Colorado Upper-Division Electrostatics diagnostic: A conceptual assessment for the junior level

Stephanie V. Chasteen,[1] Rachel E. Pepper,[2] Marcos D. Caballero,[1] Steven J. Pollock,[1] and Katherine K. Perkins[1]

[1]*Science Education Initiative, Physics Department, University of Colorado Boulder, UCB 390, Boulder, Colorado 80301, USA*

[2]*Department of Integrative Biology and Department of Civil and Environmental Engineering,*
University of California, Berkeley, California 94720, USA

(Received 2 November 2011; published 7 September 2012)

As part of an effort to systematically improve our junior-level E&M I course, we have developed a tool to assess student conceptual learning of electrostatics at the upper division. Together with a group of physics faculty, we established a list of learning goals for the course that, with results from student observations and interviews, served as a guide in creating the Colorado Upper-Division Electrostatics (CUE) assessment. The result is a 17-question open-ended post-test diagnostic (with an optional 7-question pretest) and an accompanying grading rubric. We present measures of the validation and reliability of the instrument and grading rubric, plus results from 535 students in both standard and interactive-engagement courses across seven institutions as a baseline for the instrument. Overall, we find that the CUE is a valid and reliable measure, and the data herein are intended to be of use to researchers and faculty interested in using the CUE to measure student learning.

DOI: 10.1103/PhysRevSTPER.8.020108 PACS numbers: 01.40.G−, 01.40.Di, 01.40.Fk, 01.40.gf

I. INTRODUCTION

The physics education research (PER) community has investigated student understanding of introductory physics topics, including electromagnetism (E&M), in some detail [1,2]. Efforts to improve student learning in lower-division courses have been driven in large part by data on student performance on research-based conceptual tests [3], such as the Force Concept Inventory (FCI) [4] in mechanics, or instruments which test student abilities in E&M such as the Conceptual Survey of E&M (CSEM) [5] or the Basic Electricity and Magnetism Survey (BEMA) [6]. Not only have these instruments served to identify common and persistent student difficulties, but they have also been powerful tools for supporting curricular reform. Detailed information on common student difficulties can drive more effective course reforms on the part of curriculum developers, as it has in introductory courses [7], and provide insight for faculty as to persistent student difficulties. The resulting materials may be valuable for instructors wishing to try a new approach to instruction, and student performance data on these instruments may (or may not [8]) convince faculty of the usefulness of alternative approaches.

Overall, we lack similar tools for improving instruction in the upper division. At most universities, including the University of Colorado Boulder (CU-Boulder), upper-division physics courses are often taught via traditional lecture and do not make use of effective instructional techniques [9] used at the introductory level. Research on how students understand the more mathematically and conceptually sophisticated treatment of E&M at the upper division [2,10,11] and the effectiveness of associated pedagogical strategies [12,13] is just beginning. A few concept inventories for upper-division quantum mechanics [14] and E&M [15] have been developed. The current work adds to this literature. We have undertaken a multiyear transformation of our upper-division E&M course, including identification of student difficulties, development of learning goals for the course, and student-centered instruction, the process of which is described elsewhere [16]. To assess the relative success of these transformations, as well as to document student difficulties, we developed a post-test for the course: The Colorado Upper-Division Electrostatics (CUE) assessment. Here, we report on the development, validation, and reliability of the CUE instrument to serve as a reference for the education research community.

