

Open- versus closed-response methods of measuring students' reasoning: A case study using the Physics Measurement Questionnaire

Rebeckah K. Fussell¹,¹ Jillian S. Epstein¹,¹ Gina Passante,² and N. G. Holmes¹

¹Laboratory of Atomic and Solid State Physics, *Cornell University, Ithaca, New York 14853, USA*

²Department of Physics, *California State University Fullerton, Fullerton, California 92831, USA*



(Received 9 December 2025; accepted 17 July 2026; published 15 September 2026)

We compare open- and closed-response versions of two questions that measure students' reasoning on the Physics Measurement Questionnaire, same mean different spread (SMDS), and different mean same spread (DMSS). We conducted an experiment in one class ($N = 248$) to test if asking students for their reasoning on an open-response question changes their answer to the closed-response question. To test whether answering the open-response question alters students' closed-response answers, we prompted half of the students to explain their reasoning before seeing the closed-response options, while the other half went straight to the closed-response reasoning options. We find no effect of asking students their reasoning in an open-response question on their subsequent closed-response answers. Additionally, we investigated whether we measure students' reasoning differently on the closed-response versus open-response versions of the question by evaluating the consistency between the students' open-response and closed-response answers. Although we measure similarities in students' reasoning across the two methods, the differences indicate that the open-response and closed-response questions provide different information. The results inform the need for critical judgment when deciding between open- and closed-response assessment formats.

DOI: [10.1103/z9n2-tfr9](https://doi.org/10.1103/z9n2-tfr9)

I. INTRODUCTION

Historically, education researchers have opted to use standardized assessments with closed-response questions over open-response questions in studies with large sample sizes. The primary motivation for this preference is that open-response questions require significant investments of time and effort to analyze each response, whereas responses to closed questions can be more rapidly analyzed in aggregate. The gap in the time and effort needed to analyze open-response versus closed-response data is decreasing, however, with recent innovations in social science methods that use machine learning models to scale qualitative analysis (e.g., Refs. [1–3]). As a result, there is considerable discussion in the education research community about the prospect of analyzing open-response questions more efficiently than ever before. In particular, there is excitement about leveraging large language models and other natural language processing methods that have been shown to analyze data with a level of interrater reliability comparable to human analysts (e.g., Refs. [4–6]).

The enthusiasm toward machine learning models is particularly a result of a concern that meeting the needs of 21st-century science, technology, engineering, and mathematics (STEM) students requires emphasizing more dimensions of learning than closed-response questions can assess [7]. Researchers in K-12 education, for example, have argued that tasks that ask students to construct a response provide the strongest evidence of multidimensional learning, but “a lack of resources for scoring” compels the development of closed-response assessments [8]. Indeed, a report on “Developing Assessments for the Next Generation Science Standards” notes that “selected response items will not likely be adequate for assessing the full breadth and depth of the Next Generation Science Standards (NGSS) performance expectations, particularly in assessing... scientific and engineering practices” [9]. There are, however, many closed-response assessments that exist to evaluate higher-order thinking and reasoning skills in physics, such as critical thinking [10,11], scientific reasoning [12,13], or evaluating experimental data [14].

These developments warrant a renewed discussion about closed-response and open-response questions, a discussion that has been ongoing for decades. In physics education research specifically, for example, there are many studies comparing students' performance on open-response versus closed-response formats of exams, concept inventories, and surveys [e.g., Refs. 15–19]. Two

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.

meta-analyses found that scores on closed-response questions were often higher than on open-response questions, though this gap narrowed when the two types of questions were “stem-equivalent,” that is, when the closed-response questions simply added answer choices to the same open-response question stem [20,21]. The more recent of the two meta-analyses [20] also found that scores between closed-response and open-response questions were strongly correlated, though the effect could vary based on student characteristics. The conclusions of this study focused on the policy implications as related to fairness: does the test format systematically benefit some groups over others? Related concerns also emerged about test-taking strategies (e.g., guessing).

The context of the current case study, however, is less about whether students’ scores will differ with one format or another. Rather, we are concerned about the measurements of student understanding: do the different formats provide different information about student understanding, and if so, which is the most valid approach to use?

To understand the contexts in which it is most appropriate to choose closed-response or open-response formats, we must consider what an individual survey question is intended to measure. Previous research that has directly compared open-response questions to their closed-response counterparts has varied results, suggesting that results will be dependent on the nature of what is being measured [22–24]. The types of claims that can be made about data depend on the way the question was generated, the way the data was constructed, and the context in which it was administered. Furthermore, previous studies have found that the sequence of items can influence the way students respond, particularly if the items have similar surface features [25]. Administering two versions of a question, open- and closed-response, is an intuitive way of gathering the data necessary to make a direct comparison of student responses across them, but it could alter the responses to the question asked second. As a result, in this study, we were motivated to investigate whether students select different reasoning items on closed-response questions depending on whether or not they were first asked an open-response version of the question.

We use the idea of cognitive accessibility [26], as we have in our previous work [27], to understand why open and closed-response questions measure students’ explanations differently. Cognitive accessibility refers to the relative strength with which explanations come to mind, and, as in Ref. [26], we define the relative accessibility of an explanation as the “number of times that that explanation is listed in a free-recall task.” For example, consider a question about comparing datasets, as in the Physics Measurement Questionnaire (PMQ) [28]. Students are asked whether or not two datasets agree with each other (a closed-response judgment question) and then are

asked to explain why they came to this conclusion (an open-response reasoning prompt). In answering the judgment question, students will access a subset of all the pieces of knowledge [29,30] that are “available” to them. A particular subset of this knowledge may be “accessible” depending on the context. The open-response reasoning question asks students to *recall* the explanations they used to arrive at their answer. According to Ref. [26], students will write down the explanations that are most accessible to them. Following this framework, open-response questions may be more sensitive to accessibility, whereas closed-response questions may reveal explanations that students recognize as relevant when prompted, including explanations that may have been available but not initially accessible.

Thus, the model of accessibility and availability suggests that open- versus closed-response questions likely measure different cognitive processes: open-response questions assess the pieces of knowledge accessible to students, while closed-response questions may instead provide evidence about the availability of ideas that students recognize when prompted. We must take care, then, in understanding how the proposed shift toward open-response questions as a result of machine learning opportunities may shift the nature of our measurements and the resulting claims.

Here, we reopen the discussion about how student reasoning is measured through open- and closed-response questions as part of a larger project to measure student thinking about experimental measurements. We were motivated to explore new contexts to understand and demonstrate how an open- and closed-response version of the same question affects measurements of student reasoning because we wanted to ensure we were using the best form of the question for our research. In this case study, we investigate the following:

1. Do students select different reasoning items on closed-response questions depending on whether or not they were first asked to write their reasoning in an open-response question?
2. Do we measure different student reasoning on closed-response versus open-response questions?

We base our study in the context of the PMQ [28,31,32], which we were using as part of a larger study to understand students’ thinking about measurement and uncertainty. This assessment was designed as an open-response assessment, and recent work sought to develop natural language processing algorithms to evaluate students’ responses automatically [6]. The assessment was also converted to a closed-response format [33], and other researchers have sought to develop closed-response assessments that probe similar constructs [34,35]. Our study seeks to interrogate the trade-offs between these distinct approaches, which we expect to inform discussions across assessment constructs.

II. METHODS

A. Data sources

Our survey used two questions from the PMQ [28,31,32]: same mean different spread (SMDS) (Fig. 1) and different mean same spread (DMSS) (Fig. 2). Both of these questions were designed to measure patterns in students' reasoning about measurement and uncertainty when comparing datasets. Both questions had two components: a judgment portion and a reasoning portion. In the judgment portion, students were asked to make a judgment about a pair of datasets by selecting a response from a short list of options (three options for SMDS and two options for DMSS). In the reasoning portion, students were asked about their reasoning behind the selection they made in the judgment portion.

