ℒ

Students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for experimental physics

Bethany R. Wilcox

Department of Physics, University of Colorado, 390 UCB, Boulder, Colorado 80309, USA

H. J. Lewandowski

Department of Physics, University of Colorado, 390 UCB, Boulder, Colorado 80309, USA
and JILA, National Institute of Standards and Technology and University of Colorado,
Boulder, Colorado 80309, USA

Student learning in instructional physics labs represents a growing area of research that includes investigations of students' beliefs and expectations about the nature of experimental physics. To directly probe students' epistemologies about experimental physics and support broader lab transformation efforts at the University of Colorado Boulder and elsewhere, we developed the Colorado Learning Attitudes about Science Survey for experimental physics (E-CLASS). Previous work with this assessment has included establishing the accuracy and clarity of the instrument through student interviews and preliminary testing. Several years of data collection at multiple institutions has resulted in a growing national data set of student responses. Here, we report on results of the analysis of these data to investigate the statistical validity and reliability of the E-CLASS as a measure of students' epistemologies for a broad student population. We find that the E-CLASS demonstrates an acceptable level of both validity and reliability on measures of item and test discrimination, test-retest reliability, partial-sample reliability, internal consistency, concurrent validity, and convergent validity. We also examine students' responses using principal component analysis and find that, as expected, the E-CLASS does not exhibit strong factors (a.k.a. categories).

I. INTRODUCTION

The physics education research community has a growing body of research dedicated to investigating students' attitudes and epistemologies, such as what it means to know, learn, and do physics. This attention to students' epistemologies stems, in part, from research demonstrating that students' beliefs and expectations about the nature of doing and knowing physics can be linked to both their decision to pursue physics (i.e., retention and persistence) and their content learning in a physics course [1,2]. To further investigate these links, researchers have developed several assessments designed to directly measure students' epistemologies and expectations (E&E) about both physics specifically [3–5] and the nature of science more generally [6–8].

Until recently, the available physics-specific E&E surveys have focused on assessing students' epistemologies in the context of instruction in lecture courses. However, laboratory courses also offer significant and potentially unique opportunities for students to engage in the core practices and ideas of physics. Indeed, developing students' E&E has been called out as an important goal of laboratory science courses by multiple national organizations including the American Association of Physics Teachers [9,10], the National Research Council [11], and the President's Council of Advisors on Science and Technology [12]. These calls emphasize that effective laboratory instruction should help students develop expertlike habits of mind, experimental strategies, enthusiasm, and confidence in research.

To support ongoing, nationwide initiatives to improve laboratory instruction within physics based on these recommendations, researchers at the University of Colorado Boulder (CU) recently developed the Colorado Learning Attitudes about Science Survey for experimental physics (E-CLASS) [13]. E-CLASS is a 30-item, Likert-style survey designed to measure students' epistemologies and expectations about experimental physics (see Supplemental Material for a list of all item prompts [14]). Items on the E-CLASS feature a paired question structure in which students are presented with a statement and asked to rate their level of agreement from both their personal perspective and that of a hypothetical experimental physicist (see Fig. 1) [15]. The instrument was developed to target

Calculating uncertainties usually helps me under-stand my results better.

| | Strongly disagree | 1 | 2 | 3 | 4 | 5 | Strongly agree |

What do YOU think when doing experiments for class?

What would experimental physicists say about their research?
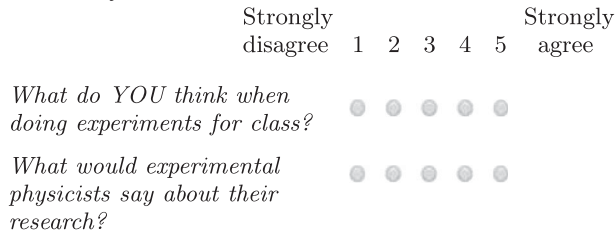
FIG. 1. An example item from the E-CLASS. Students are asked to rate their agreement with the statement from their own perspective and that of an experimental physicist. See Supplemental Material for a list of all item prompts [14].

explicit learning goals articulated as part of laboratory course transformations taking place at CU [16].

The development and initial validation of the E-CLASS included both iterative faculty review and student think-aloud interviews [13]. Twenty-three physics experts reviewed and responded to the E-CLASS survey. These responses were used to confirm that the instrument effectively targeted the desired learning goals and determine the consensus, expertlike response. Additionally, 42 interviews were conducted in which students completed the survey while explaining their reasoning out loud. The E-CLASS was modified based on these interviews in order to ensure the prompts were clear and correctly interpreted. For example, the prompt "What do YOU think when doing experiments for class?" was reworded from the original, simpler version "What do YOU think?" The addition of the phrase "when doing experiments for class" was motivated by evidence from interviews that students with prior research experience would sometimes pull from their personal research experiences, rather than their laboratory course experience, when responding to the survey [13].

The current version of the E-CLASS has now been administered as an online pre- and postsurvey for five semesters at CU and multiple other institutions across the US. In this article, we build on the initial validation of the E-CLASS through analysis of this national data set in order to establish the full statistical validity and reliability of the instrument for a broad student population. After a general overview of the data collection and scoring of the E-CLASS (Sec. II), we present the results of a detailed statistical analysis of students' responses, including test and item scores (Sec. III A), reliability (Sec. III B), validity (Sec. III C), and principal component analysis (Sec. IV). We end with a discussion of limitations and future work (Sec. V). Throughout this article, we will focus exclusively on analyses related to establishing the statistical validity of the E-CLASS. Analysis of the data to address broader research questions (e.g., impact of different pedagogical techniques on E-CLASS scores) will be the subject of future publications.