II. METHODS AND CUE DEVELOPMENT

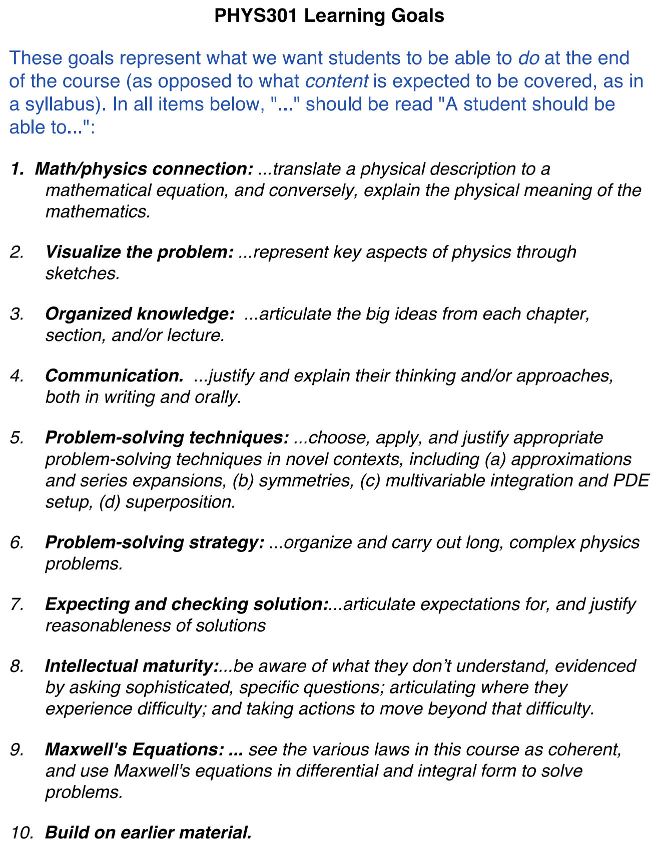
A. About the course

In 2007, the CU physics department chose to transform [17] one of the core courses that defines what it means to learn physics as a major—upper-division Electricity & Magnetism I (E&M I), typically taken in the Fall semester of the junior year. The course is offered every semester (with Fall representing the largest enrollment) and taught by a variety of instructors who rotate through course assignments. The content of the course is canonical for E&M I: electro- and magnetostatics, including techniques for solving for the potential and fields in matter. This covers Chaps. 1–6 of the text by Griffiths [18].

B. Faculty discussions and learning goals

We began the process of course transformation by forming a working group of 13 faculty with an interest in the course. This group included PER faculty ($N = 4$) and faculty involved in more traditional research ($N = 9$). We held seven biweekly informal ''brown bag'' meetings of faculty (described elsewhere [19]) and 13 individual faculty interviews. In these sessions we asked faculty to help us determine the learning goals, or what students need to be able to do by the end of the course. There was early broad agreement on the topical content of the course—namely, the topics covered in Chaps. 1–6 of Griffiths' textbook [18]. Upon this topical agreement, we turned to development of broader, course-scale learning goals. These learning goals, shown in Fig. 1, are intended to describe the skills expected of a junior-level physics student, and have been used as the basis for learning goals in several other courses [14,21]. The full set of learning goals is available online [20].

While some of these skills are tested in traditional exams (such as setting up and executing integrals), many of these metalevel techniques and their conceptual underpinnings are not. The CUE was developed to address this gap. Faculty in the working group were then consulted on the question development and refinement (see Sec. II D).



PHYS301 Learning Goals

These goals represent what we want students to be able to *do* at the end of the course (as opposed to what *content* is expected to be covered, as in a syllabus). In all items below, "..." should be read "A student should be able to...":

1. ***Math/physics connection:*** *...translate a physical description to a mathematical equation, and conversely, explain the physical meaning of the mathematics.*
2. ***Visualize the problem:*** *...represent key aspects of physics through sketches.*
3. ***Organized knowledge:*** *...articulate the big ideas from each chapter, section, and/or lecture.*
4. ***Communication.*** *...justify and explain their thinking and/or approaches, both in writing and orally.*
5. ***Problem-solving techniques:*** *...choose, apply, and justify appropriate problem-solving techniques in novel contexts, including (a) approximations and series expansions, (b) symmetries, (c) multivariable integration and PDE setup, (d) superposition.*
6. ***Problem-solving strategy:*** *...organize and carry out long, complex physics problems.*
7. ***Expecting and checking solution:****...articulate expectations for, and justify reasonableness of solutions*
8. ***Intellectual maturity:****...be aware of what they don't understand, evidenced by asking sophisticated, specific questions; articulating where they experience difficulty; and taking actions to move beyond that difficulty.*
9. ***Maxwell's Equations: ...*** *see the various laws in this course as coherent, and use Maxwell's equations in differential and integral form to solve problems.*
10. ***Build on earlier material.***

FIG. 1 (color online). Abbreviated course-scale learning goals for PHYS301, developed by the faculty working group. The full set of learning goals is available at [20].

