

Context of authority may affect students' evaluations of measurement

Emily M. Smith,^{1,2} Nicole Chodkowski,³ and N. G. Holmes¹

¹*Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca NY 14853*

²*Department of Physics, Colorado School of Mines, Golden CO 80401*

³*School of Integrative Plant Science & L.H. Bailey Hortorium, Cornell University, Ithaca, NY 14853*

Recent research in introductory physics labs suggests that most students judge the quality of a measurement based on a comparison with theory. To probe this dimension of students' judgments based on authority, we sought to evaluate whether students' responses about evaluations of measurement depended on contextual cues. We asked students which measurement of the acceleration due to gravity was 'better:' (1) one given with uncertainty and found by 'you and your friend' or 'you and your research group' or (2) a textbook value with no reported uncertainty but more significant figures. By deliberately structuring multiple possible forms of authority (e.g., precision, expertise, equipment, theory) we intended to draw out nuances in how students draw upon authority in evaluating the quality of measurements. Our results suggest that contextual cues may influence students' judgments about measurement and the authority that they draw upon more than lab instruction aimed at developing students' experimentation skills.

I. INTRODUCTION

Physics is an experimental science and so common conceptual knowledge introduced in a physics course should include understanding about measurements and the process of experimentation. In this paper, we focus on the role of authority in students' understandings of measurements, gauged from their comparisons of measurements embedded in different contexts.

Previous work has broadly categorized students' reasoning about measurements into two paradigms: point and set [1, 2]. In *point* reasoning, measurements are viewed such that any measurement could be measuring the 'true' value. To contrast, in *set* reasoning, measurements are viewed as an approximation that is obtained by several trials that form a distribution around the 'true' value. The Physics Measurement Questionnaire was specifically developed to probe these broad ways of reasoning about measurement [1–3]. Research with the questionnaire has demonstrated that many students tend toward presenting mixed reasoning—perhaps due to contextual cues—that includes elements from both point and set reasoning [1, 2, 4–8].

Recent research suggests that many introductory students judge the quality of a measurement based on a comparison with theory [9]. These results suggest that students' ideas about measurement may depend on forms of authority in experimentation. To begin to probe the interactions between point and set reasoning and authority, we sought to evaluate the role of authority in students' ideas about measurement by deliberately structuring multiple possible forms of authority (e.g., precision, expertise, equipment, theory) within a single prompt.

We asked students to choose which of two measurements was 'better' and to explain their reasoning. Within the same population of students in a first-semester college physics lab, we adjusted contextual cues to test the role of authority in students' evaluations of the measurements, though the measurement values were unchanged. Additionally, a subset of the students provided responses after an academic year of lab instruction, allowing us to evaluate the impact on student thinking due to instruction. Our results suggest that in this prompt, explicit point reasoning is rare and that the context of authority in measurement plays a greater role in students' judgments than instruction.

II. DATA AND METHODS

The data in this study are from two sequential introductory courses that are aimed at prospective physics majors and minors. Students who enroll tend to be highly prepared; most have taken a college-level physics course in high school, such as Advanced Placement Physics, and all were expected to be concurrently enrolled in a multi-variable calculus course at the beginning of the course sequence. Typical demographics include about 20% women and 80% men, 10% students from underrepresented racial

or ethnic groups in physics, and 50% students who intend to pursue physics majors.

The lab course required students to make decisions about the design, execution, and extension of their experiments. Quantitative tools were introduced throughout two semesters, including (in order of introduction): (1) t' (compare two measurements with uncertainty), (2) weighted χ^2 (to check consistency of data with a model), (3) linearization through semi-log and log-log plots (to identify exponential and power law models from data), and (4) propagation of uncertainty (to calculate quantities using measurements).

We measured students' reasoning about measurements before and after instruction. Data were collected from students' individually completed prelab assignments. The pre-response was collected prior to the first lab activity of the first introductory course and prompted students to use their own reasoning and intuition rather than outside sources. The post-response was due concurrent with the final lab activity of the second introductory course, leaving nearly an entire academic year between students' responses. Cohort I provided both pre- and post-responses to the same version of the prompt. Cohort II provided only a pre-response.