II. METHODS

In this section, we describe the methods used for collection, scoring, and analysis of student responses to the E-CLASS. We also present the general demographic breakdown of our final data set.

A. Data sources

All data for this study were collected between January 2013 and June 2015, and, in all courses, the E-CLASS was administered online. The survey was hosted exclusively at CU, and all student responses were collected directly by the CU research team without needing to be collected by the course instructor first. To administer the E-CLASS, participating instructors completed a course information survey [17] in which they provided basic information about their course, institution, and pedagogy. After completing the course information survey, instructors received a link to the online pre- and postinstruction E-CLASS to distribute to their students. In most cases, the pre- and postsurvey remained open for the first and last seven days of the course, respectively.

Instructors were recruited to administer the E-CLASS in their courses through a variety of methods, including presentations at national professional meetings and Emails to professional list serves. Information on the instrument was also available on the E-CLASS Web site [17], providing an additional avenue for interested instructors to learn about the survey. Over the five semesters of data collection, we aggregated student responses to the pre- and/or postinstruction E-CLASS from 80 distinct courses spanning 49 institutions including CU. We have matched pre- and postinstruction data from 71 of these courses spanning 44 institutions. These institutions include a variety of different types from four-year colleges to Ph.D.-granting institutions (see Table I). Additionally, several institutions administered the E-CLASS multiple times in the same course during the five semesters of data collection resulting in a total of 103 separate instances of the E-CLASS in our matched set and 114 in the overall pre- and postinstruction data sets. Moreover, these courses span the full space of introductory and advanced labs. Table II shows the breakdown between first-year and beyond-first-year courses in the matched data set.

Certain student responses were eliminated from the final pre- and postsurvey data sets because they were identified

TABLE I. Number of institutions of different types for which we have either pre- *or* post-test responses or matched pre- *and* post-test responses to the E-CLASS.

	Two-year college	Four-year college	Master's granting	Ph.D. granting	Total
All	0	25	4	20	49
Matched	0	23	4	17	44

TABLE II.   Number of first-year and beyond-first-year courses in the matched data set. The number of students in the beyond-first-year courses is smaller in part because of the smaller class sizes typical of more advanced physics labs. The number of separate instances of the E-CLASS accounts for courses that administered E-CLASS more than once in the five semesters of data collection.

	Distinct courses	Separate instances	Number of students
First year	35	49	2433
Beyond first year	36	54	1158

TABLE IV.   Breakdown of students by major in the matched data sets ($N = 3591$). Note that "Physics" includes both physics and engineering physics majors, and "nonscience" includes both declared nonscience majors and students who are open option or undeclared. The exact distribution of majors in any specific course varies significantly based on the type and level of the course.

	Physics	Engineering (nonphysics)	Other science and math	Nonscience
Percent	23%	29%	40%	7%

as invalid. For a response to be considered valid, the student must (1) include at least two of the three student identifiers (first name, last name, and student ID number), (2) respond to multiple survey questions, and (3) respond appropriately to a filtering question. This filtering question asks students to select the option "Agree" (not "Strongly Agree") and was included in the survey to help eliminate responses from students who randomly selected answers without reading the questions. Valid student responses were then matched pre- to postinstruction using student ID number or, when student ID matching failed, first and last name.

The final breakdown of valid student responses overall and by gender is given in Table III. Table III does not include a breakdown of the racial demographics of our sample because these data were not collected. Starting in Fall 2015, the postinstruction E-CLASS will collect data on the race of the participants. Data on students' current major were also collected on the postsurvey. Table IV reports the breakdown of students by major in the matched data set. Here, engineering physics majors are included in the "Physics" category and undeclared majors are included in the "Nonscience" category.

To get an idea of the overall response rate, we compare the number of E-CLASS responses to the number of students enrolled in the course. Total enrollment was reported by the instructor on the course information survey; however, this number represents only a rough estimate of the actual enrollment because many instructors complete the survey prior to the start of the term when the number of students may still fluctuate. Moreover, the reported enrollment is likely an overestimate as it reflects the initial

enrollment of the course and does not take into account those students who drop the course during the term. The overall, average response rate for our sample was 75% for the pretest and 64% for the post-test. Response rates for individual courses varied significantly. Factors that may have impacted the response rate include the level of incentive provided by the instructor (i.e., normal course participation credit, extra credit, or no credit for completion), when the instructor distributed the link, and how the instructor framed the activity to the students.

In addition to students' responses to the E-CLASS, we also collected student grade data from two semesters of the four core physics laboratory courses at CU in order to address the relationship between E-CLASS scores and laboratory course grade (Sec. III C 3). These courses span both the lower- and upper-division level, and, in all cases, students were awarded only participation credit for completing the survey. Final letter grades were collected for all students who agreed to release their grade data. Only students with matched postinstruction E-CLASS scores and final course grades were included in the grade analysis ($N = 873$).

B. Scoring

Students' numerical E-CLASS scores are determined only by their responses to the prompt targeting their personal beliefs, rather than their prediction of what an experimental physicist would say (see Fig. 1). Moreover, for the purpose of validating the survey, we will focus exclusively on students' pre- and postinstruction scores rather than their shifts from pre- to posinstruction. The motivation for this is to minimize the potential for confounding the validation of the survey with the impact of instruction. This also allows individual implementations of the pre- or postinstruction E-CLASS to provide valid and comparable results that can be used, for example, as baseline data.

For scoring purposes, students' responses to each 5-point Likert item are condensed into a standardized, 3-point scale in which the responses "(dis)agree" and "strongly (dis)agree" are collapsed into a single category. Responses are then given a numerical score based on whether they are

TABLE III.   Number of students in the final pretest, post-test, and matched data sets. Note that gender data were collected only on the postinstruction E-CLASS, and a small number of students opted not to provide their gender.