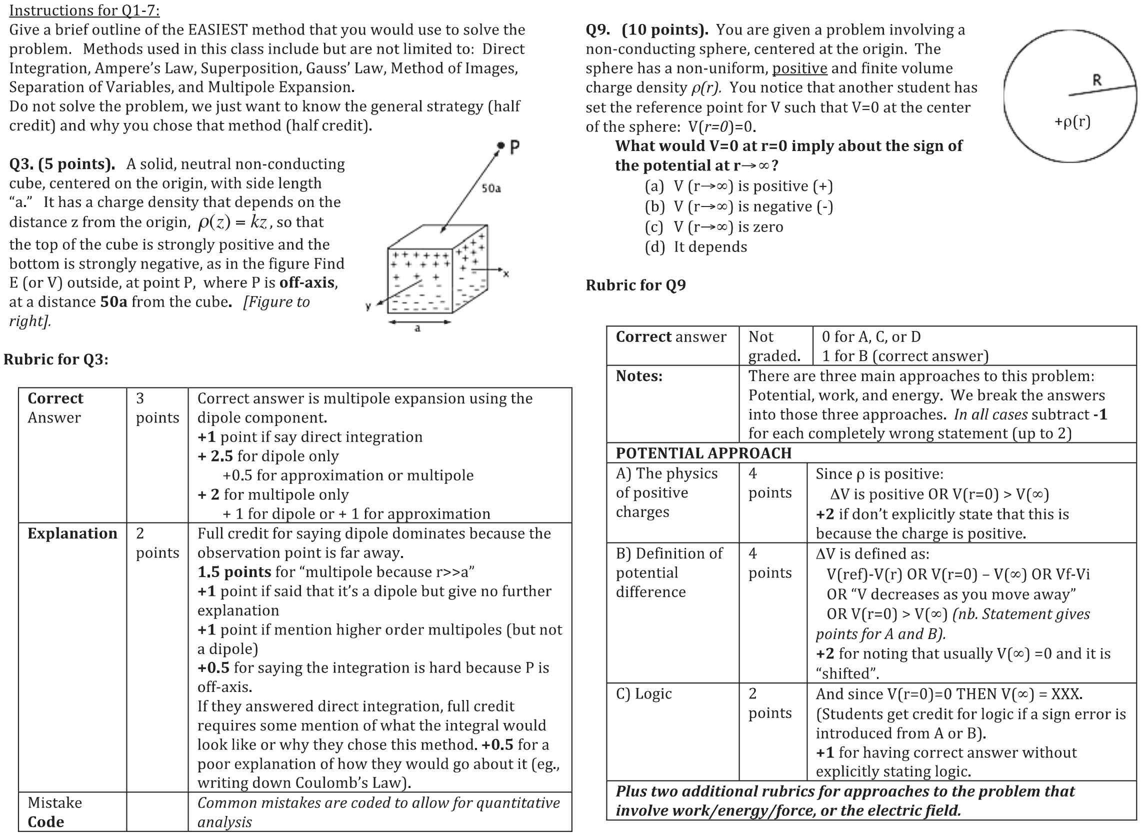
These learning goals capture what CU faculty think that E&M I is about, and what they care that students take away from the course. The final CUE exam was consistently guided by these learning goals, which defined the ''concept space'' that is probed by the exam. In the end, however, the instrument is intended to stand on its own as a general proxy of student achievement of faculty's learning goals, rather than providing a measurement of student achievement of any individual learning goal. For example, a series of questions (Q1–Q7; see Fig. 2 for an example) probe students' ability to recognize which problem-solving method is most appropriate for a given physical situation. These questions were motivated by learning goal no. 5 (''choose and justify appropriate problem-solving techniques''). In order to probe learning goal no. 2 (''visualize the problem''), one question (Q10) asks students to sketch both the induced charge on and induced electric field around a conductor in an external field. However, some learning goals are not addressed by the CUE. Learning goal no. 6 (''organize and carry out complex physics problems'') is, we felt, better probed by traditional calculational exam problems; learning goal no. 8 (''intellectual maturity'') is better tapped by student and faculty surveys and interviews. Overall, learning goals 3, 6, 8, and 9 are not addressed by the CUE instrument. CUE questions and associated learning goals are listed later in the paper (Table I).