In the PMQ, these questions refer to an experimental context that is described using text and an image. The context is an imagined scenario in which different groups of students in a lab class collect data for a projectile motion experiment where they place a ball on a ramp, the ball rolls down the ramp and off the table, and the students measure the horizontal displacement x that the ball travels. In our survey, before being presented with the two questions, the students watched a short video embedded in the survey (or read a transcript with an image) that explains the shared context for both questions.

In SMDS, the students were shown two datasets, which they were told were collected by two of the groups in the

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

Which group's measurements do you think are better?

- ☐ Group A
- ☐ Group B
- ☐ Group A and Group B have equally good measurements

FIG. 1. Judgment question for SMDS.

Release	Group D (mm)	Group E (mm)
1	440	432
2	438	444
3	433	426
4	422	433
5	432	440
Average:	433	435

Do the results of these two groups agree?

- ☐ Yes
- ☐ No

FIG. 2. Judgment question for DMSS.

imagined lab class, Group A and Group B (Fig. 1). Both datasets had five data points (from five releases of the ball). Both datasets had the same average measurement of 435 mm, but the two datasets had different spreads (a quantity not given to students, but inferable from the values). In the judgment portion, students were asked to answer “Which groups’ measurements do you think are better?” by selecting one of the following options: “Group A,” “Group B,” or “Groups A and B have equally good measurements.” After selecting their answer to the judgment question, students were asked about their reasoning.

In DMSS, the students were shown two new datasets, which they were told were collected by two more groups in the imagined lab class, Group D and Group E (Fig. 2). Both datasets had five data points (from five releases of the ball). The two groups’ data had similar spreads but different means: the average of Group D’s data was 433 mm, and the average of Group E’s data was 435 mm. In the judgment portion, the students were asked, “Do the results of these two groups agree?” and could select either “Yes” or “No.” As in SMDS, after answering the judgment question, students were asked about their reasoning.

1. Two versions of the questions

We tested both an open-response and closed-response version of the reasoning portion of the two questions. In the closed-response version, we prompted students to “Please choose 1 or 2 of the following statements that are most relevant to your reasoning for choosing [Response].” Each statement in the list of items was a potential explanation that a student might provide for making the judgment they did. These closed-response statements are presented in Table I (SMDS) and Table II (DMSS). The methods used to develop these items are described below. In the open-response version, we directly asked “Please explain your choice” and provided an open text box, as in the original administration of the PMQ. All students who received an open-response version of the question also received the closed-response version on the next page.

We analyzed the open-response answers to the reasoning portion of SMDS and DMSS by identifying which of the closed-response reasoning items were expressed in the open-response answers. This analysis is described in more detail in Sec. II C. The data for this study are available at Ref. [36].

2. Survey administration

We collected data by administering the survey as a homework assignment in an introductory experimental physics course at Cornell University. The homework assignment was given three-quarters of the way through the semester, so the students in this population had already received instruction relevant to these survey questions. We chose to administer the survey as a homework assignment to get a higher response rate compared to an optional

TABLE I. Closed-response reasoning items for the SMDS probe. Students were prompted to choose one or two of these statements that were most relevant to their reasoning for choosing “Group A,” “Group B,” or “Group A and Group B have equally good measurements.”

Group A	Group B	Group A and B
A1: Group A’s range is smaller than Group B’s range.	B1: Group B’s range is larger than Group A’s range.	AB1: The range of Group A’s measurements (424–440 mm) is about the same as the range of Group B’s measurements (410–460 mm).
A2: Group A’s last measurement is 435 mm, which is the average.	B2: Group B’s data are more spread, so they are more likely to capture the true value.	AB2: Both groups measured the same average (435 mm).
A3: Group A probably used better methods than Group B.	B3: Group B probably used better methods than Group A.	AB3: Group A and B used the same methods.
A4: The spread of Group A’s measurements (e.g., standard deviation) is smaller than Group B’s.	B4: Group A’s data are spread more narrowly, so they may have manipulated their data.	AB4: The spread of each group’s measurements (e.g., standard deviation) are roughly the same.
A5: Group A measured 435 mm twice.	B5: The spread of Group A’s measurements (e.g., standard deviation) is smaller than Group B’s.	AB5: Both groups measured 441 mm, 424 mm, and 440 mm.
AO: Other (please describe).	BO: Other (please describe).	ABO: Other (please describe).

survey. The homework assignment was not graded for correctness, so the students perceived the assignment as low-stakes. The 413 students who took the survey were randomly placed in one of two groups, such that approximately half of the students were prompted to provide open-response reasoning for DMSS but not SMDS, and the other half were prompted to provide open-response reasoning for SMDS but not DMSS. All students (both groups) were prompted to select reasoning items from the closed-response question regardless of whether or not they had received an open-response reasoning question. Our dataset

consists of 248 students, after removing nonconsenting students and students under the age of 18 from the dataset. As indicated in Table III, 130 of these students were in the group that gave an open-response answer to SMDS (group 1) and 118 of these students were in the group that gave an open-response answer to DMSS (group 2).

B. Development of closed-response explanations

Most previous implementations of the PMQ have used the open-response question to measure students’

TABLE II. Closed-response reasoning items for the DMSS probe. Students were prompted to choose one or two of these statements that were most relevant to their reasoning for choosing “Yes” the groups’ measurements agree or “No” the groups’ measurements do not agree.

Yes	No
Y1: Both groups measured outliers. If we discard the outliers, the two groups’ measurements are similar.	N1: The averages are different (433 mm and 435 mm).
Y2: Both groups got the values 440, 433, and 432 mm.	N2: The percent difference (or percent error) between the two groups’ measurements is large.
Y3: The ranges in the groups’ measurements overlap.	N3: The ranges of each group’s measurements don’t overlap enough.
Y4: The percent difference (or percent error) between the two groups’ measurements is small.	N4: Some of the groups’ measurements are different.
Y5: The difference between the two averages (2 mm) is small.	N5: The difference between the two averages (2 mm) is large.
Y6: The difference between the two averages (2 mm) is small considering the spread of the data (e.g., standard deviation).	N6: The difference between the two averages (2 mm) is large relative to the ranges of each group’s measurements.
Y7: The difference between the two averages (2 mm) is small considering the size of the average measurements (433 or 435 mm).	N7: The difference between the two averages (2 mm) is large compared to the size of the average measurements (433 or 435 mm).
Y8: The difference between the two averages (2 mm) is small relative to the ranges of each group’s measurements.	N8: The difference between the two averages (2 mm) is large considering the spread of the data (e.g., standard deviation).
YO: Other (please describe)	NO: Other (please describe)

TABLE III. Description of the two randomly assigned groups.

	Group 1	Group 2
N	130	118
SMDS open-response?	Yes	No
DMSS open-response?	No	Yes

reasoning. We developed a preliminary set of closed-response reasoning items using the closed-response options in Ref. [33] as well as the open-response coding schemes in the original papers reporting on the PMQ [28,31,32]. We added several options so that multiple point and set reasoning ideas [32] were available and so that there were parallel statements between the different answers to the judgment questions. We then iteratively revised the statements over the course of three rounds of interviews with students at Cornell University and California State University, Fullerton. In the interviews, all students were prompted to explain their reasoning out loud before proceeding to the next page that contained the closed-response reasoning items. The students were also asked if they had any questions or comments about any of the closed-response statements. After each round of approximately four student interviews, the research team discussed the interviews and made edits to the closed-response reasoning items, reaching saturation before the final round. A total of 11 interviews were conducted.