	Total N	Female	Male
Pretest	5658	…	…
Post-test	4732	35%	62%
Matched pre- and post-test	3591	36%	62%

consistent with the consensus expert response. The expert response can be either agree or disagree depending on particular item [13], and thus, student responses to individual items are coded simply as favorable (+1), neutral (0), or unfavorable (−1). The collapsing of the 5-point scale to 3 points is common in the analysis of Likert-style items and is motivated, in part, by the inherently ordinal, rather than interval, nature of the Likert response scale [18]. The use of the 3-point scale is also supported by previous literature suggesting that the threshold between "Agree" and "Strongly agree" is not always consistent between individual students or groups with different cultural backgrounds [19].

Previous literature on E&E surveys has often defined a student's overall score as the fraction of items to which they responded favorably [3]. This 2-point scoring scheme treats neutral and unfavorable responses the same. However, we consider the distinction between a neutral and an unfavorable response to be valuable and argue that the overall score should include this distinction. Thus, students' overall E-CLASS scores are given by the sum of their scores on the individual items on the 3-point scale described above. This results in a range of possible scores from −30 to 30 points. To explore the impact of different scoring conventions, we performed all of the analyses described in Sec. III using both the 2-point and 3-point scoring schemes and found that the two schemes resulted in the same conclusions with respect to the validity and reliability of the E-CLASS.

C. Analysis

There are multiple potential approaches to the analysis of multiple-choice tests [20]. We utilized two of these approaches here: classical test theory (CTT) [21] and principal component analysis [22]. CTT establishes the validity and reliability of a multiple-choice assessment based on the assumption that a student's score is composed solely of both their true score along with some random error [20]. Validation of an assessment via CTT involves analysis of student responses to calculate multiple test statistics and evaluation of these statistics relative to accepted thresholds (see Sec. III). One major drawback to CTT stems from the fact that all test statistics are calculated using student responses and, thus, are dependent on the specific population of students. As a consequence, there is no guarantee that test statistics calculated for one student population (e.g., undergraduate students) will hold for another population (e.g., high school students). For this reason, scores on assessments validated through the use of CTT can only be clearly interpreted to the extent that the student population matches the population with which the assessment was validated [23].

One alternative to CTT for establishing the statistical validity of a multiple-choice assessment is item response theory (IRT). IRT addresses many of the shortcomings of CTT by providing a method for producing population-independent estimates of both item and student parameters [20,23]. In the simplest IRT models, a student's performance on an item is assumed to depend only on their latent ability and the item's difficulty. For test items that fit this model, all item and student parameters calculated via IRT are independent of both population and test form [20,24]. However, IRT models also assume that the assessment is unidimensional (i.e., designed to measure a single construct). The E-CLASS, on the other hand, was explicitly designed to target multiple, potentially nonoverlapping aspects of students' beliefs about the nature of experimental physics, including modeling, statistical analysis, troubleshooting, etc. [13]. For this reason, we have chosen to utilize CTT, rather than IRT, to establish the statistical validity of the E-CLASS.

Literature on existing E&E surveys often groups questions into categories (a.k.a. factors) and reports students' scores in each of these categories. In some cases, this categorization was based on a priori criteria imposed by the developer (see, e.g., Ref. [4]). In other cases, the categorization was based on statistical analyses such as factor analysis, which identified statistically robust categories of questions (see, e.g., Refs. [3,25]). However, items on the E-CLASS were not specifically developed to match a specific categorization scheme, but rather to target a wide range of individual learning goals. In other words, the E-CLASS was not designed with specific categories in mind; thus, we have no a priori cause to believe that the E-CLASS would exhibit strong factors.

In order to determine whether or not the E-CLASS can be adequately characterized by a relatively small number of question groups, we utilize principal component analysis (PCA). PCA is a data reduction technique used to reduce the number of variables in a data set while still capturing a significant fraction of the variance [22,25]. PCA is typically used in data sets where there is reason to believe that there is significant redundancy among the variables [22] and uses an interitem correlation matrix to identify groups of items that appear to vary together. We performed a PCA on students' responses to both the pre- and postinstruction E-CLASS from the matched data set using the statistical software package R [26].

III. RESULTS: STATISTICAL VALIDITY

This section presents evidence for the statistical validity and reliability [21] of the E-CLASS using the preinstruction, postinstruction, and matched data sets described in Sec. II A. The results of the principal component analysis will be discussed in Sec. IV.

A. Test and item scores

As described in Sec. II B, a student's overall E-CLASS score is given by the sum of their scores on each of the
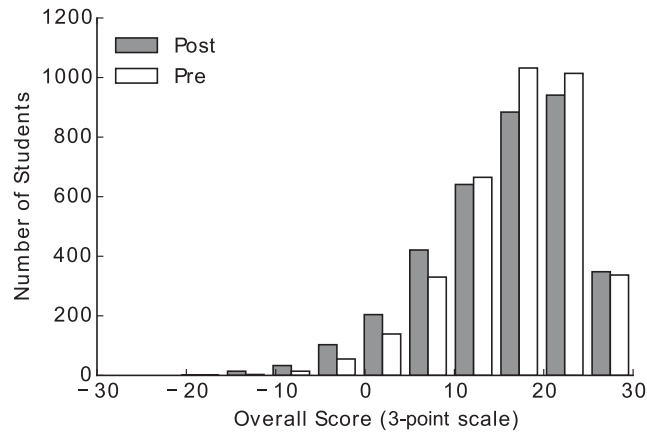
FIG. 2. Distribution of students pre- and postinstruction E-CLASS scores for students with matched pre- and postinstruction scores ($N = 3591$). The average E-CLASS score was 16.5 ± 0.1 points ($\sigma = 6.8$) for the pretest and 15.4 ± 0.1 points ($\sigma = 7.9$) for the post-test. The difference between the two distributions is statistically significant (Mann-Whitney U [27], $p \ll 0.05$).

