data are presented in full in Table II. We also provide the data corresponding to the results figures in the main text in Tables III–V. In each case, we include the total number of responses in each category, rather than percentages.

TABLE II. Demographic information for the students considered in this analysis.

Institution	No. of students
California State University Fullerton	8
Cornell University	83
Michigan State University	30
University of Colorado Boulder	26
University of St. Andrews	3
Year of college	
Second year (sophomore)	9
Third year (junior)	75
Fourth year + (senior)	49
Graduate student	11
Unspecified	6
Gender	
Female	40
Male	104
Nonbinary	2
Unspecified	4
Race or ethnicity	
American Indian or Alaska Native	1
Asian or Asian American	44
Black or African American	2
Hispanic or Latinx	18
Native Hawaiian or other Pacific Islander	1
Prefer to self-describe	4
White	78
Unspecified	16

TABLE III. Raw frequency data corresponding to Fig. 3.

Question	Scenario	Response	Frequency
More	Projectile motion	Single value	0
More	Projectile motion	More narrow	64
More	Projectile motion	Same	81
More	Projectile motion	More wide	5
More	Stern-Gerlach	Single value	0
More	Stern-Gerlach	More narrow	21
More	Stern-Gerlach	Same	33
More	Stern-Gerlach	More wide	6
More	Brownian motion	Single value	0
More	Brownian motion	More narrow	10
More	Brownian motion	Same	40
More	Brownian motion	More wide	8
More	Single-slit	Single value	1
More	Single-slit	More narrow	16
More	Single-slit	Same	38
More	Single-slit	More wide	5
Better	Projectile motion	Single value	22
Better	Projectile motion	More narrow	114
Better	Projectile motion	Same	14
Better	Projectile motion	More wide	0
Better	Stern-Gerlach	Single value	4
Better	Stern-Gerlach	More narrow	31
Better	Stern-Gerlach	Same	19
Better	Stern-Gerlach	More wide	6
Better	Brownian motion	Single value	1
Better	Brownian motion	More narrow	32
Better	Brownian motion	same	25
Better	Brownian motion	More wide	2
Better	Single-slit	Single value	3
Better	Single-slit	More narrow	29
Better	Single-slit	Same	26
Better	Single-slit	More wide	2

TABLE IV. Raw frequency data associated with Fig. 4.

Scenario	Response	Code	Frequency
Projectile motion	More narrow	Blank	6
Projectile motion	More narrow	Statistics	43
Projectile motion	More narrow	Principles	3
Projectile motion	More narrow	Principles + Limitations	0
Projectile motion	More narrow	Limitations	11
Projectile motion	More narrow	Other	1
Projectile motion	More wide	Blank	1
Projectile motion	More wide	Statistics	2
Projectile motion	More wide	Principles	0
Projectile motion	More wide	Principles + Limitations	0
Projectile motion	More wide	Limitations	1
Projectile motion	More wide	Other	1
Projectile motion	Same	Principles	3
Projectile motion	Same	Blank	8
Projectile motion	Same	Principles + Limitations	0
Projectile motion	Same	Statistics	30

(Table continued)

TABLE IV. *(Continued)*

Scenario	Response	Code	Frequency
Projectile motion	Same	Other	7
Projectile motion	Same	Limitations	33
Projectile motion	Single value	Principles	0
Projectile motion	Single value	Blank	0
Projectile motion	Single value	Principles + Limitations	0
Projectile motion	Single value	Statistics	0
Projectile motion	Single value	Other	0
Projectile motion	Single value	Limitations	0
Stern-Gerlach	More narrow	Other	1
Stern-Gerlach	More narrow	Principles	1
Stern-Gerlach	More narrow	Blank	4
Stern-Gerlach	More narrow	Principles + Limitations	0
Stern-Gerlach	More narrow	Statistics	13
Stern-Gerlach	More narrow	Limitations	2
Stern-Gerlach	More wide	Other	2
Stern-Gerlach	More wide	Principles	1
Stern-Gerlach	More wide	Blank	1
Stern-Gerlach	More wide	Principles + Limitations	0
Stern-Gerlach	More wide	Statistics	2
Stern-Gerlach	More wide	Limitations	0
Stern-Gerlach	Same	Other	0
Stern-Gerlach	Same	Blank	6
Stern-Gerlach	Same	Statistics	6
Stern-Gerlach	Same	Principles	6
Stern-Gerlach	Same	Principles + Limitations	1
Stern-Gerlach	Same	Limitations	14
Stern-Gerlach	Single value	Other	0
Stern-Gerlach	Single value	Blank	0
Stern-Gerlach	Single value	Statistics	0
Stern-Gerlach	Single value	Principles	0
Stern-Gerlach	Single value	Principles + Limitations	0
Stern-Gerlach	Single value	Limitations	0