C. Open-response coding

One of our goals was to evaluate whether we measure different student reasoning from closed-response and open-response questions, so we developed a coding scheme to analyze the open-response data for evidence of whether or not each reasoning item (from the closed-response version) was included. Each closed-response option was considered independently, such that there were no constraints on how many items could be coded for a given open-response answer. Within an open-response answer, however, any distinct idea articulated could only relate to one code, not multiple. For example, if a student said, “the two averages are close together,” this single idea would be related to one code; in this example, “the difference between the two averages (2 mm) is small” [Y5]. The idea would not also be related to the code, “the percent difference (or percent error) between the two groups’ measurements is small” [Y4], because the student response did not explicitly mention percent difference or percent error. In cases where a distinct idea was ambiguous in a way that could overlap with multiple ideas, we maintained the rule of one code per idea by making a judgment call for which code is the best fit. For example, if a student expressed the idea that Group A’s data were “closer together,” but with no other information, the response could be consistent with options A1 (“Group

A’s range is smaller than Group B’s range”) or A4 (“The spread of Group A’s measurements (e.g., standard deviation) is smaller than Group B’s”). Of these two reasoning options, A4 was chosen to better match the student response because “closer together” evokes the idea of spread more than range and because the student did not explicitly express that *range* was an important factor (a key element in option A1). Students had to explicitly mention range or the upper or lower limits of the dataset to be coded as A1. While coding the data, the coders had no access to information about which closed-response items the students chose on the subsequent closed-response question.

When coding DMSS “other” responses, we noticed a common reasoning item that was not present in any of the closed-response reasoning items; 19% of students articulated the idea that “Group A’s spread is similar to/different from Group B’s spread.” We identified this idea as YO-S and separated these responses from those that articulated a different, other response. Compared with all the closed-response choices, this idea does not answer the posed question, even though it is a factually true statement.

We evaluated interrater reliability of our coding scheme with 10% of the data. The overall Cohen’s kappa for SMDS was 0.87 and for DMSS was 0.86 between two independent raters, indicating excellent agreement between raters [37]. The two raters then divided the remaining data to code individually.

D. Data analysis

To answer our research questions, we report percentages of students who were coded as writing (open-response item) or chose (closed-response item) each reasoning item. We report uncertainty of these proportions as the 95% confidence intervals for binomial proportions, estimated using the Wilson score interval [38]. To answer RQ1, we compare the closed-response selections of group 1 and group 2 on each question, evaluating whether the selections differed depending on whether they saw the open-response item before the closed-response item. To answer RQ2, we compare each group’s proportion of coded open-response reasoning items and selected closed-response reasoning choices.

We primarily interpret the results by evaluating the relative size of effects, that is, by comparing the frequencies of reasoning items generated or selected. In the Appendix, we report two-tailed proportion tests to compare the proportions of students in each group who selected or generated each reasoning item. We report the outcome of each proportion test as both an unadjusted p value and a Bonferroni corrected p value that accounts for the number of tests in each analysis (i.e., the number of reasoning items across which comparisons are being made). We also calculate effect size using Cohen’s h for binomial data [39]. Although we report statistical significance, we avoid reliance on statistical significance

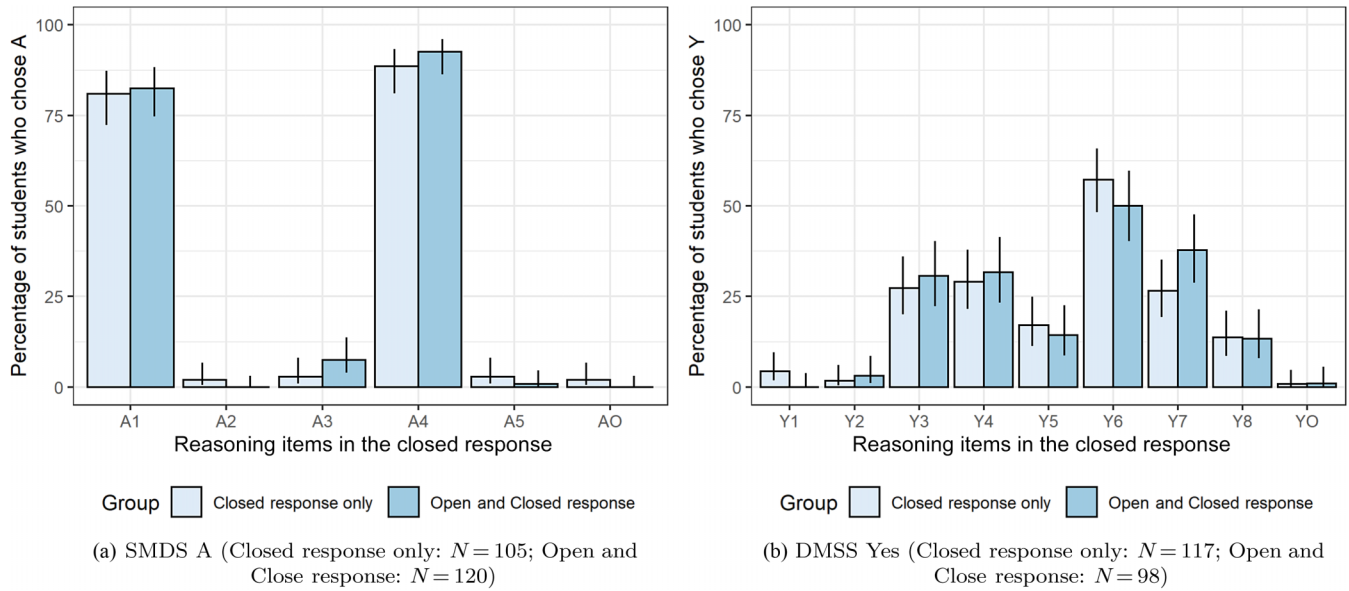


FIG. 3. The graphs compare the closed-response reasoning choices of students who received either the closed-response question only or both the open- and closed-response questions. The percentages come from fractions of the students who chose the most common judgment (A for SMDS and Y for DMSS).

to guide interpretations, as per recommendations in the literature [40–42].

III. RESULTS

Each judgment question had a most common answer selected by a large majority of students. In the SMDS judgment question, the most common choice was “Group A” (selected by 120 students in group 1 and 105 students in group 2; 91% of all students). In the DMSS judgment question, the most common choice was “Yes” (selected by 117 students in group 1 and 98 students in group 2; 87% of all students). Thus, we focus our analysis of students’ reasoning on the students who selected these two choices for the judgment questions. The analyses for the responses associated with the less common judgment choices are provided in the Appendix. In each case, the percent of responses that included each reasoning item does not add up to 100% across all reasoning items because students could generate or select more than one reasoning item in their response.

A. The effect of answering the open-response question on closed-response reasoning

Our first research question investigates the impact of asking students to first write their reasoning in an open-response format on their selections to a subsequent closed-response question. To do so, we compare the reasoning items selected by both group 1 and group 2 students on the closed-response reasoning questions.

For SMDS, students selecting “Group A” on the judgment question largely selected reasoning options A1

and A4 [Fig. 3(a)], with no discernible difference between groups. The results of the proportion tests (Table IV) show that we do not detect differences between group 1 and group 2’s selections for any “Group A” SMDS reasoning item.

For DMSS, students selecting “Yes” on the judgment question selected a wide range of reasoning options [Fig. 3(b)]. Despite small differences between the groups, we again do not detect statistical differences between group 1 and group 2’s selections for any “Yes” DMSS reasoning item (Table V).

The Appendix (Sec. A 1) contains a similar analysis of the students who made the less common judgments: “Groups A and B are equally good measurements” for SMDS ($N=9$ for group 1, $N=12$ for group 2) and “No” for DMSS ($N=13$ for group 1, $N=20$ for group 2). These sample sizes are too small to make claims about reasoning between the two groups, but the trends in the results do not contradict the pattern of no effect that we see from the more common judgments. We were unable to analyze students who made the judgment “Group B” on SMDS, as only one student in group 1 and only one student in group 2 selected this option.

To summarize, we find no effect on the closed-response question if students are first asked for their open-ended reasoning before asking them the closed-response question for both SMDS and DMSS.

B. Measuring student reasoning on closed-response versus open-response questions

Our second research question investigates whether we measure different student reasoning on closed-response

TABLE IV. Proportion test results for students who selected “Group A” on SMDS. The proportion test compares the proportion of respondents in group 1 vs group 2 who selected each reasoning item. The corrected p value is Bonferroni-adjusted. The effect size is calculated with Cohen’s h .