30 items where each item is scored on a 3-point scale (favorable $= +1$, neutral $= 0$, unfavorable $= -1$). Figure 2 shows the distribution of overall E-CLASS scores for the matched data set ($N = 3591$). The difference between the pre- and postdistributions is statistically significant (Mann-Whitney U [27], $p \ll 0.05$), though the effect size is small (Cohen's $d = 0.14$ [28]). The students' overall performance on the E-CLASS can also be summarized by looking at the average fraction favorable relative to the average fraction unfavorable. Table V reports these statistics for the matched pre- and postinstruction E-CLASS.

In addition to looking at the overall E-CLASS score, we can also examine students' scores on each item individually. A sorted graph of the average pre- and post-test scores for each of the 30 E-CLASS items is given in Fig. 3. Figure 3 also highlights questions for which the difference between the pre- and postinstruction scores is statistically significant. Statistical significance was determined at the $\alpha = 0.05$ level, and p values were corrected for multiple-testing effects using the Holm-Bonferroni method [29]. While there is no standard criteria for acceptable item scores on Likert items [30], it is typically argued that ideal

TABLE V. Average fraction of items with favorable and unfavorable responses for the matched pre- and postinstruction E-CLASS.

		Average	Standard error	(Standard deviation)
Favorable	Pre	0.69	0.002	(0.15)
	Post	0.68	0.003	(0.16)
Unfavorable	Pre	0.15	0.002	(0.10)
	Post	0.17	0.002	(0.12)

item scores should be targeted towards maximizing the potential discriminatory power of each item and the test as a whole [31]. Figure 3 shows that the average item scores for the 30 E-CLASS items all fall between -0.4 and 0.96. This wide range suggests that the E-CLASS is capturing a significant amount of variation in students' epistemologies and expectations about experimental physics.

B. Reliability

The reliability of an assessment relates to the instrument's ability to measure the targeted construct(s) *consistently*. In this section, we present measures of several different aspects of the reliability of the E-CLASS including: test-retest reliability, time-to-completion reliability, partial-sample reliability, internal consistency, and testing environment (i.e., in-class versus online) reliability.

1. Test-retest reliability

The test-retest reliability of an assessment looks at the stability of students' scores. In other words, if students were to take the test twice with no intervention, they should, in theory, get the same score. Straightforward measures of test-retest reliability are difficult to achieve for a number of logistical reasons, including maturation and retest effects. One proxy that can be used to establish the test-retest reliability of the E-CLASS is the stability of preinstruction scores from semester to semester of the same course. As the students in a particular course are drawn from the same population each semester, we can reasonably expect that the preinstruction scores for that course should remain relatively stable over time.

During the five semesters of data collection, we collected matched pre- and postinstruction data for two or more semesters of 16 different courses at 11 institutions. Of these, only five courses showed statistically significant differences between the average preinstruction scores between semesters. In all cases, these differences stemmed from a single semester with anomalously high or low scores, and the effect sizes ranged from small (two courses, $d = 0.2$) to large (three courses, $d > 0.5$). However, while we expect the preinstruction population of a course to be relatively stable, it is also reasonable to expect that there would be small, legitimate variations in this population from cohort to cohort. The small number of statistically significant variations detected in our pretest data is consistent with this, thus supporting the test-retest reliability of the E-CLASS.

Another proxy for the test-retest reliability of the E-CLASS looks at whether students' scores shift from the beginning to the end of the semester when they are not enrolled in a laboratory course. To investigate the stability of students' scores under the null condition, we administered the E-CLASS pre- and postsurveys to one instance of the middle-division classical mechanics course at
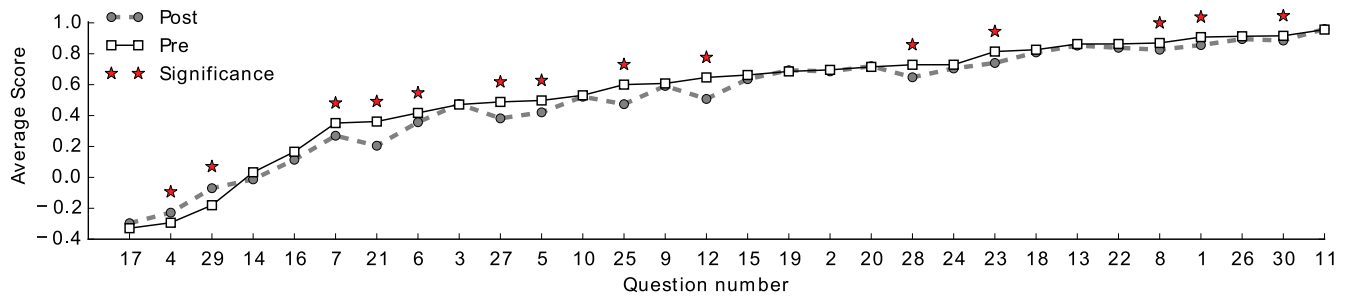
FIG. 3.   Average item scores using the 3-point scoring scale for each of the 30 E-CLASS items. Items are sorted in ascending order by scores on the preinstruction E-CLASS. Statistically significant differences between the pre- and postinstruction averages are indicated by a red star (Holm-Bonferroni corrected $p < 0.05$). See Supplemental Material for a list of individual prompts [14].