A. Prompt design

We expected that students developed ideas about how to compare and evaluate the quality of measurements prior to college physics labs. Therefore, we constructed three prompts to probe characteristics of students' comparisons of measurements of the same quantity and their corresponding reasoning. From students' answers to the prompts in early semesters, we suspected that contextual cues of the prompts—especially the role of authority—were affecting students' answers, so we made slight adjustments to these cues in a subsequent semester.

In this paper, we focus on one of the prompts (including the changes in the contextual cues) that is provided in Table I. There is no single correct (nor incorrect) answer, and we have not collected responses from experts. We report characteristics of students' responses to the prompt to discuss how contextual cues and lab experiences may affect students' evaluations of measurements.

We use italics to distinguish the two measurements in the prompt: (1) *(own) measurement* refers to the measurement done by 'you and your friend' or 'you and your research group' and (2) *textbook* refers to the value listed in the textbook. The students' *own measurement* of the acceleration due to gravity was intentionally selected to be lower than the value in the *textbook*. This selection ensured that the *measurement* was farther from 9.81 m/s^2 (the 'canonical value,' which we assumed students were familiar with) and the value listed on a modern source (e.g., Wolfram Alpha) than the *textbook* value. To test the role of context in students' evaluations of the measurements, we adjusted the forms of authority of the *measurement* for Cohort II to include expertise ('you and

TABLE I. Prompt used to characterize students' sense of authority in measurement.

Cohort I	Cohort II
You and your friend decide to measure the acceleration due to gravity in Ithaca. You measure the acceleration due to gravity to be $g=9.790 \pm 0.005 \text{ m/s}^2$. However, when you look up the value for Ithaca, you find that it is listed in a textbook as $g=9.8029 \text{ m/s}^2$ (W.D. Henderson, <i>Physics in Everyday Life</i> , p. 27, 1921). Which measurement do you think is better? Explain your reasoning.	You join a research group at Cornell that decides to measure the acceleration due to gravity in Ithaca with brand new, top-of-the-line equipment. You and your research group measure the acceleration due to gravity to be $g=9.790 \pm 0.005 \text{ m/s}^2$. However, when you look up the value for Ithaca, you find that it is listed in a textbook as $g=9.8029 \text{ m/s}^2$ (W.D. Henderson, <i>Physics in Everyday Life</i> , p. 27, 1921). Which measurement do you think is better? Explain your reasoning.

your research group at Cornell”) and equipment (“brand new, top-of-the-line equipment”), compared to Cohort I where the authority was in “you and your friend.”

B. Analysis of responses

We combined and anonymized responses for both cohorts prior to coding for *judgment* and *reasoning*, and N.C. was not made aware of the conditions. E.M.S. emergently coded for the *judgments* that students made in the situations (i.e., whether the students' *own measurement* or *textbook* value was judged to be better), and E.M.S. and N.C. separately and emergently coded for the *reasoning* to support judgments (i.e., characteristics of the explanation that support a judgment). Following our separate coding, we discussed the *reasoning* codes and settled on a scheme that accommodated both raters' codes, in App. A. Then, we independently applied the new coding scheme (both *judgments* and *reasoning*) to a common subset of the data that included responses from each condition. Cohen's kappa was 0.92 for *judgments* and 0.87 for *reasoning*, so E.M.S. independently coded the remaining responses. We then identified students and matched pre- and post-responses for Cohort I.

C. Examples of coding

Here, we provide examples to show how responses were coded. These responses were selected because they were among the most difficult to code, and therefore, provide more insight to the decisions that we made while coding than typical responses. The first example: “*I think the value listed in the textbook is better, [Judgment: textbook] because that result has been tested by scientific experts [Authoritative] with standard lab equipment, and has been peer reviewed [Authoritative] and replicated to pinpoint [Multiple confirmations] the exact value [True/exact value]. Therefore, it is reasonable to say that that value is accurate and all sources of error were accounted for, something not necessarily true for our measurement.*” We did not code for Methods/equipment because the student did not claim that standard equipment is better than student equipment; instead, the student's response suggests that people who conducted the experiment for the textbook value are experts, which is coded

as Authoritative.