Reasoning	Group	Number	p value	Corrected p value	Effect size
A1: Group A’s range is smaller than Group B’s range.	Group 1 ($N = 120$)	99	0.8	1	−0.04
	Group 2 ($N = 105$)	85			
A2: Group A’s last measurement is 435 mm, which is the average.	Group 1 ($N = 120$)	0	0.13	0.8	0.28
	Group 2 ($N = 105$)	2			
A3: Group A probably used better methods than Group B.	Group 1 ($N = 120$)	9	0.12	0.8	−0.22
	Group 2 ($N = 105$)	3			
A4: The spread of Group A’s measurements (e.g., standard deviation) is smaller than Group B’s.	Group 1 ($N = 120$)	111	0.3	1	−0.13
	Group 2 ($N = 105$)	93			
A5: Group A measured 435 mm twice.	Group 1 ($N = 120$)	1	0.3	1	0.16
	Group 2 ($N = 105$)	3			
AO: Other (please describe)	Group 1 ($N = 120$)	0	0.13	0.8	0.28
	Group 2 ($N = 105$)	2			

TABLE V. Proportion test results for students who selected “Yes” on DMSS. The proportion test compares the proportion of respondents in group 1 vs group 2 who selected each reasoning item. The corrected p value is Bonferroni-adjusted. The effect size is calculated with Cohen’s h .

Reasoning	Group	Number	p value	Corrected p value	Effect size
Y1: If we discard the outliers, the two group’s measurements are similar.	Group 1 ($N = 117$)	5	0.04	0.4	0.42
	Group 2 ($N = 98$)	0			
Y2: Both groups got the values 440, 433, and 432 mm.	Group 1 ($N = 117$)	2	0.5	1	−0.09
	Group 2 ($N = 98$)	3			
Y3: The ranges in the groups’ measurements overlap.	Group 1 ($N = 117$)	32	0.6	1	−0.07
	Group 2 ($N = 98$)	30			
Y4: The percent difference (or percent error) between the two groups’ measurements is small.	Group 1 ($N = 117$)	34	0.7	1	−0.06
	Group 2 ($N = 98$)	31			
Y5: The difference between the two averages (2 mm) is small.	Group 1 ($N = 117$)	20	0.6	1	0.08
	Group 2 ($N = 98$)	14			
Y6: The difference between the two averages (2 mm) is small considering the spread of the data (e.g., standard deviation).	Group 1 ($N = 117$)	67	0.3	1	0.15
	Group 2 ($N = 98$)	49			
Y7: The difference between the two averages (2 mm) is small considering the size of the average measurements (433 or 435 mm).	Group 1 ($N = 117$)	31	0.08	1	−0.24
	Group 2 ($N = 98$)	37			
Y8: The difference between the two averages (2 mm) is small relative to the ranges of each group’s measurements.	Group 1 ($N = 117$)	16	0.9	1	0.01
	Group 2 ($N = 98$)	13			

versus open-response questions. To do so, we compare the reasoning items generated by students on the open-response question (according to hand coding) to those selected by the same students on the closed-response question. Figure 4 represents a possible Venn diagram relating students’ generated versus selected reasoning items. If the closed-response and open-response questions measured student reasoning in the same way, we would expect the blue and red circles to overlap completely.

Overall, students generated fewer reasoning items on the open-response question than they selected on the closed-response question. Students’ responses to the open-response question contained an average of 1.2 ideas (standard deviation of 0.50). On the closed-response question, the average number of responses selected by students was 1.78 (standard deviation of 0.50).

The proportions of students selecting (closed-response) or generating (open-response) each reasoning item (Fig. 5,

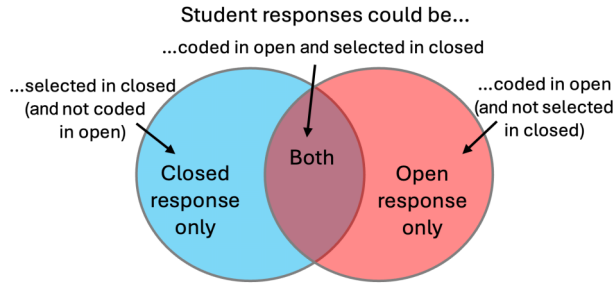


FIG. 4. Venn Diagram depicting the options for how students may answer the closed-response and open-response reasoning questions.

Tables VIII and IX) are similar for most—but not all—reasoning items. Consistent with the average number of generated and selected explanations per student, slightly more students selected any given reasoning item on the closed-response question than generated that same item on the open-response question. There were several exceptions to these general trends.

First, more students generated ideas coded as Other (AO, YO, and YO-S in Fig. 5) and “Y5” (“The difference between the two averages [2 mm] is small”) on the open-response question than selected them on the closed-response question.

Second, several reasoning items were selected by *many* more students on the closed-response question than were generated on the open-response question. For example, 80% of students selected A1 (“Group A’s range is smaller than Group B’s range”) on the SMDS closed-response

question, compared with 20% generating it on the open-response question. We also observe that Y4 (“The percent difference [or percent error] between the two groups’ measurements is small”) and Y7 (“The difference between the two averages [2 mm] is small considering the size of the average measurements [433 or 435 mm]”) are selected much more frequently (at least 4 times as often) in the closed-response question than generated in the open-response question.

To further explore these exceptions, we evaluate, of the students who generated the most common reasoning items (on the open-response question), what proportion selected each of the closed-response reasoning items (Fig. 6). This group of students represents the overlapping region of the Venn diagram in Fig. 4.

In SMDS [Fig. 6(a)], 21% of students generated A1 on the open-response question, 92% of these students selected A1 and 96% selected A4. Similarly, 83% of students generated A4, 93% of them selected A4 and 84% of them selected A1. That is, most students generated A4 (but not A1) on the open-response question, and most of these students went on to select *both* A1 and A4 on the closed-response question.

In DMSS [Fig. 6(b)], 51% of the students generated Y5 in the open-response question, yet only 24% of these students selected Y5 on the closed-response question. More students who generated Y5 on the open-response question selected Y3, Y4, Y6, or Y7 than selected Y5 on the closed-response question. By contrast, 34% of the students generated Y6 in the open-response question, 70% of whom also selected Y6 on the closed-response question. Finally, 19% of students