C. Observations and interviews

While the CUE is intended to provide insight into student difficulties, we needed some informed ideas of where students struggle in order to write meaningful CUE questions. To give us a starting point on student difficulties for the creation of the CUE, one of the authors (S. V. C.) observed class sessions of a standard lecture-based course, ran homework help sessions, and conducted student interviews. A total of 47 student interviews were conducted in a think-aloud format to (a) identify common difficulties or refine our understanding of their difficulties and (b) pilot test and validate the CUE as questions were modified. Out of all interviews, 19 specifically focused on understanding student responses on the CUE or validating the questions on the instrument.

D. Creation and refinement of questions

We then collaborated with the faculty working group to develop initial questions for the CUE. Questions were based on these learning goals, brainstorming, as well as common difficulties observed during student interviews and homework help sessions [23]. While the learning goals guided the development of the CUE questions, the CUE was not designed to provide a definitive test of each learning goal. That is, the CUE does not provide a score for each


P
50a
x
y
a
R
+ρ(r)

Instructions for Q1-7:
Give a brief outline of the EASIEST method that you would use to solve the problem. Methods used in this class include but are not limited to: Direct Integration, Ampere's Law, Superposition, Gauss' Law, Method of Images, Separation of Variables, and Multipole Expansion.
Do not solve the problem, we just want to know the general strategy (half credit) and why you chose that method (half credit).

Q3. (5 points). A solid, neutral non-conducting cube, centered on the origin, with side length "a." It has a charge density that depends on the distance z from the origin, $\rho(z) = kz$, so that the top of the cube is strongly positive and the bottom is strongly negative, as in the figure Find E (or V) outside, at point P, where P is **off-axis**, at a distance **50a** from the cube. *[Figure to right].*

Rubric for Q3:

Correct Answer	3 points	Correct answer is multipole expansion using the dipole component. **+1** point if say direct integration **+ 2.5** for dipole only +0.5 for approximation or multipole **+ 2** for multipole only + 1 for dipole or + 1 for approximation
Explanation	2 points	Full credit for saying dipole dominates because the observation point is far away. **1.5 points** for "multipole because r>>a" **+1** point if said that it's a dipole but give no further explanation **+1** point if mention higher order multipoles (but not a dipole) **+0.5** for saying the integration is hard because P is off-axis. If they answered direct integration, full credit requires some mention of what the integral would look like or why they chose this method. **+0.5** for a poor explanation of how they would go about it (eg., writing down Coulomb's Law).
Mistake **Code**		*Common mistakes are coded to allow for quantitative analysis*

Q9. (10 points). You are given a problem involving a non-conducting sphere, centered at the origin. The sphere has a non-uniform, positive and finite volume charge density *ρ(r).* You notice that another student has set the reference point for V such that V=0 at the center of the sphere: V(*r=0*)=0.

What would V=0 at r=0 imply about the sign of the potential at r→∞?

(a) V (r→∞) is positive (+)
(b) V (r→∞) is negative (-)
(c) V (r→∞) is zero
(d) It depends

Rubric for Q9

Correct answer	Not graded.	0 for A, C, or D 1 for B (correct answer)
Notes:	There are three main approaches to this problem: Potential, work, and energy. We break the answers into those three approaches. *In all cases* subtract **-1** for each completely wrong statement (up to 2)	
POTENTIAL APPROACH		
A) The physics of positive charges	4 points	Since ρ is positive: ΔV is positive OR V(r=0) > V(∞) **+2** if don't explicitly state that this is because the charge is positive.
B) Definition of potential difference	4 points	ΔV is defined as: V(ref)-V(r) OR V(r=0) – V(∞) OR Vf-Vi OR "V decreases as you move away" OR V(r=0) > V(∞) *(nb. Statement gives points for A and B).* **+2** for noting that usually V(∞) =0 and it is "shifted".
C) Logic	2 points	And since V(r=0)=0 THEN V(∞) = XXX. (Students get credit for logic if a sign error is introduced from A or B). **+1** for having correct answer without explicitly stating logic.
Plus two additional rubrics for approaches to the problem that involve work/energy/force, or the electric field.		