Scenario	Response	Code	Frequency
Brownian motion	More narrow	Statistics	6
Brownian motion	More narrow	Other	0
Brownian motion	More narrow	Blank	3
Brownian motion	More narrow	Limitations	1
Brownian motion	More narrow	Principles	0
Brownian motion	More narrow	Principles + Limitations	0
Brownian motion	More wide	Statistics	0
Brownian motion	More wide	Other	0
Brownian motion	More wide	Blank	0
Brownian motion	More wide	Limitations	3
Brownian motion	More wide	Principles	5
Brownian motion	More wide	Principles + Limitations	0
Brownian motion	Same	Statistics	14
Brownian motion	Same	Limitations	10
Brownian motion	Same	Other	1
Brownian motion	Same	Principles	9
Brownian motion	Same	Blank	6
Brownian motion	Same	Principles + Limitations	0
Brownian motion	Single value	Statistics	0
Brownian motion	Single value	Limitations	0

(Table continued)

TABLE IV. *(Continued)*

Scenario	Response	Code	Frequency
Brownian motion	Single value	Other	0
Brownian motion	Single value	Principles	0
Brownian motion	Single value	Blank	0
Brownian motion	Single value	Principles + Limitations	0
Single-slit	More narrow	Statistics	8
Single-slit	More narrow	Principles + Limitations	0
Single-slit	More narrow	Limitations	2
Single-slit	More narrow	Other	0
Single-slit	More narrow	Principles	1
Single-slit	More narrow	Blank	5
Single-slit	More wide	statistics	2
Single-slit	More wide	Principles + Limitations	0
Single-slit	More wide	Limitations	2
Single-slit	More wide	Other	0
Single-slit	More wide	Principles	1
Single-slit	More wide	Blank	0
Single-slit	Same	Principles + Limitations	0
Single-slit	Same	Statistics	15
Single-slit	Same	Other	0
Single-slit	Same	Blank	6
Single-slit	Same	Limitations	12
Single-slit	Same	Principles	5
Single-slit	Single value	Principles + Limitations	0
Single-slit	Single value	Statistics	0
Single-slit	Single value	Other	0
Single-slit	Single value	Blank	1
Single-slit	Single value	Limitations	0
Single-slit	Single value	Principles	0

TABLE V. Raw frequency data associated with Fig. 5.

Scenario	Response	Code	Frequency
Projectile motion	More narrow	Blank	10
Projectile motion	More narrow	Statistics	3
Projectile motion	More narrow	Principles	1
Projectile motion	More narrow	Principles + Limitations	3
Projectile motion	More narrow	Limitations	81
Projectile motion	More narrow	Other	16
Projectile motion	More wide	Blank	0
Projectile motion	More wide	Statistics	0
Projectile motion	More wide	Principles	0
Projectile motion	More wide	Principles + Limitations	0
Projectile motion	More wide	Limitations	0
Projectile motion	More wide	Other	0
Projectile motion	Same	Principles	0
Projectile motion	Same	Blank	6
Projectile motion	Same	Principles + Limitations	0
Projectile motion	Same	Statistics	1
Projectile motion	Same	Other	3
Projectile motion	Same	Limitations	4

(Table continued)

TABLE V. *(Continued)*

Scenario	Response	Code	Frequency
Projectile motion	Single value	Principles	0
Projectile motion	Single value	Blank	2
Projectile motion	Single value	Principles + Limitations	0
Projectile motion	Single value	Statistics	1
Projectile motion	Single value	Other	2
Projectile motion	Single value	Limitations	17
Stern-Gerlach	More narrow	Other	0
Stern-Gerlach	More narrow	Principles	0
Stern-Gerlach	More narrow	Blank	6
Stern-Gerlach	More narrow	Principles + Limitations	3
Stern-Gerlach	More narrow	Statistics	1
Stern-Gerlach	More narrow	Limitations	21
Stern-Gerlach	More wide	Other	1
Stern-Gerlach	More wide	Principles	2
Stern-Gerlach	More wide	Blank	1
Stern-Gerlach	More wide	Principles + Limitations	0
Stern-Gerlach	More wide	Statistics	1
Stern-Gerlach	More wide	Limitations	1
Stern-Gerlach	Same	Other	0
Stern-Gerlach	Same	Blank	4
Stern-Gerlach	Same	Statistics	0
Stern-Gerlach	Same	Principles	9
Stern-Gerlach	Same	Principles + Limitations	0
Stern-Gerlach	Same	Limitations	6
Stern-Gerlach	Single value	Other	1
Stern-Gerlach	Single value	Blank	0
Stern-Gerlach	Single value	Statistics	0
Stern-Gerlach	Single value	Principles	0
Stern-Gerlach	Single value	Principles + Limitations	0
Stern-Gerlach	Single value	Limitations	3