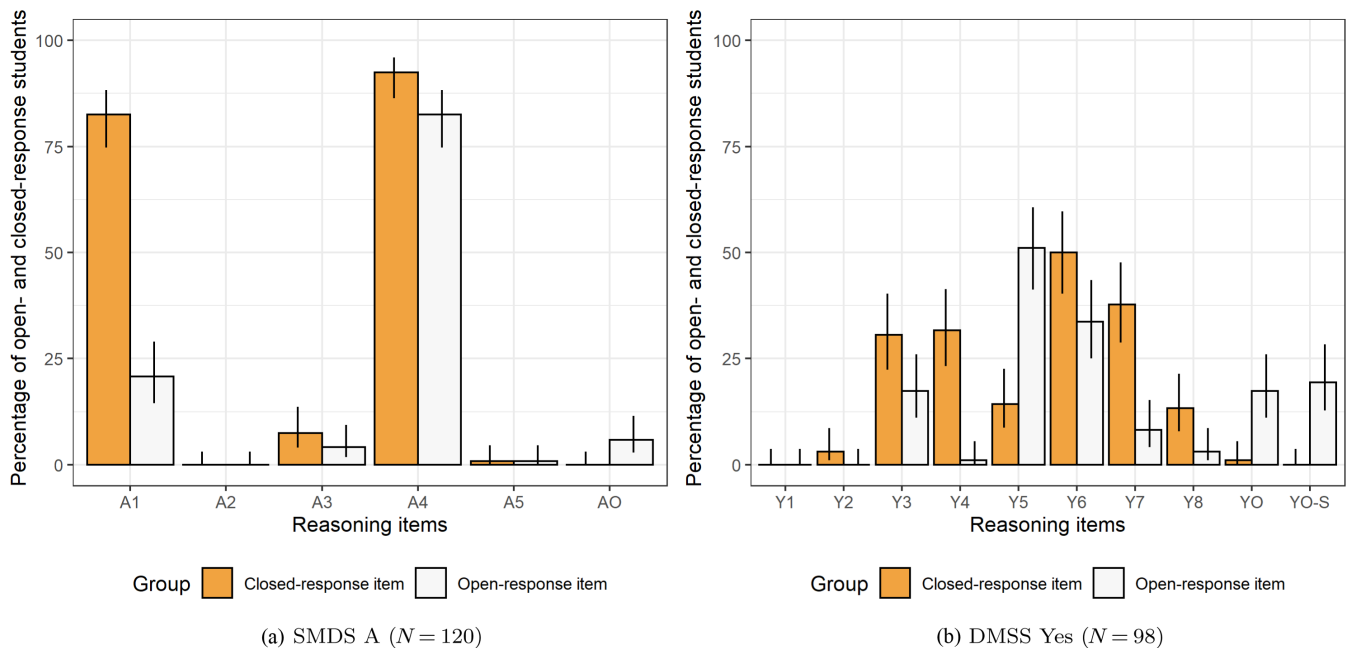


FIG. 5. The graphs compare the reasoning items either selected (on the closed-response item) or generated (on the open-response item) by the students who received both the open and closed-response options.

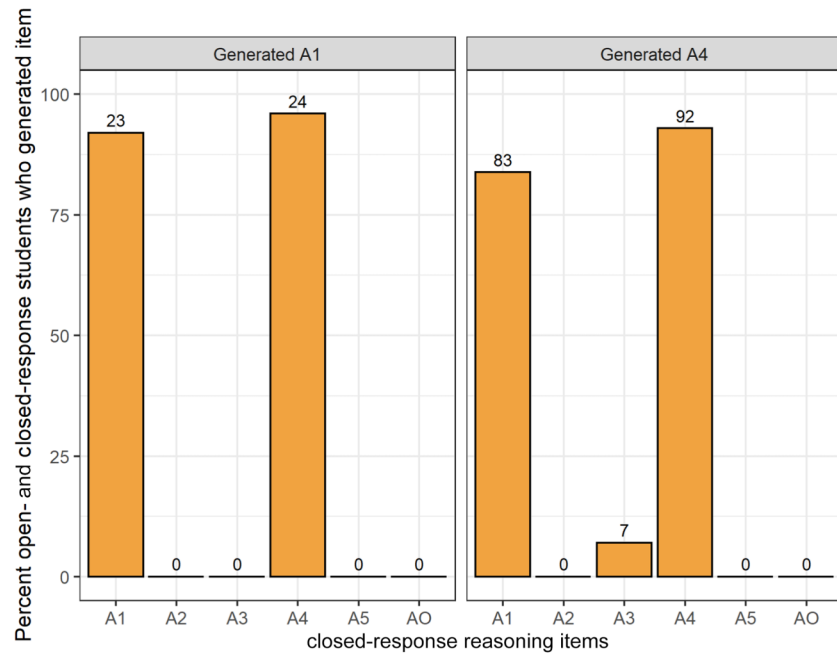
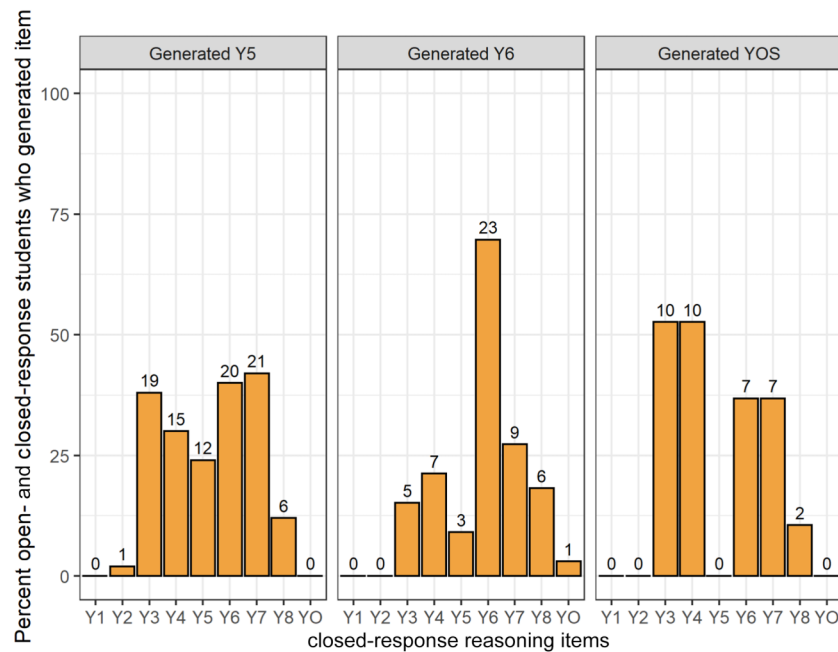
(a) SMDS A (Generated A1: $N = 25$; Generated A4: $N = 99$)(b) DMSS Yes (Generated Y5: $N = 50$; Generated Y6: $N = 33$; Generated YOS: $N = 19$)

FIG. 6. The graphs compare the reasoning items selected on the closed-response item by the students who generated common reasoning items on the open-response question.

generated YO-S (“Group A’s spread is similar to/different from Group B’s spread”) and none of these students selected YO (Other) on the closed-response question. Compared with the closed-response choices, it does not answer the posed question, even though it is a factually true statement. As a result, this difference does represent a genuine difference in measuring using open response and closed response, as our closed-response format made it less

likely that students would provide us with irrelevant justifications of their reasoning.

IV. DISCUSSION

In this case study, we compared open- and closed-response versions of two questions that measured students’ reasoning about measurement uncertainty from the PMQ

[28,31,32]. We found that first asking students their reasoning in an open-response question did not change their answer to the closed-response question. We also found that students' reasoning is measured differently between the open-response and closed-response questions in a number of ways.

We explore these differences (elaborated below) through two lenses. First is that of accessibility and availability [26]; accessibility is the extent to which a reasoning item "comes to mind" for a student, in particular as measured by a *recall* task, while availability refers to whether or not a reasoning item represents a piece of knowledge that is in the student's brain (i.e., an idea they recognize as relevant when they see it written down). Second is that of framing [43,44]; students come to a task with a set of expectations and beliefs about what is happening, in this case, expectations and beliefs about closed-response questions. We then expand to consider the potential implications of these results on research and assessment.

A. Discussion of results

We find across two different questions that students selected more reasoning items on the closed-response question than they generated on the open-response question. Through the lens of cognitive accessibility [26], we expect, in some cases, that the closed-response options cued ideas that students recognize as relevant, including ideas that may have been available but not generated in the open-response question. This interpretation is supported by the evidence that many students selected both the reasoning item they generated and an additional reasoning item(s) they did not generate. For example, on SMDS, most students generated item A4, and while it continued to be selected most often in the closed-response version, nearly as many students also selected item A1 (Fig. 6). Through the lens of accessibility and availability, we would conclude that, for most students, the ideas in A1 and A4 were viewed as acceptable or relevant reasoning items for this question, even though A4 was the most accessible on the open-response question.

In some cases, however, students did not select the reasoning item(s) they generated and only selected different closed-response items (e.g., Y5 on DMSS and Other on both questions, Figs. 5 and 6). DMSS reasoning item Y5 ("The difference between the two averages [2 mm] is small") was generated in 50% of open-response answers but was only selected in 14% of closed-response answers. In this case, the reasoning item is perhaps a vague or shortened version of three other closed-response options that compare the difference between the two averages to a reference point (Y6: "... the spread of the data," Y7: "... the size of the average measurements," and Y8: "... the range of the measurements"). Similarly, students who generated ideas that were coded as Other rarely went on to select Other on the closed-response items. One explanation for these results is that students' more accessible but

relatively vague explanations to the open-response format were replaced by more specific, available explanations in the closed-response format.