FIG. 2. Two problems from the CUE, Q3 (5 points) and Q9 (10 points), and their accompanying grading rubrics.

student learning outcome, but rather represents a holistic measure that taps into student expertise as defined by faculty. Additionally, the CUE was designed to test a selection of the material in the course (the early electrostatics content), to serve as a litmus test for the overall performance of a student in the course, much as the Force Concept Inventory tests only a subset of the material in introductory mechanics. Thus, the CUE addresses only a limited subset of the essential material that faculty believe belongs in the course.

The selection and wording of questions was refined using the expertise of faculty and education researchers. The resulting instrument was pilot tested with a small group of five students, using a think-aloud format, and the test was administered to students enrolled in one course.

Questions were modified as we reviewed student performance on the diagnostic and evaluated the quality of responses on the test (i.e., did they understand what we were asking?) over six semesters of administration. After the second administration, six students were interviewed regarding their responses on the questionnaire. The penultimate version was validated by interviewing eight students using a think-aloud protocol, resulting in the final, validated, version of the test [24]. All reliability and validity data (see Sec. V) are reported on this version of the test, which was administered to 103 students in three courses. Previous versions of the instrument were administered to 432 students in 14 courses.

These steps match those listed by others [25,26] as the standard steps in instrument design, particularly the key importance of the identification of expert-thinking strategies and of student interviews in iterative test development and validation [25]. In the course of development, a total of five iterated versions of the test have been administered over time to various courses. Since the first version, a total of seven questions were dropped, two questions were added, and five questions were substantially modified, with most questions reworded for clarity.

As an example of a question that was substantially modified, Q16 shows a quadrupole distribution (see the Appendix). Early versions of the exams asked students "At point P, a distance r from the origin, how does V depend on r for r ≫ a?" We found that student responses did not allow us to differentiate between students who did not recognize the distribution as a quadrupole and students who did not know the $1/r^3$ dependence of a quadrupole. Thus, in later iterations, we simplified the question to "Is the dipole moment of this distribution zero" (Yes/No/Not Sure). Briefly explain your reasoning." However, student explanations often contradicted their choice, and interviews confirmed that students found this wording of the question confusing. Thus, the final version reads, "The

TABLE I. All problems on the CUE, including point allocations, Cohen's kappa (see Sec. III B), and learning goal alignment. Question numbers in italics indicate items which are on the pretest. Cohen's kappa values are designated as "moderate" (*; 0.41–0.60), "substantial" (**; 0.6–0.80), or "almost perfect" (***; 0.81–1.0), as determined by Landis and Koch [22]. Relevant learning goals are provided for illustrative purposes and are not intended to be comprehensive.