CU ($N = 49$). Given the standard course sequence at CU, the majority of the students in this course are physics and engineering physics majors and are not taking a physics laboratory course at the same time. To identify any off-sequence students, the postinstruction E-CLASS for this course was modified to ask students to report which (if any) laboratory courses they were taking that semester.

We collected 38 matched responses to the E-CLASS from the classical mechanics course, ten of whom were concurrently enrolled in a physics laboratory course. The remaining students ($N = 28$) had an average preinstruction score of 17.6 ± 1.6 ($\sigma = 8.5$) and a postinstruction score of 17.6 ± 1.5 ($\sigma = 7.8$). The difference between these scores is not statistically significant, suggesting that these students' epistemologies did not shift over the course of a semester in which they were not enrolled in a laboratory course. Alternatively, the ten students who were concurrently enrolled in a physics laboratory course showed a slight decrease in their scores (22.2 points to 21.6 points) over the semester. With only ten students, this decrease is not statistically significant; however, the direction and magnitude of the shift are consistent with that in the overall population (Fig. 2). This finding also preliminarily supports the test-retest reliability of the E-CLASS; however, additional data from other courses and institutions will be necessary to robustly establish the stability of students' E-CLASS scores under the null condition.

2. Time-to-completion reliability

It is also worth examining whether there is any correlation between the amount of time each student spends on the survey and their final score. In most cases, the E-CLASS was administered as an out-of-class online assignment, and start and stop times for each student were automatically collected by our online system for all semesters except one. The time elapsed is defined as the amount of time between when the student first clicked on the survey link to their final submission. Between these two times, however, a student may, for example, step away from the computer or close out the survey and start again at a later time. For this reason, the time elapsed does not

necessarily represent the amount of time the student actually spent completing the survey. The median time elapsed was 8 min for the pretest and 12 min for the post-test. The correlation between overall score and time elapsed was $|r| < 0.02$ for both pre- and post-tests. This correlation is neither statistically nor practically significant, suggesting that there was no link between time to completion and E-CLASS score for the matched data set.

3. Partial sample reliability

For any assessment with less than 100% response rate, it is also important to keep in mind the potential for selection effects when interpreting students scores. Because students essentially self-select into those who respond and those who do not, the sample of respondents may end up overrepresenting certain subpopulations (e.g., high performers) and underrepresenting others (e.g., low performers). To examine one aspect of the partial-sample reliability of the E-CLASS, we compare the average overall score for the matched data set with the average overall score of students who took only either the pre- or postsurvey. We found no statistically significant difference in average preinstruction scores for the matched and unmatched data sets. This same difference for the postinstruction scores was statistically significant ($p \ll 0.05$) with the unmatched students scoring slightly below the matched, but the effect size was small ($d = 0.2$).

In addition to looking at the partial-sample reliability of the matched data relative to the unmatched data, we can also examine the population of respondents and nonrespondents relative to other measures of student performance. In order to establish the convergent validity of the E-CLASS (see Sec. III C 3), we collected course grade data for all students in CU's four core laboratory courses over two semesters during which we also administered the E-CLASS. From these data, we can compare the average course grade of students who completed both the pre- and postinstruction E-CLASS ($N = 875$) with those who completed only one or neither ($N = 459$). We found that students in the matched data set had an average laboratory course grade of 3.3 (on a 4-point scale) while the average

for those who completed only one or neither was 2.7. The difference in final grade between the matched and unmatched students is statistically significant and represents a relatively large effect size ($d = 0.7$). This suggests that low response rates likely result in an underrepresentation of lower performing students.

This selection effect is unsurprising, but is an important factor for instructors and researchers to keep in mind when interpreting the results of the E-CLASS for courses with lower response rates. The oversampling of the higher performing students may suggest that results from courses with low response rates should be interpreted as a best case scenario snapshot of the overall student population's epistemologies.

4. Internal consistency

Another aspect of test reliability is the consistency of students' scores on arbitrary subsets of items. For a unidimensional test, Cronbach's alpha is a measure of this type of internal consistency [32]. Cronbach's alpha can be loosely interpreted as the average correlation between all possible split-half exams. For the purposes of low-stakes testing of individuals, a minimum value of $\alpha = 0.8$ is considered acceptable [21]. For the matched data set, we found $\alpha = 0.76$ for the presurvey and $\alpha = 0.83$ for the postsurvey. However, the interpretation of Cronbach's alpha generally assumes a unidimensional test targeting a single construct, and multidimensionality within the assessment will tend to drive alpha downward [32]. The E-CLASS, alternatively, was not designed to be unidimensional, but rather to target multiple, potentially overlapping, aspects of students' ideas; thus, Cronbach's alpha represents a conservative estimate of the internal consistency of the E-CLASS. The potential multidimensionality of the E-CLASS will be discussed in more detail in Sec. IV.

5. Testing environment reliability

In the majority of courses, students complete the E-CLASS outside of class time; however, a subset of instructors have students complete the assessment during class time, usually in order to increase participation. Giving the E-CLASS in class is most viable in courses that take place in a computer lab in which there is one student per computer. To investigate students' scores in different testing environments (i.e., in class versus out of class), we administered the E-CLASS during class time to the first-year lab course at CU in the fall of 2015 ($N = 521$). We compared the average overall score from this course with that from the same course in the five previous semesters ($N = 1568$). We found a small (Cohen's $d = 0.16$) but statistically significant (Mann-Whitney U, $p = 0.003$) increase in students' preinstruction E-CLASS scores from the full preinstruction data set for this course. However, the difference between preinstruction averages was not statistically significant for the subset of the

preinstruction data set for this course that had matched postinstruction responses. This suggests that for this population of students, taking the E-CLASS during class time had, at most, a small positive impact on their performance, though that increase did not persist in the matched sample. While this finding preliminarily supports the testing environment reliability of the E-CLASS, additional data from other institutions and courses will be necessary to clearly establish the impact of testing environment on students' scores.