Assessment framing provides a complementary explanation to these results [44]. Students may have held expectations about closed-response assessments related to picking what appear to be the "better" items. In the case of Y5 on DMSS, the alternatively selected items are longer and more specific than Y5, which may be viewed as more expertlike responses. In the case of A1 and A4 on SMDS, both items are relatively long and specific and may have both been viewed as more expertlike responses. In the case of Other, students may have assumed that the listed choices were better selections than to write in their own other idea, expecting that the test writers would not have left out the best idea.

B. Implications for research and assessment

Together, our results suggest that (1) providing an open-response question before a closed-response question has no measurable bearing on how students answer the closed-response question compared to answering the closed-response question alone and (2) open-response and closed-response questions provide different information about students' reasoning. The first finding implies that researchers need not pair open-response and closed-response questions if the closed-response questions are the ones that will be ultimately analyzed. The second finding, however, reinforces the idea that open-response and closed-response questions may lead to different conclusions about student thinking [45], even though the responses may be scored similarly [20,21].

Consider, for example, if we had only administered the open-response version of the SMDS question. We may have claimed that students think that spread is much more important than range in evaluating which of two datasets is better, unaware that the students consider both to be good answers when presented with both items. If we had only administered the closed-response version of this question, in contrast, we may have claimed that students consider spread and range as equally important, unaware that the idea of comparing range is less accessible than the idea of comparing spread.

Similarly, if we had only administered DMSS as an open-response question, we may have concluded that students consider raw differences and differences relative to spread to be most important and that they do not see the value of reasoning about overlapping ranges, percent error, or differences relative to the average. A closed-response-only version of DMSS would have instead measured that students frequently recognize or endorse differences relative to spread, overlapping ranges, percent error, and differences relative to the average when these explanations are explicitly presented. Additionally, a closed-response-only version would have completely missed the observation

that a lot of students express the two other ideas that the individual data points are similar or the spreads are similar.

These findings underscore the importance of matching the type of measurement to the research questions [22–24]. If the goal is to make claims about which ideas come to mind (accessibility), an open-response question may be more appropriate. If the goal is to determine which ideas students recognize or endorse as relevant when explicitly presented, a closed-response question may be more appropriate. A key risk with open-response questions, however, that emerges from our results is that many students struggled to articulate their reasoning in a way that could be precisely and accurately coded by the research team. We expect that, in some cases, the closed-response options provided students with more precise language to express ideas they recognize as relevant, potentially identifying aspects of their reasoning that may not have been captured in an open-response format. A key risk with the closed-response questions, however, is that some students reject their initially most accessible ideas in favor of ideas that may “look better,” perhaps because they appear more expertlike or mathematically sophisticated. In these cases, students’ selections may reflect their recognition or endorsement of a reasoning statement after seeing it, rather than the reasoning that initially informed their judgment. As a result, closed-response selections may provide insight into which explanations students view as relevant or convincing when cued, but not necessarily which explanations were spontaneously activated during their original reasoning.

Researchers across physics education are exploring the effectiveness of machine learning and generative artificial intelligence (AI) to replace or augment human coding [2,3,46,47], in order to scale measurements of student reasoning through open-response questions. The goal, in many cases, is to replace closed-response questions in favor of more open-ended formats. As an example, this study is part of a larger project where we are administering a survey to measure students’ reasoning about uncertainty across at least 50 classes at 30 or more institutions. A previous version of this survey was comprised of a series of open-response questions that were analyzed individually by the research team [27,48,49]. As we scale up the deployment of this survey, we require methods to measure these ideas more efficiently. For some questions, we are using machine learning to scale this analysis, leaving the questions as open-response (as has been done for the PMQ questions in this study by other researchers as well [6]). For other questions, we hope to scale up analysis by converting a previously open-response question to a closed-response question. Our results from this study show that both of these options result in some information gained and other information lost. This analysis demonstrates the need to evaluate trade-offs in the context of each

question and determine what information is the most important to measure.

Our results also underscore that there are reasons other than efficient scoring to convert an open-response question to a closed-response question. The cognitive accessibility rule would suggest that, on an open-response task, survey respondents are less likely to consider multiple explanations for a phenomenon once they have provided the explanation that is most accessible to them [26]. If we had only asked our questions as open-response questions, we would limit the precision of our measurement of student reasoning and our ability to measure how students think about less accessible reasoning statements. The emergence of generative AI and chatbots also means that take-home open-response questions may become less accurate measurement methods, though we do not currently know to what extent students look up answers to the closed-response questions as compared to asking a chatbot to generate an answer to an open-response question. Future work should seek to probe this explicitly.

On the other hand, there are also reasons to move to automated coding of open-response questions instead of using closed-response questions to scale measurement. Open-response questions will identify the most accessible reasoning items and the extent to which students can articulate them, while also limiting the testing strategies to selected so-called “better” items.

The important task, therefore, when designing a measurement tool is to evaluate whether you want to measure students’ most accessible ideas or the broader set of ideas that may become available when they are presented with them. The answer to this question is likely context-dependent [22–24]. It may not always be possible to know *a priori* if an open-response or closed-response question is the better choice for a particular question context. In such a case, it may be desirable to pilot the question in both the open-response and closed-response formats to preview the differences between the results you would measure with each method. Our results suggest that sequentially asking both the open-response question and closed-response question to the same students will not alter the students’ selections on the closed-response question.

In this case study, we only evaluated two questions with one population of students. Future studies should evaluate the robustness of these results with different question contexts (e.g., not measurement uncertainty), different student populations, and different closed-response formats. Plausibly, results from a choose-one multiple-choice question may have looked more similar to the generated ideas on the open-response question, given that many students only generated one idea. Future studies should also explore the effectiveness of responsive generative AI questions that ask students to clarify potentially vague or underspecified open-response question reasoning.

ACKNOWLEDGMENTS

We would like to thank Matt Thomas for statistical consulting support and the UNC Chapel Hill teaching faculty for commenting on a draft of this paper. We acknowledge support from National Science Foundation under (Grants No. DUE-2336135, No. DUE-2336136) and the Cornell University Provost Diversity Fellowship.

DATA AVAILABILITY

The data that support the findings of this article are openly available [24].

APPENDIX

The appendix includes the data and analysis for students who selected the less common judgment questions on the two PMQ questions.

1. RQ1

In the SMDS judgment question, the second most common choice was “Groups A and B have equally good measurements” (selected by 9 students in group 1 and 12 students in group 2; 8.5% of all students). A list of the closed-response reasoning items provided to students who selected this option, along with the number of students in each group who selected each reasoning item, is provided in Table VI.

Similarly, in the DMSS judgment question, the second most common (and only other) choice was “No” (selected

by 13 students in group 1 and 20 students in group 2; 13% of all students). A list of the closed-response reasoning items provided to students who selected this option, along with the number of students in each group who selected each reasoning item, is provided in Table VII.

These sample sizes are too small to make claims about reasoning between the two groups, though the results of these proportion tests do not contradict the pattern of no effect that we see among students who answered “Group A” for SMDS and “Yes” for DMSS. We are unable to analyze students who made the judgment “Group B” on SMDS, as only one student in group 1 and only one student in group 2 selected this option.

2. RQ2

In an analysis that parallels the analysis in Sec. III B, we measured the students’ reasoning on both the open-response and closed-response reasoning questions by analyzing the percent of responses that included each reasoning item for these less common judgment question responses (Tables X and XI). The measurements from the open-response and the closed-response questions are similar for most reasoning items, though slightly more students selected items on the closed-response question with the exception of Other. These trends are consistent with the general trend we observed in the analysis of students who selected the more common judgment question responses in Sec. III B.

TABLE VI. Proportion test results for students who selected “Group A and Group B have equally good measurements” on SMDS. The proportion test compares the proportion of respondents in group 1 vs group 2 who selected each reasoning statement. The corrected p value is Bonferroni-adjusted.