Scenario	Response	Code	Frequency
Brownian motion	More narrow	Statistics	3
Brownian motion	More narrow	Other	1
Brownian motion	More narrow	Blank	4
Brownian motion	More narrow	Limitations	16
Brownian motion	More narrow	Principles	1
Brownian motion	More narrow	Principles + Limitations	7
Brownian motion	More wide	Statistics	0
Brownian motion	More wide	Other	0
Brownian motion	More wide	Blank	1
Brownian motion	More wide	Limitations	0
Brownian motion	More wide	Principles	1
Brownian motion	More wide	Principles + Limitations	0
Brownian motion	Same	Statistics	2
Brownian motion	Same	Limitations	1
Brownian motion	Same	Other	0
Brownian motion	Same	Principles	13
Brownian motion	Same	Blank	6
Brownian motion	Same	Principles + Limitations	3
Brownian motion	Single value	Statistics	1
Brownian motion	Single value	Limitations	0
Brownian motion	Single value	Other	0
Brownian motion	Single value	Principles	0
Brownian motion	Single value	Blank	0

(Table continued)

TABLE V. *(Continued)*

Scenario	Response	Code	Frequency
Brownian motion	Single value	Principles + Limitations	0
Single-slit	More narrow	Statistics	3
Single-slit	More narrow	Principles + Limitations	3
Single-slit	More narrow	Limitations	13
Single-slit	More narrow	Other	2
Single-slit	More narrow	Principles	3
Single-slit	More narrow	Blank	5
Single-slit	More wide	Statistics	0
Single-slit	More wide	Principles + Limitations	0
Single-slit	More wide	Limitations	0
Single-slit	More wide	Other	0
Single-slit	More wide	Principles	1
Single-slit	More wide	Blank	1
Single-slit	Same	Principles + Limitations	3
Single-slit	Same	Statistics	3
Single-slit	Same	Other	0
Single-slit	Same	Blank	6
Single-slit	Same	Limitations	3
Single-slit	Same	Principles	11
Single-slit	Single value	Principles + Limitations	0
Single-slit	Single value	Statistics	0
Single-slit	Single value	Other	0
Single-slit	Single value	Blank	2
Single-slit	Single value	Limitations	1
Single-slit	Single value	Principles	0

[1] P. Heron and L. E. McNeil, Phys21: Preparing physics students for 21st century careers, American Physical Society, Technical Report, 2016.

[2] American Association of Physics Teachers, AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum, Technical Report, 2014.

[3] N. G. Holmes and E. M. Smith, Instructional strategies that foster experimental physics skills, in *International Handbook of Physics Education Research: Learning Physics*, edited by M. F. Taşar and P. R. L. Heron (AIP Publishing, Melville, New York, 2023), pp. 18–1–18–20.

[4] 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).

[5] 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).

[6] S. Allie, A. Buffler, F. Lubben, and B. Campbell, Point and set paradigms in students' handling of experimental measurements, in *Research in Science Education—Past, Present, and Future* (Kluwer Academic Publishers, Dordrecht, 2001), pp. 331–336.

[7] F. Lubben, S. Allie, and A. Buffler, *Experimental Work in Science* (Springer, Netherlands, 2010), pp. 135–152.

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

[9] R. L. Kung, Analyzing students' use of metacognition during laboratory activities, in *Proceedings of AREA Meeting* (American Educational Research Association, New Orleans, LA, 2002).

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

[12] J. Leach, R. Millar, J. Ryder, M.-G. Séré, D. Hammelev, H. Niedderer, and V. Tselfes, Survey 2: Students' images of science as they relate to labwork learning, Centre for Studies in Science and Mathematics Education, Technical Report, 1998.