Q no.	Points	Short name	Description	Cohen's kappa	Related learning goals
Q1–Q7 all ask the student to identify the easiest method to solve for E or V at point P					
Q1	5	V of theta	Insulating sphere with $V(\theta) = k\cos(3\theta)$, P on surface.	0.81**	4, 5, 5c
Q2	5	Cube	Neutral polarized cube, P on axis.	0.57*	4, 5, 5c, 10
Q3	5	Cube far away	Neutral polarized cube, P off axis and far away.	0.67**	4, 5, 5a
Q4	5	Images	Grounded conducting plane with charge Q. P on same side as Q.	0.80**	4, 5
Q5	5	Superposition	Charged insulating sphere with off-center spherical cavity. P outside.	0.57*	4, 5, 5b, 5d, 10
Q6	5	Current loop	Current loop, point P off axis and far away.	0.73**	4, 5, 5a
Q7	5	Gauss	Nonconducting sphere with charge density $\rho(r) = \rho_0 e^{-r^2/a^2}$. P outside.	0.59*	4, 5, 5b, 10
Q8	5	Delta function	Mass density:$\rho(r) = m_1\delta^3(r - r_1) + m_2\delta^3(r - r_2)$. What is integral of ρ over all space, and what does this represent physically?	0.82***	1, 4
Q9	10	Reference for V	$V = 0$ is set at the center of a charged sphere. What is sign of the potential at infinity?	0.41*	1, 4, 7, 10
Q10	10	Sketch E cylinder	Sketch induced charge and E field for conducting cylinder placed in E field.	0.65**	2, 5d, 10
Q11	6	BCs on E and V	List boundary conditions for V and E needed to solve Q10 by separation of variables	0.62**	1, 5c
Q12	22	Graph E and V of disk	A uniformly charged infinitely thin disk. (A) What is E_z near the center of disk; (B) how does E_z behave as you move away from disk; (C) draw E_z vs z; (D) draw V vs z.	0.45*	2, 5a, 7, 10
Q13	5	Cartesian BCs	For a 2D box with specified potentials, identify whether sinusoidal and exponential form of solutions should be in X or Y.	0.73**	1, 4, 5b
Q14	8	Dielectric	Describe what happens to a dielectric when inserted into a capacitor, plus limiting case of an "infinitely polarizable" dielectric.	0.68**	2, 4, 5a, 7, 10
Q15	5	Circle BCs	Circle the appropriate boundary conditions to solve for V on spherical charged surface.	1.0***	5c
Q16	5	Multipole	Quadrupole distribution is shown; explain whether the dipole moment is zero.	0.55*	1, 4, 5b, 5d
Q17	7	Ampere	For an infinite cylinder with current density J, where is B field maximum?	0.72**	1, 4, 5, 5c, 10

dipole moment of this distribution is: (Zero/Non-zero/Not sure). Briefly explain your reasoning."

An example of a more minor wording change is on Q2 which asked students how they would find E at a point P "at a distance $\mathbf{r} = \mathbf{a}$" from a cube with side-length a. Interviews indicated that this wording caused some students to stumble, and was simplified to "at a distance $\mathbf{a}$ from the cube" (and similarly for Q3).

Questions that were dropped were typically those which we found were very problematic in their wording, with no clear path towards revising them to be more effective, or ones which we found were tapping concepts and skills that we decided were not our main focus. For example, Q12 asks students to sketch a graph of E and V above and below a charged disk. Originally, we also asked students to "explain the physical origin of the behavior of the graph near z = 0," with the intention of eliciting the math and physics connection that the discontinuity in the graph indicated the presence of a surface charge. However, students consistently misinterpreted the question, and we removed it from the exam. One question that was dropped in its entirety gave students a charge +Q in the hollow center of a two-dimensional metal square "frame" and asked students for the net charges on the interior and exterior surfaces of the "frame," to explain why the charges had that exact magnitude, and what would happen if the charge were moved off center. This item was dropped in order to limit the length of the test and because we felt it was more of a "clever" problem rather than directly tapping the learning goals developed by faculty for this course.

In order to provide a measure of student learning gains, a total of seven questions from the CUE were chosen (in collaboration with faculty) for the pretest. These seven questions use vocabulary and concepts that incoming students have previously encountered and thus may be able to attempt, though our data show that these questions are still very challenging for students after introductory physics, as well as incoming juniors entering E&MI.

E. About the CUE

The resulting exam is a 17-question conceptual test consisting of 15 electrostatics and 2 magnetostatics questions. Different numbers of points are assigned to different portions of the exam to represent their importance as learning goals, or the difficulty of the question, resulting in a total possible 118 points on the exam. All CUE scores reported are normalized out of 100%. A brief description of all questions is given in Table I along with learning goals related to that question. Two sample questions from the CUE are shown in Fig. 2 and the full instrument is given in the Appendix.

The assessment tests students' ability to choose a problem-solving method and defend that choice, sketch electric field patterns, graph electric field strength and potentials, and explain the physics and mathematics underlying steps in common problems. With the exception of one multiple-choice question the exam is open ended; three additional questions give students a multiple-choice alternative and require students to explain their answer to receive credit. The CUE post-test is typically given at the end of the semester, and requires approximately 50 minutes. The pretest is designed to be given during the first week of class, and requires approximately 20 minutes to complete. In-class administration is recommended on both tests to encourage consistent testing conditions. Typical student results on the CUE are discussed in Sec. V, and student performance on each question is given in the Appendix.