Reasoning	Group	Number	p value	Corrected p value
AB1: The range of Group A’s measurements (424–440 mm) is about the same as the range of Group B’s measurements (410–460 mm).	Group 1 ($N = 9$)	1	0.8	1
	Group 2 ($N = 12$)	1		
AB2: Both groups measured the same average (435 mm).	Group 1 ($N = 9$)	5	0.3	1
	Group 2 ($N = 12$)	4		
AB3: Groups A and B used the same methods.	Group 1 ($N = 9$)	5	0.6	1
	Group 2 ($N = 12$)	8		
AB4: The spread of each group’s measurements (e.g., standard deviation) is roughly the same.	Group 1 ($N = 9$)	2	0.7	1
	Group 2 ($N = 12$)	2		
AB5: Both groups measured 441 mm, 424 mm, and 440 mm.	Group 1 ($N = 9$)	0	...	...
	Group 2 ($N = 12$)	0		
ABO: Other (please describe)	Group 1 ($N = 9$)	1	0.8	1
	Group 2 ($N = 12$)	1		

TABLE VII. Proportion test results for students who selected “No” on DMSS. The proportion test compares the proportion of respondents in group 1 vs group 2 who selected each reasoning statement. The corrected p value is Bonferroni-adjusted.

Reasoning	Group	Number	p value	Corrected p value
N1: The averages are different (433 mm and 435 mm).	Group 1 ($N = 13$)	9	0.6	1
	Group 2 ($N = 20$)	12		
N2: The percent difference (or percent error) between the two groups’ measurements is large.	Group 1 ($N = 13$)	3	0.3	1
	Group 2 ($N = 20$)	2		
N3: The ranges of each group’s measurements don’t overlap enough.	Group 1 ($N = 13$)	1	0.3	1
	Group 2 ($N = 20$)	4		
N4: Some of the groups’ measurements are different.	Group 1 ($N = 13$)	2	0.07	0.6
	Group 2 ($N = 20$)	0		
N5: The difference between the two averages (2 mm) is large.	Group 1 ($N = 13$)	0	...	...
	Group 2 ($N = 20$)	0		
N6: The difference between the two averages (2 mm) is large relative to the ranges of each group’s measurements.	Group 1 ($N = 13$)	2	1	1
	Group 2 ($N = 20$)	3		
N7: The difference between the two averages (2 mm) is large compared to the size of the average measurements (433 or 435 mm).	Group 1 ($N = 13$)	1	0.2	1
	Group 2 ($N = 20$)	0		
N8: The difference between the two averages (2 mm) is large considering the spread of the data (e.g., standard deviation).	Group 1 ($N = 13$)	2	0.2	1
	Group 2 ($N = 20$)	7		

TABLE VIII. Proportion test results for students who selected “Group A” on SMDS ($N = 120$). The proportion test compares the proportion of respondents who generated the reasoning statement on open-response (OR) versus the proportion who selected the reasoning statement on closed-response (CR). The corrected p value is Bonferroni-adjusted. The effect size is calculated with Cohen’s h .

Reasoning	OR/CR	Number	p value	Corrected p value	Effect size
A1: Group A’s range is smaller than Group B’s range.	OR	25	< 0.001	< 0.001	1.33
	CR	99			
A2: Group A’s last measurement is 435 mm, which is the average.	OR	0	...	...	0
	CR	0			
A3: Group A probably used better methods than Group B.	OR	5	0.3	1	0.14
	CR	9			
A4: The spread of Group A’s measurements (e.g., standard deviation) is smaller than Group B’s.	OR	99	0.02	0.12	0.31
	CR	111			
A5: Group A measured 435 mm twice.	OR	1	1	1	0
	CR	1			
AO: Other (please describe)	OR	7	0.007	0.04	−0.49
	CR	0			

TABLE IX. Proportion test results for students who selected “Yes” on DMSS ($N = 98$). The proportion test compares the proportion of respondents who generated the reasoning statement on open-response (OR) vs the proportion who selected the reasoning statement on closed-response (CR). The corrected p value is Bonferroni-adjusted. The effect size is calculated with Cohen’s h .

Reasoning	Group	Number	p value	Corrected p value	Effect size
Y1: If we discard the outliers, the two groups’ measurements are similar.	OR	0	...	...	0
	CR	0			
Y2: Both groups got the values 440, 433, and 432 mm.	OR	0	0.08	0.9	0.35
	CR	3			
Y3: The ranges in the groups’ measurements overlap.	OR	8	<0.001	<0.001	0.31
	CR	30			
Y4: The percent difference (or percent error) between the two groups’ measurements is small.	OR	2	<0.001	<0.001	0.99
	CR	31			
Y5: The difference between the two averages (2 mm) is small.	OR	37	<0.001	0.002	−0.82
	CR	14			
Y6: The difference between the two averages (2 mm) is small considering the spread of the data (e.g., standard deviation).	OR	34	0.03	0.3	0.33
	CR	49			
Y7: The difference between the two averages (2 mm) is small considering the size of the average measurements (433 or 435 mm).	OR	8	<0.001	<0.001	0.74
	CR	37			
Y8: The difference between the two averages (2 mm) is small relative to the ranges of each group’s measurements.	OR	3	0.009	0.1	0.39
	CR	13			

TABLE X. Proportion test results for students who selected “Group A and Group B have equally good measurements” on SMDS ($N = 9$). The proportion test compares the proportion of respondents who generated the reasoning statement on the open-response (OR) question versus the proportion who selected the reasoning statement on the closed-response (CR) question. The corrected p value is Bonferroni-adjusted.

Reasoning	Group	Number	p value	Corrected p value
AB1: The range of Group A’s measurements (424–440 mm) is about the same as the range of Group B’s measurements (410–460 mm).	OR	1	1	1
	CR	1		
AB2: Both groups measured the same average (435 mm).	OR	3	0.3	1
	CR	5		
AB3: Groups A and B used the same methods.	OR	3	0.3	1
	CR	5		
AB4: The spread of each group’s measurements (e.g., standard deviation) are roughly the same.	OR	0	0.13	0.8
	CR	2		
AB5: Both groups measured 441 mm, 424 mm, and 440 mm.	OR	0	...	...
	CR	0		
ABO: Other (please describe)	OR	4	0.11	0.7
	CR	1		

TABLE XI. Proportion test results for students who selected “No” on DMSS ($N = 20$). The proportion test compares the proportion of respondents who generated the reasoning statement on the open-response (OR) question versus the proportion who selected the reasoning statement on the closed-response (CR) question. The corrected p value is Bonferroni-adjusted.

Reasoning	Group	Number	p value	Corrected p value
N1: The averages are different (433 mm and 435 mm).	OR	12	1	1
	CR	12		
N2: The percent difference (or percent error) between the two groups' measurements is large.	OR	0	0.15	1
	CR	2		
N3: The ranges of each group's measurements don't overlap enough.	OR	1	0.15	1
	CR	4		
N4: Some of the groups' measurements are different	OR	1	0.3	1
	CR	0		
N5: The difference between the two averages (2 mm) is large.	OR	0	...	...
	CR	0		
N6: The difference between the two averages (2 mm) is large relative to the ranges of each group's measurements.	OR	1	0.3	1
	CR	3		
N7: The difference between the two averages (2 mm) is large compared to the size of the average measurements (433 or 435 mm).	OR	0	...	...
	CR	0		
N8: The difference between the two averages (2 mm) is large considering the spread of the data (e.g., standard deviation).	OR	1	0.02	0.16
	CR	7		