A second example mentions that both measurements may be valid, however, does not provide reasoning for the alternative judgment, and so, is coded with only the judgment and reasoning in favor of the *textbook*: “*I think the textbook measurement is better, [Judgment: textbook] because it has been scrutinized over and over before, [Authoritative] however our data is valid in its own sense. It would be worthwhile to test gravity again and again given the significance of this difference, however the experimental errors could have easily led to this difference.*”

A third example is in favor of the measurement and nearly makes a case for the textbook, however, never provides an explicit statement that the textbook value may be a preferable selection. The final example: “*I would think that the measurement taken from the lab is more accurate [Judgment: Own measurement] because since the book is published in 1921, our methods and techniques for collecting data must have improved since then [Methods/equipment]. Therefore it is more reasonable to trust more recent data with newer technology. Also, even though the value in the textbook has more decimal places indicating more precision, I think it is more reasonable to take a measurement with justified random error [Includes uncertainty] than using a value that might be inaccurate.*” The student's comment about the precision of the value in the textbook was not coded as reasoning because it implied that there was an advantage to the textbook value but never explicitly said that value could be judged to be better. We debated whether the Recency code applied but decided that the Methods/equipment was the driving reason why students trusted the more recent data.

III. RESULTS

Throughout this section, we refer to Table II to show frequencies of codes for responses from Cohorts I and II. Cohorts I and II received different contextual cues in the prompt (given in Table I).

A. Contextual cues affect students' judgments

The majority of students in Cohort I preferred the *textbook* value in the pre-responses. Therefore, we inferred that students tended to regard the authority of the text-

book or the people who contributed to the textbook to be greater than the authority of students obtaining their own measurements. For example, one student expressed that “*without knowing how each value was obtained, it’s hard to say. But I’d be more inclined to trust the text simply because it’s official and verified.*” Apart from outwardly describing the authority held by the textbook, students frequently expected experts to have access to better equipment or methods, despite referencing a textbook from nearly a century ago. Interestingly, many students also assumed that the value listed in the textbook underwent experimental confirmations and a minority of students brought up significant figures (found to be an influential form of authority in Ref. [10]).

Only a quarter of Cohort I judged their *own measurement* to be better than the *textbook* value. The most frequent reason for this judgment was that uncertainty is necessary in a measurement. One student succinctly summarized “*I think that my measurement of gravity in Ithaca is better because it is more recent than 1921, and also I know what the uncertainty is.*” This suggests that including uncertainty is of greater importance to some students than any authority provided by the textbook.

For Cohort II, the contextual cues were adjusted to impose greater authority on the *measurement*, and about two-thirds of Cohort II judged the *measurement* to be better than the *textbook* value. The dominant reasoning shifted to the methods or equipment available to the research group. For example, “*I think the measurement my research group conducts is more accurate, because the modern equipment I used should be more precise and accurate than the equipment used in 1921.*” Students’ shift in reasoning suggests that some students may view modern methods and equipment as having greater authority in measurements than, possibly outdated, values in textbooks. However, for students who preferred the *textbook* value, the most frequent reasons were Multiple confirmations and Significant figures (in contrast to Methods/equipment and Authoritative common among Cohort I), which suggests that students shift their reasoning in response to contextual cues.

We also identified responses that explicitly suggested a student understood a measurement to have a true or exact value. Only 13 responses were identified, and of those, 12 responses included judgments in favor of the textbook value (including *Textbook* and *Both (Textbook)* judgments) and 10 were pre-responses from Cohort I. For example, one student wrote “*I think the value listed in the textbook is better. Because when we try to measure the acceleration during experiment, the result is going to differ from the true value due to air friction, equipment deficiency and many other kinds of confounding factors. The value we get also has a high uncertainty. Therefore (sic) the value on the textbook should be better and more accurate.*” Similarly, few students in either cohort support their *textbook* judgment with Excludes uncertainty, which may be similar to research demonstrating the rarity of explicit point reasoning in other populations [7, 8].

TABLE II. Code frequency for responses to the prompt. Appendix A provides code definitions.