F. CUE administration

Some version of the CUE was given to a total of 535 students in standard (lecture-based) and transformed courses, both at CU-Boulder (seven courses) and at six external institutions (nine courses). Most of these students took early versions of the CUE developed prior to the ultimate version [24], and so "common" exam scores are created to allow us to compare student performance across different versions of the exams (see Sec. V). Typically, students were not informed of the test in advance, and it was always administered in class as part of the course, but not for course credit. Student data are excluded for students who dropped the course. The students in the different courses at CU were similar on many measures of preparation, such as cumulative grade-point average (GPA) (3.1–3.2), GPA in physics courses (2.9–3.2), and prerequisite courses (standard error of the mean ~0.1 for most courses). More information about the courses in the study is given in our other publications [13].

A total of 103 students in three courses taught by three different faculty at CU took the final form of the CUE. Students were given the exam at the end of the semester (with one exception, below). Course A ($N = 34$) was taught with some interactive methods, including some clicker questions and optional tutorials. Course B ($N = 22$) was taught mostly traditionally with semiregular clicker use, primarily for review. Course C ($N = 47$) was taught in a traditional lecture fashion, and students took the CUE at the beginning of the next semester.

G. CUE rubric and inter-rater reliability

Because the CUE is an open-ended exam, a detailed grading rubric was developed in order to explicitly define what points should be assigned to each question. This was a challenging task; the rubric needed to clearly specify the points to be assigned for a variety of student responses. What is important is not that all parties agree with our point scheme, but rather that independent graders will achieve consistent results using the rubric (see Sec. IV).

The final rubric includes a point allocation for each section of the problem, descriptions of common student mistakes with accompanying point allocations, and codes that are used to categorize student mistakes on that question (to assist with research). The rubric underwent significant revision over time as independent graders compared results on student exams. The final rubric is a concise document and graders undergo significant training to use it consistently. See Fig. 2 for the rubric for sample problems and the Appendix for the full rubric. This inter-rater grading informed subsequent revisions of the rubric (typically involving explicit point allocations for particular types of student responses), resulting in improved inter-rater reliability over time.

To check inter-rater reliability on the final rubric, a set of CUE exams was scored by independent graders and the scores compared. Prior to grading the set of common exams, graders were trained on the rubric. Training occurred by providing graders with a set of earlier CUE exams with scores agreed upon by two of the authors (S. V. C. and S. J. P.) and comparing grader scores to the agreed-upon scores.

Inter-rater reliability data were collected on a total of 47 exams, over the course of two different studies described below. All final data in Sec. V use CUE version 16 and the final version, which are nearly identical.

The first grading study used two graders (A and B) on a set of 37 exams (CUE version 16). Only minor adjustments were made to the wording on all questions (except for question 14, see below) after version 16, so these inter-rater reliability data are included in the final data used to validate the rubric. The use of the inter-rater reliability data from this study on CUE version 16 will tend to *overestimate* the difference between graders—modifications made to the rubric after this time only added additional precision.

The second study used two different graders (C and D) on an additional 15 exams from the final version of the CUE. Thus, we collected inter-rater reliability data on a total of 52 exams for each question.

Question 14 was substantially modified after version 16, and so two graders (C and D) graded only question 14 on an additional 32 exams from the final version of the CUE. Combined with the 15 exams described in the paragraph above, we collected inter-rater reliability data on a total of 47 exams for question 14. These data replaced those for question 14 on older student exams, resulting in complete exam data on 47 exams.

While this detailed rubric serves the purposes of the exam—to provide consistent measurements of student performance on the CUE questions—it has at least two limitations. First, the final detailed rubric will not be able to account for all possible student answers and thought processes (though repeated use has shown that it accounts for the vast majority of student answers).

Second, because the number of points allocated to each question is subjective, the CUE offers only an ordinal, rather than interval, measure of student performance. That is, we know that a student score of 60% is meaningfully higher than a student score of 30% on the instrument, but we cannot claim that a score of 60% indicates that that student is *twice* as successful on the instrument. The rubric does not allow us to make such arguments, as indicated by the results in Sec. IV.