[13] S. Allie, A. Buffler, B. Campbell, and F. Lubben, First-year physics students' perceptions of the quality of experimental measurements, Int. J. Sci. Educ. **20**, 447 (1998).
[14] E. M. Stump, M. Dew, G. Passante, and N. G. Holmes, Context affects student thinking about sources of uncertainty in classical and quantum mechanics (to be published).
[15] M. M. Stein, C. White, G. Passante, and N. G. Holmes, Student interpretations of uncertainty in classical and quantum mechanics experiments, presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Stein.
[16] E. M. Stump, C. White, G. Passante, and N. Holmes, Student reasoning about sources of experimental measurement uncertainty in quantum versus classical mechanics, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.Stump.
[17] C. White, E. M. Stump, N. Holmes, and G. Passante, Student evaluation of more or better experimental data in classical and quantum mechanics, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.White.
[18] C. Baily and N. D. Finkelstein, Development of quantum perspectives in modern physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010106 (2009).
[19] R. Steinberg, M. C. Wittmann, L. Bao, and E. F. Redish, *The influence of student understanding of classical physics when learning quantum mechanics*, Research on Teaching and Learning Quantum Mechanics (National Association for Research in Science Teaching, Boston, MA, 1999).
[20] E. Marshman and C. Singh, Framework for understanding the patterns of student difficulties in quantum mechanics, Phys. Rev. ST Phys. Educ. Res. **11**, 020119 (2015).
[21] K. Krijtenburg-Lewerissa, H. J. Pol, A. Brinkman, and W. R. van Joolingen, Insights into teaching quantum mechanics in secondary and lower undergraduate education, Phys. Rev. Phys. Educ. Res. **13**, 010109 (2017).
[22] B. M. Zwickl, D. Hu, N. Finkelstein, and H. J. Lewandowski, Model-based reasoning in the physics laboratory: Framework and initial results, Phys. Rev. ST Phys. Educ. Res. **11**, 020113 (2015).
[23] J. Cohen, The earth is round ($p < .05$)., Am. Psychol. **49**, 997 (1994).
[24] G. Cumming, The new statistics: Why and how, Psychol. Sci. **25**, 7 (2013).
[25] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, The preregistration revolution, Proc. Natl. Acad. Sci. USA **115**, 2600 (2018).
[26] R. F. Lippmann, Students' understanding of measurement and uncertainty in the physics laboratory: Social construction, underlying concepts, and quantitative analysis, Ph.D. thesis, The University of Maryland, College Park, MD, 2003.
[27] A. A. DiSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[28] A. Bakker and J. Derry, Lessons from inferentialism for statistics education, Math. Think. Learn. **13**, 5 (2011).
[29] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).
[30] A. D. Robertson, L. C. Bauman, Y. M. Abraham, B. Hansen, H. Tran, and L. M. Goodhew, Resources-oriented instruction: What does it mean, and what might it look like?, Am. J. Phys. **90**, 529 (2022).
[31] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, Phys. Rev. Phys. Educ. Res. **18**, 010141 (2022).
[32] J. D. Bransford, J. J. Franks, N. J. Vye, and R. D. Sherwood, New approaches to instruction: Because wisdom can't be told, in *Similarity and Analogical Reasoning*, edited by S. Vosniadou and A. Ortony (Cambridge University Press, Cambridge, England, 1989), pp. 470–497.
[33] D. L. Schwartz and J. D. Bransford, A time for telling, Cognit. Instr. **16**, 475 (1998).
[34] D. L. Schwartz and T. Martin, Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction, Cognit. Instr. **22**, 129 (2004).
[35] C. C. Chase, J. T. Shemwell, and D. L. Schwartz, Explaining across contrasting cases for deep understanding in science: An example using interactive simulations, in *Proceedings of the 9th International Conference of the Learning Sciences, ICLS 2010* (International Society of the Learning Sciences, Chicago, IL, 2010), pp. 153–160.
[36] D. L. Schwartz, C. C. Chase, M. A. Oppezzo, and D. B. Chin, Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer, J. Educ. Psychol. **103**, 759 (2011).
[37] M. T. Chi, I. Dohmen, J. T. Shemwell, D. B. Chin, C. C. Chase, and D. L. Schwartz, Seeing the forest from the trees: A comparison of two instructional models using contrasting cases, in *Proceedings of the American Educational Research Conference, Vancouver, BC* (American Educational Research Association, Vancouver, BC, 2012).
[38] J. Roelle and K. Berthold, Effects of Comparing Contrasting Cases on Learning From Subsequent Explanations, Cognit. Instr. **33**, 199 (2015).
[39] A. B. Heim, C. Walsh, D. Esparza, M. K. Smith, and N. G. Holmes, What influences students' abilities to critically evaluate scientific investigations?, PLoS One **17**, e0273337 (2022).