	Cohort I		Cohort II
	Pre (N=91)	Post (N=51)	Pre (N=31)
<i>Judgment & reasoning</i>			
<i>Own measurement</i>	27% (25)	31% (16)	68% (21)
Includes uncertainty	48% (12)	56% (9)	29% (6)
Methods/equipment	28% (7)	6% (1)	67% (14)
Varies in location	28% (7)	38% (6)	19% (4)
Recency	16% (4)	25% (4)	24% (5)
<i>Textbook</i>	55% (50)	35% (18)	26% (8)
Excludes uncertainty	8% (4)	6% (1)	13% (1)
Methods/equipment	40% (20)	50% (9)	0%
Multiple confirmations	36% (18)	50% (9)	50% (4)
Authoritative	40% (20)	50% (9)	13% (1)
Significant figures	22% (11)	17% (3)	38% (3)
<i>Both (Own)</i>	2% (2)	10% (5)	3% (1)
<i>Both (Textbook)</i>	9% (8)	16% (8)	0%
<i>None</i>	6% (5)	8% (4)	3% (1)
<i>Other</i>	1% (1)	0%	0%

		Post-assignment				
		Own measurement	Textbook	Both (Own measurement)	Both (Textbook)	None
Pre-assignment	Own measurement	9	3	2	0	0
	Textbook	5	11	1	3	1
	Both (Own measurement)	0	0	0	0	0
	Both (Textbook)	0	0	1	2	0
	None	0	0	0	1	1

FIG. 1. Judgments for matched students in Cohort I (N=40), a subset of those in Table II. Yellow corresponds to a change in judgment and green indicates consistency. To identify trends in the reasoning, we focused on students’ responses in darker shades.

B. Judgments appear to be relatively stable

Next, we compared pre- and post-responses to the same version of the prompt after an academic year of instruction for Cohort I. The frequency of students favoring the value listed in the *textbook* dropped from pre- to post-responses. Meanwhile, there was no clear gain in the percentage of students preferring their *own measurement*. Instead, a larger percentage of students provided reasoning for both judgments, which we initially interpreted to be students’ increased comfort in ambiguity. To test this, we matched responses to identify shifts in individual students’ judgments across the academic year; there were 40 matched responses. As shown in Fig. 1, the majority of students remained consistent with their judgments despite an academic year of lab instruction. This suggests that students’ judgments were, perhaps, unaffected by instruction or that we have too few responses to discern meaningful comparisons by instruction at this grain-size of analysis.

For the students who made major shifts from *own measurement* to *textbook* or vice versa, we identified the rea-

soning that they used to support their judgments. Three students chose the *measurement* on their pre-response but shifted to the *textbook* value, and all discussed the methods or equipment on post-responses. For example, one student wrote “*although my measurement has a larger confidence value due to its interval, since I likely do not have access to precision technology the value in the book is likely more accurate considering it has probably been tested using highly precise equipment and checked multiple times.*”

There was greater variation among the reasons provided for post-responses for the five students who shifted from a judgment in favor of the *textbook* to their *own measurement*. In their post-responses, three students provided reasoning that the measurement included uncertainty, and three students (with overlap) mentioned the relevancy of the location. For example, one student expressed “*the first one. It provides the error. Different locations in Ithaca might have different values so we can't just take the measurement in the book to be correct.*”

IV. DISCUSSION

These results suggest that contextual cues may influence students' judgments about measurement more than lab instruction aimed at developing experimentation skills. We tentatively claim that within this population of students, authority is ranked: Measurements by research groups > Textbook values > measurements by student groups. However, the contextual cues elicit different forms of authority in students' justifications of a judgment. Most obvious was the role of methods and equipment: Students' responses suggest that methods or equipment of an unknown expert (from a century ago) are of greater authority than those of a student group but are outdone by a research group's. This particular view of authority may be robust to lab instruction because students may view the methods and equipment of instructional labs to be different from those of expert experimentalists.