C. Validity

The validity of an assessment relates to the instrument's ability to *accurately* measure the targeted construct(s). In this section, we present measures of several different aspects of the validity of the E-CLASS, including discrimination (whole test and by item), concurrent validity, and convergent validity.

1. Discrimination

To examine the ability of the E-CLASS overall to discriminate between students, we used Ferguson's delta [20]. Roughly speaking, Ferguson's delta looks at how well scores are distributed over the full range of possible point values, in this case, -30 to 30 points. Delta can range from [0, 1], and anything above 0.9 indicates good discriminatory power [30]. For both the pre- and postinstruction E-CLASS, we found $\delta > 0.98$, well above the standard threshold.

We also examined the discrimination of each individual item by looking at students' scores on individual items relative to their performance on the E-CLASS as a whole. Figure 4 shows item-test correlations for each of the 30 E-CLASS items. Here, we adopt the standard threshold for an item-test correlation for dichotomously scored items, $r > 0.2$ [30]. The majority of the E-CLASS items and the average postinstruction item-test correlation ($r = 0.33$) fall above this threshold. These results support the conclusion that the E-CLASS demonstrates adequate whole-test and item discrimination.

2. Concurrent validity

Concurrent validity examines the extent to which E-CLASS scores are consistent with certain expected results. For example, it is reasonable to expect that students' scores will vary between different levels of courses. In particular, first-year courses are often service courses catering primarily to engineering, biology, or nonscience majors, rather than physics majors. Thus, the learning goals of these courses may be less aligned with some of the learning goals targeted by E-CLASS, which were developed in collaboration with physics faculty to capture their desired learning outcomes for physics students in their upper-division lab courses [13]. Moreover, we
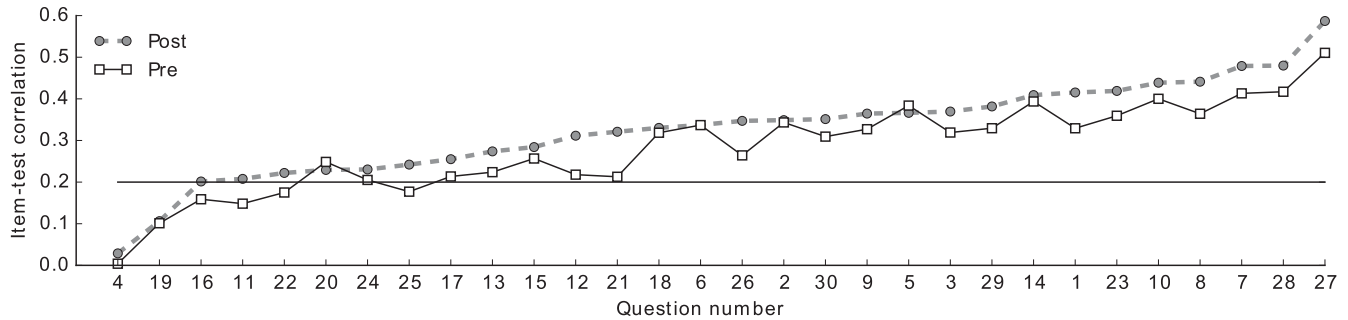
FIG. 4.   Item-test correlations using the 3-point scoring scale for each of the 30 E-CLASS items. Items are sorted in ascending order by postinstruction r values. The standard threshold for an acceptable item-test correlation is noted as a horizontal line at $r = 0.2$. See Supplemental Material for a list of individual prompts [14].

anticipated that students in the higher level courses would have more expertlike responses, due to either selection effects or the cumulative impact of instruction, or some combination of both. To investigate this, we divided the students into two subgroups composed of those in first-year (FY) labs ($N = 2433$) and those in second-, third-, and fourth-year labs ($N = 1158$), which we will refer to as the beyond-first-year (BFY) labs. This division between FY and BFY labs is consistent with the classification of courses used in the community of laboratory instructors and provides a relatively clear distinction between courses that is applicable at both CU and other institutions.

The average pre- and postinstruction E-CLASS scores for the FY and BFY courses are given in Table VI. The difference between the E-CLASS scores of the FY and BFY courses is statistically significant ($p \ll 0.05$) for both the pre- and postsurveys. Thus, students in the BFY courses both begin and end the semester with more expertlike views than students in the FY course. In addition to expecting students in the BFY courses to score higher on the E-CLASS, we might also anticipate that physics majors will score higher than nonphysics majors even within the FY courses. Here, we include engineering physics majors in our population of physics majors. Average pre- and postinstruction E-CLASS scores for FY physics and nonphysics majors are given in Table VII and show that FY physics majors both begin and end with significantly more expertlike views than nonmajors. Both of these findings are consistent with

expectations and support the concurrent validity of the E-CLASS.

3. Convergent validity

Convergent validity looks at whether scores on an assessment correlate with other, related student outcomes [33]. For conceptual assessments, convergent validity is typically established relative to students' final course grades. However, for E&E assessments like the E-CLASS, it is reasonable to ask if we expect the same level of correlation with course performance [4,34], particularly given that promoting expertlike attitudes and beliefs is rarely an *explicit* goal of physics courses. Of the available E&E assessments, only the VASS (Views about Science Survey) has published results reporting a modest but significant correlation ($r \sim 0.3$) between VASS score and final course grades in high school and college physics courses [19]. For the purposes of convergent validity, we will focus exclusively on students' postinstruction E-CLASS scores as these are the most likely to be consistent with final course grades. Grade data were collected from two semesters of the four core laboratory courses at CU, and students were assigned a standard grade point value for each letter grade (i.e., A = 4.0, A− = 3.7, B+ = 3.3, B = 3.0, etc.). Aggregating across all students in all courses ($N = 873$), we found an overall correlation coefficient of $r = 0.04$ between final course grade and postinstruction E-CLASS score. This correlation is neither practically nor statistically significant ($p = 0.2$).