- [1] L. K. Nelson, Computational grounded theory: A methodological framework, *Sociol. Methods Res.* **49**, 3 (2020).
- [2] T. O. B. Odden, H. Tyseng, J. T. Mjaaland, M. F. Kreutzer, and A. Malthe-Sørenssen, Using text embeddings for deductive qualitative research at scale in physics education, [arXiv:2402.18087](https://arxiv.org/abs/2402.18087).
- [3] R. K. Fussell, E. M. Stump, and N. G. Holmes, Method to assess the trustworthiness of machine coding at scale, *Phys. Rev. Phys. Educ. Res.* **20**, 010113 (2024).
- [4] R. K. Fussell, M. Flynn, A. Damle, M. F. J. Fox, and N. G. Holmes, Comparing large language models for supervised analysis of students' lab notes, *Phys. Rev. Phys. Educ. Res.* **21**, 010128 (2025).
- [5] Z. Chen and T. Wan, Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy, *Phys. Rev. Phys. Educ. Res.* **21**, 010126 (2025).
- [6] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, *Phys. Rev. Phys. Educ. Res.* **18**, 010141 (2022).
- [7] K. F. Stanger-Hall and E. H. Chudler, Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes, *CBE—Life Sci. Educ.* **11**, 294 (2012).
- [8] J. T. Lavery, S. M. Underwood, R. L. Matz, L. A. Posey, J. H. Carmel, M. D. Caballero, C. L. Fata-Hartley, D. Ebert-May, S. E. Jardeleza, and M. M. Cooper, Characterizing college science assessments: The three-dimensional learning assessment protocol, *PLOS One* **11**, e0162333 (2016).
- [9] *Developing Assessments for the Next Generation Science Standards* (National Academies Press, Washington, D.C., 2014), p. 18409.
- [10] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, *Phys. Rev. Phys. Educ. Res.* **15**, 010135 (2019).
- [11] A. B. Heim, D. Esparza, N. G. Holmes, and M. K. Smith, Comparing study features is easy but identifying next steps is hard: Evaluating critical thinking through the biology lab inventory of critical thinking in ecology, *Ecol. Evol.* **13**, e10071 (2023).
- [12] A. E. Lawson, The development and validation of a classroom test of formal reasoning, *J. Res. Sci. Teach.* **15**, 11 (1978).
- [13] L. Bao, K. Koenig, Y. Xiao, J. Fritchman, S. Zhou, and C. Chen, Theoretical model and quantitative assessment of scientific thinking and reasoning, *Phys. Rev. Phys. Educ. Res.* **18**, 010115 (2022).

- [14] G. Geschwind, M. Vignal, M. D. Caballero, and H. J. Lewandowski, Evidence for validity and reliability of a research-based assessment instrument on measurement uncertainty, *Phys. Rev. Phys. Educ. Res.* **20**, 020125 (2024).
- [15] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020124 (2014).
- [16] Q. Chen, G. Zhu, Q. Liu, J. Han, Z. Fu, and L. Bao, Development of a multiple-choice problem-solving categorization test for assessment of student knowledge structure, *Phys. Rev. Phys. Educ. Res.* **16**, 020120 (2020).
- [17] M. Scott, T. Stelzer, and G. Gladding, Evaluating multiple-choice exams in large introductory physics courses, *Phys. Rev. ST Phys. Educ. Res.* **2**, 020102 (2006).
- [18] S.-Y. Lin and C. Singh, Can free-response questions be approximated by multiple-choice equivalents?, *Am. J. Phys.* **81**, 624 (2013).
- [19] M. M. Wooten, A. M. Cool, E. E. Prather, and K. D. Tanner, Comparison of performance on multiple-choice questions and open-ended questions in an introductory astronomy laboratory, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020103 (2014).
- [20] S. Breuer, T. Scherndl, and T. M. Ortner, Effects of response format on achievement and aptitude assessment results: Multi-level random effects meta-analyses, *R. Soc. Open Sci.* **10**, 220456 (2023).
- [21] M. C. Rodriguez, Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations, *J. Educ. Meas.* **40**, 163 (2003).
- [22] S. Connor Desai and S. Reimers, Comparing the use of open and closed questions for web-based measures of the continued-influence effect, *Behav. Res. Methods* **51**, 1426 (2019).
- [23] V. Baburajan, J. de Abreu e Silva, and F. C. Pereira, Open vs closed-ended questions in attitudinal surveys—comparing, combining, and interpreting using natural language processing, *Transp. Res. Part C: Emerging Technol.* **137**, 103589 (2022).
- [24] V. Baburajan, J. de Abreu e Silva, and F. C. Pereira, Do open-ended questions influence the measurement of attitudes? An investigation, *Transp. Res. Proc.* **76**, 608 (2024).
- [25] M. R. Federer, R. H. Nehm, J. E. Opfer, and D. Pearl, Using a constructed-response instrument to explore the effects of item position and item features on the assessment of students' written scientific explanations, *Res. Sci. Educ.* **45**, 527 (2015).
- [26] A. F. Heckler and A. M. Bogdan, Reasoning with alternative explanations in physics: The cognitive accessibility rule, *Phys. Rev. Phys. Educ. Res.* **14**, 010120 (2018).
- [27] E. M. Stump, M. Dew, G. Passante, and N. G. Holmes, Context affects student thinking about sources of uncertainty in classical and quantum mechanics, *Phys. Rev. Phys. Educ. Res.* **19**, 020157 (2023).
- [28] 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).
- [29] A. A. diSessa, Knowledge in pieces, in *Constructivism in the Computer Age, the Jean Piaget Symposium Series* (Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1988), pp. 49–70.
- [30] D. Hammer, Student resources for learning introductory physics, *Am. J. Phys.* **68**, S52 (2000).
- [31] 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).
- [32] 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).
- [33] R. F. Lippmann, Students' Understanding of Measurement and Uncertainty in the Physics Laboratory: Social Construction, Underlying Concepts, and Quantitative Analysis, Ph. D. thesis, University of Maryland, 2003.
- [34] D. L. Deardorff, Introductory Physics Students' Treatment of Measurement Uncertainty, Ph. D. thesis, North Carolina State University, 2001.
- [35] M. Vignal, G. Geschwind, B. Pollard, R. Henderson, M. D. Caballero, and H. J. Lewandowski, Survey of physics reasoning on uncertainty concepts in experiments: An assessment of measurement uncertainty for introductory physics labs, *Phys. Rev. Phys. Educ. Res.* **19**, 020139 (2023).
- [36] R. Fussell, J. Epstein, N. Holmes, and G. Passante, Open versus closed-response methods of assessing students' reasoning (2025). .
- [37] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33**, 159 (1977).
- [38] E. B. Wilson, Probable inference, the law of succession, and statistical inference, *J. Am. Stat. Assoc.* **22**, 209 (1927).
- [39] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Routledge, New York, 2013).
- [40] J. Cohen, The earth is round ($p < .05$)., *Am. Psychol.* **49**, 997 (1994).
- [41] G. Cumming, The new statistics: Why and how, *Psychol. Sci.* **25**, 7 (2013).
- [42] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, The preregistration revolution, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2600 (2018).
- [43] D. Tannen, *Framing in Discourse* (Oxford University Press on Demand, Oxford, UK, 1993).
- [44] K. Shar, R. S. Russ, and J. T. Laverty, Student epistemological framing on paper-based assessments, *Phys. Rev. Phys. Educ. Res.* **16**, 020101 (2020).
- [45] J. K. Hubbard, M. A. Potts, B. A. Couch, and K. E. Perez, How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats, *CBE—Life Sci. Educ.* **16**, ar26 (2017).
- [46] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, *Phys. Rev. Phys. Educ. Res.* **19**, 020123 (2023).

-
- [47] J. T. Mjaaland, M. F. Kreutzer, H. Tyseng, R. K. Fussell, G. Passante, N. G. Holmes, A. Malthe-Sørenssen, and T. O. B. Odden, Scalable and consistent few-shot classification of survey responses using text embeddings, [arXiv:2508.19836](#).
- [48] E. M. Stump, M. Hughes, G. Passante, and N. G. Holmes, Comparing introductory and beyond-introductory students' reasoning about uncertainty, [Phys. Rev. Phys. Educ. Res. **19**, 020147 \(2023\)](#).
- [49] A. Schang, M. Dew, E. M. Stump, N. G. Holmes, and G. Passante, New perspectives on student reasoning about measurement uncertainty: More or better data, [Phys. Rev. Phys. Educ. Res. **19**, 020105 \(2023\)](#).