Many introductory students may claim to judge the quality of measurements based on their comparison to theory [9], but only one student's response in our data compared the measurements to theory (9.81 m/s²). Furthermore, few students identified the lack of uncertainty in the textbook value as indicating quality or used as a form of authority in making their judgment (and several indicated the inclusion of uncertainty is important). This suggests students' views of theory and comparisons of measurements and theory may also be context-dependent.

Future work will aim to further characterize the role of authority in students' understandings of measurement. How does the context of a lab activity induce students' views of authority? In what ways do students characterize their own measurements with varying precision? What reasoning do students draw upon when identifying features of their measurements?

ACKNOWLEDGMENTS

We thank Saaj Chattopadhyay whose identification of patterns in students' evaluations of measurements informed the prompt. This work was partially supported by the Cornell University Active Learning Initiative.

Appendix A: Code descriptions

The ‘reasoning’ codes that emerged from the data are structured by ‘judgment’ (italicized).

Own measurement: Reasoning provided for favoring the measurement done by the student or research group.

Includes uncertainty: The measurement includes uncertainty or the textbook does not include uncertainty, implying uncertainty is an important aspect.

Methods/equipment: The methods and/or equipment available more recently are newer, better, more sophisticated, and/or improved precision. Students may refer to better accuracy of equipment, however, we are interpreting this to indicate improved instrumental precision.

Varies in location: Measurements of gravitational acceleration vary due to location and/or geology.

Recency: Measurement is more recent and therefore better and/or measurement is taken in real-time and may provide a sense of knowing the truth.

Textbook: Reasoning provided for favoring the value listed by the textbook.

Excludes uncertainty: Textbook value excludes uncertainty or is known exactly, and/or the measurement includes uncertainty or is not known exactly.

Methods/equipment: Equipment used to measure the textbook value is better than equipment available to students and/or a research group.

Multiple confirmations: Multiple people or an expert made repeated measurements for the textbook value and/or the value has been confirmed multiple times (different people or an expert).

Authoritative: Textbooks are trustworthy (e.g., peer-reviewed, good resources, published work, etc.) and/or values in textbooks are produced or measured by professionals.

Significant figures: Significant figures and/or precision of the textbook value indicates a more reliable or precise measurement.

Both (Own): Reasoning is providing for both the measurement and value but ultimately favors the measurement done by the student and/or research group. All reasoning codes may apply.

Both (Textbook): Reasoning is providing for both the measurement and value but ultimately favors the value listed by the textbook. All reasoning codes may apply.

None: No judgment is ultimately made though may provide reasoning for one or both cases.

True/exact value: Response suggests that the student has an understanding of a measurement as having a true, exact, or real value.

-
- [1] 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).
- [2] A. Buffler and S. Allie, The development of first year physics students' ideas about measurement in terms of point and set paradigms, *Int. J. Sci. Educ.* 23(11), 1137 (2001).
- [3] A. Buffler, F. Lubben, and 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(9), 1137 (2009).
- [4] D. S. Abbott, Assessing student understanding of measurement and uncertainty, Ph.D. Thesis, North Carolina State University, Raleigh, NC (2003).
- [5] R. Lippmann, Students' understanding of measurement and uncertainty in the physics laboratory: Social construction, underlying concepts, and quantitative analysis, Ph.D. Thesis, University of Maryland, College Park, MD (2003).
- [6] R. Lippmann Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, *NorDiNa* 4, 40 (2006).
- [7] B. Pollard, R. Hobbs, J.T. Stanley, D. R. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab on students' understanding of measurement uncertainty, *Physics Education Research Conference Proceedings* p. 312–315 (2017).
- [8] H. J. Lewandowski, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and B. Pollard, Student reasoning about measurement uncertainty in an introductory lab course, *Physics Education Research Conference Proceedings* p. 244–247 (2017).
- [9] D. Hu and B. M. Zwickl, Examining students' views about validity of experiments: From introductory to Ph.D. students, *Phys. Rev. Phys. Educ. Res.* 14, 010121 (2018).
- [10] K. Kok, B. Priemer, W. Musold and A. Masnick, Students' conclusions from measurement data: The more decimal place, the better? *Phys. Rev. Phys. Educ. Res.* 15, 010103 (2019).