TABLE VI.   Average overall scores for FY and BFY courses. Differences between FY and BFY averages are statistically significant (Mann-Whitney U, $p \ll 0.05$).

		Average (points)	Standard error	Significance	Effect size
FY	Pre	15.8	0.1	$p \ll 0.05$	0.2
($N = 2433$)	Post	14.4	0.2		
BFY	Pre	17.7	0.2	$p = 0.3$	0.04
($N = 1158$)	Post	17.5	0.2		

TABLE VII.   Average overall scores for first-year physics and nonphysics majors. Differences between the majors are all statistically significant (Mann-Whitney U, $p \ll 0.05$).

		Average (points)	Standard error	Effect size
Physics	Pre	19.4	0.4	0.1
($N = 215$)	Post	18.6	0.5	
Nonphysics	Pre	15.5	0.1	0.2
($N = 2218$)	Post	14.0	0.2	

However, consistent with the previous section, which showed higher overall scores in BFY courses, it is also reasonable to expect that this correlation might vary between courses. To investigate this potential variability across courses, we divided the students into two subgroups composed of those in the FY lab ($N = 717$) and those in the second-, third-, and fourth-year (BFY) labs ($N = 156$). For students in the FY lab, the correlation between overall E-CLASS score and final grade is small and not statistically significant ($r = 0.004$, $p = 0.9$). However, for the BFY labs, this correlation increased to $r = 0.17$ and is statistically significant ($p = 0.03$). This correlation, while still weak, is similar in magnitude to the correlations reported between CLASS or Maryland Physics Expectations (MPEX) survey scores and conceptual learning gains as measured by conceptual assessments such as the Force Concept Inventory [35–37]. We are not arguing that the relationship between E-CLASS scores and final grades is a causal one; however, these results do suggest that the link between course performance and epistemological stance is stronger in more advanced lab courses than in the first-year lab.

Overall, these findings suggest that, even for BFY courses, E-CLASS scores are not good predictors of a student's course grade (or vice versa). One interpretation of this is that students' epistemological knowledge and beliefs are not being effectively captured by their final course score. This finding is also consistent with results from the CLASS that have demonstrated neutral or even negative shifts in CLASS scores from courses with significant conceptual learning gains [38].

IV. PRINCIPAL COMPONENT ANALYSIS

In addition to looking at students' scores overall and by item, existing E&E surveys often examine students' aggregate scores on smaller groups of items. These item groupings are typically either based on an *a priori* categorization created by the survey developer [4] or empirically derived using a statistical data reduction technique [3]. As previously discussed, the E-CLASS was not developed to match any particular *a priori* categorization of questions; however, it is still possible that different questions on the E-CLASS target the same latent variable. In order to determine if this is the case, we utilize principal component analysis.

PCA is a type of exploratory factor analysis that attempts to reduce the number of variables in a data set by examining interitem correlations in order to identify groups of items that appear to vary together [22,25]. We first performed a PCA on students' responses to the postinstruction E-CLASS from the matched data set. The initial exploratory PCA produced 30 components (a.k.a. factors) along with associated eigenvalues. To determine how many of these components to extract, we adopted the Guttman-Kaiser criterion [39], which states that all components with eigenvalues greater than 1 should be kept. This criteria

resulted in seven components that together explained 45% of the overall variance in the survey. However, it is generally accepted that, to sufficiently represent the overall data set, the retained components should account for at least 70% of the variance. Meeting this threshold for the E-CLASS data would require retaining 16 of the 30 components. This factor of 2 decrease does not represent a useful reduction in the number of variables in the data set.

To determine if the factors identified in the postinstruction data were robust, we performed the same PCA on student responses to the preinstruction E-CLASS. We found that there was little overlap between the items that made up specific factors in the presurvey data compared to those from the postsurvey. This result, along with the fact that the results of the PCA accounted for less than half the overall variance, suggests that the E-CLASS does not exhibit a clear factorization. This lack of strong factors is not surprising given that PCA is used to identify cases in which there are multiple items targeting a single latent variable. The E-CLASS, on the other hand, was designed to target a relatively large number of distinct, though potentially overlapping, learning goals. The PCA suggests that, consistent with this design, the E-CLASS does not appear to contain groups of items targeting a single latent variable. Given this result, we strongly recommend that instructors do not focus only on their students' overall E-CLASS score, as it does not represent students' performance around a single well-defined construct. Rather, instructors should examine their students' responses to the questions individually with a particular focus on those questions that are most aligned with their learning goals for that course.

V. SUMMARY AND FUTURE WORK

We previously created an attitudinal survey—known as the E-CLASS—targeting students' epistemologies and expectations about the nature of experimental physics. Prior work established the content validity of this assessment through expert review and student interviews. To build on this initial validation, we collected student responses to the E-CLASS from roughly 80 courses spanning approximately 45 institutions. Analysis of these data supports the statistical validity and reliability of the E-CLASS as measured by multiple test statistics, including item and whole-test discrimination, internal consistency, partial-sample reliability, test-retest reliability, and concurrent validity. A principal component analysis of these data also showed that the E-CLASS does not exhibit robust factors that can be used to accurately and reliably describe students' responses using a smaller number of statistically consistent groups of questions.

Future work will include analysis of our growing national data set of student responses to the E-CLASS to answer broader research questions regarding students' ideas about the nature of experimental physics. For example, this data set includes some longitudinal data that

could begin to provide insight into how students' epistemologies change over the course of their undergraduate career. Additionally, the course information survey, completed by all instructors prior to using the E-CLASS, collects information on both pedagogy and learning goals (e.g., learning physics content versus developing lab skills). These data can be used to determine if certain pedagogical strategies or learning goals promote more expertlike epistemologies and expectations. Future research will also include investigating if and how instructors use their students' E-CLASS results to inform their instruction or course transformation efforts.

ACKNOWLEDGMENTS

[1] L. Lising and A. Elby, The impact of epistemology on learning: A case study from introductory physics, Am. J. Phys. **73**, 372 (2005).

[2] D. Hammer, Epistemological beliefs in introductory physics, Cognit. Instr. **12**, 151 (1994).

[3] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[4] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

[5] A. Elby, J. Frederiksen, C. Schwarz, and B. White, Epistemological beliefs assessment for physical science (EBAPS) (2001).

[6] I. Halloun, Views about Science and Physics Achievement: The VASS Story, AIP Conf. Proc. **399**, 605 (1996).

[7] F. Abd-El-Khalick, N. G. Lederman, R. L. Bell, and R. S. Schwartz, Views of Nature of Science Questionnaire (VNOS): Toward Valid and Meaningful Assessment of Learners' Conceptions of Nature of Science, J. Res. Sci. Teach. **39**, 497 (2001).

[8] S. Chen, Development of an instrument to assess views on nature of science and attitudes toward teaching science, Sci. Educ. **90**, 803 (2006).

[9] AAPT Committee on Laboratories, AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum (2015).

[10] R. Feynman, Goals of the introductory physics laboratory, Am. J. Phys. **66**, 483 (1998).

[11] National Research Council Undergraduate Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century *et al.*, *BIO2010: Transforming Undergraduate Education for Future Research Biologists* (National Academies Press, Washington, DC, 2003).

[12] S. Olson and D. G. Riordan, Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics, Executive Office of the President (2012).

[13] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).

[14] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.12.010123 for a list of all E-CLASS prompts along with an example to demonstrate how the questions are formatted.

[15] K. E. Gray, W. K. Adams, C. E. Wieman, and K. K. Perkins, Students know what physicists believe, but they don't agree: A study using the CLASS survey, Phys. Rev. ST Phys. Educ. Res. **4**, 020106 (2008).

[16] B. M. Zwickl, N. Finkelstein, and H. Lewandowski, The process of transforming an advanced lab course: Goals, curriculum, and assessments, Am. J. Phys. **81**, 63 (2013).

[17] http://tinyurl.com/ECLASS-physics.

[18] M. Lovelace and P. Brickman, Best practices for measuring students' attitudes toward learning science, CBE Life Sci. Educ. **12**, 606 (2013).

[19] I. Halloun, Student views about science: A comparative survey, in *International Conference on Undergraduate Physics Education, Educational Research Center, Lebanese University, Beirut, Lebanon* 5, 3 (2001).

[20] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

[21] P. Engelhardt, An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests, *Getting Started in PER* (American Association of Physics Teachers, College Park, 2009), Vol. 2.

[22] N. O'Rourke, R. Psych, and L. Hatcher, *A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling* (SAS Institute, Cary, 2013).

[23] C. S. Wallace and J. M. Bailey, Do concept inventories actually measure anything, Astron. Educ. Rev. **9**, 010116 (2010).

[24] F. B. Baker, *The Basics of Item Response Theory* (ERIC Clearinghouse on Assessment and Evaluation, College Park, 2001).

[25] C. Lindstrøm and M. D. Sharma, Initial development of a Physics Goal Orientation survey using factor analysis, in *Proceedings of The Australian Conference on Science and Mathematics Education* (UniServe Science, Sydney, 2012).

[26] https://www.r-project.org/.

[27] H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math Stat. **18,** 50 (1947).

[28] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, 2013).

[29] S. Holm, A simple sequentially rejective multiple test procedure, *Scand. J. Stat. Theory Appl.* **6,** 65 (1979).

[30] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2,** 010105 (2006).

[31] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction* (ERIC, 1980).

[32] J. M. Cortina, What is coefficient alpha? An examination of theory and applications, J. Appl. Psych. **78,** 98 (1993).

[33] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33,** 1289 (2011).

[34] D. Hammer, Two approaches to learning physics, Phys. Teach. **27,** 664 (1989).

[35] K. Perkins, W. Adams, S. Pollock, N. Finkelstein, and C. Wieman, Correlating student beliefs with student learning using the Colorado Learning Attitudes about Science Survey, AIP Conf. Proc. **790,** 61 (2005).

[36] S. J. Pollock, No single cause: learning gains, student attitudes, and the impacts of multiple effective reforms, AIP Conf. Proc. **790,** 137 (2005).

[37] G. Kortemeyer, Correlations between student discussion behavior, attitudes, and learning, Phys. Rev. ST Phys. Educ. Res. **3,** 010101 (2007).

[38] S. Pollock and N. Finkelstein, Sustaining Change: Instructor Effects in Transformed Large Lecture Courses, in *Proceeding of the Physics Education Research Conference, Syracuse, NY, 2006* (AIP Conference Proceedings, 2006), Vol. 883, pp. 109–112.

[39] K. A. Yeomans and P. A. Golder, The Guttman-Kaiser criterion as a predictor of the number of common factors, The Statistician **31,** 221 (1982).