III. RELIABILITY OF THE CUE

One desirable characteristic of a conceptual assessment is that the instrument be reliable, that it measure the construct of interest consistently. In our case, we would like to determine that a student who does well on one item on the test tends to do well on other items on the test (internal consistency), and that different graders will assign similar CUE scores to the same student (inter-rater reliability). In this section, we indicate that the CUE adequately meets these two measures of reliability. For more information on reliability and validity test statistics, we refer the interested reader to related publications [6,25,27].

A. Internal consistency

In order to test internal consistency, or how well different items on the CUE give consistent results with one another, we calculated Cronbach's alpha for the test as a whole. Cronbach's alpha is a correlational measure and can be interpreted as the average of the correlations of all possible split-half exams[28]. Values range from 0 to 1, with larger numbers indicating a greater correlation between test items, and thus the degree to which test items measure the same, or related, constructs. A common cutoff value for a good value of alpha is 0.80 [29].

We take each question as one test item. Calculating Cronbach's alpha using the actual score on each question (e.g., 5 points, 7 points), we obtain $\alpha = 0.82$. This number indicates high internal statistical reliability. We recognize, however, that the CUE does not measure a single unidimensional construct, thus violating the assumptions of Cronbach's alpha. However, this means that our computed alpha is likely an underestimate of the true alpha, suggesting that internal consistency is still high [30]. These values are highly sample dependent [27], as Cronbach's alpha relies heavily on the total score variance.

B. Inter-rater reliability

Inter-rater reliability for the CUE exam as a whole is high. The absolute value of grader differences on the "total CUE score" are shown in Fig. 3. Averaged across all exams, the difference in the total CUE score between graders was 1% (out of 100%) with the standard deviation (of the difference) of 3.6%. Graders agreed on overall CUE scores within 10% for all students, and within 5% for the majority of students (74%).

As an additional measure of inter-rater reliability, we computed Cohen's kappa [31] for the CUE. Cohen's kappa calculates how often raters give an exam, or a question, the same score, compared to the proportion expected by chance. Using Cohen's kappa to calculate agreement on the actual points earned on the CUE is not particularly useful: Given the large spectrum of points (118 possible), it is unlikely that graders will agree within a single point. Thus, student scores were binned using five-point increments [32] and Cohen's kappa computed for these binned scores to be 0.41, which is defined as "moderate" [22] agreement. When we bin scores within 10% agreement, Cohen's kappa is calculated to be 0.62, indicating "substantial" agreement. We note that this agreement

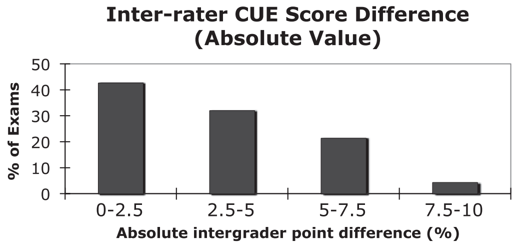


FIG. 3. Absolute value grader differences in "total CUE score" for a total of 47 exams graded in common by independent graders using a common rubric. Categories are exclusive of the lowest number in the bin except for the lowest bin: E.g., 0–2.5 can be read as $0 \leq X \leq 2.5$ and 2.5 to 5 can be read as $2.5 < X \leq 5$.

closely matches a pooled kappa estimate [33] computed by taking a weighted average (by question point value) of individual item kappa scores (see below).

We note that Cohen's kappa is an overly conservative measure of inter-rater reliability [34]; this fact, combined with the restrictive binning cutoffs required for agreement, has likely resulted in an underestimate of the true Cohen's kappa score.

These Cohen's kappa calculations suggest that raters agree moderately for scores within roughly 5% and substantially for scores within 10%, and further substantiate the results shown in Fig. 3; more than 70% of the overall scores given by two graders agreed within 5%. Thus, for any given exam, different graders will assign a similar score, suggesting that the rubric is good and the CUE score is reliable.

We also examined inter-rater reliability on the individual questions on the exam. All questions are normalized to 100 points for this analysis. One long question (Q12), worth 26 points, is split into three parts, for a total of 19 questions. The average inter-rater agreement is high for individual questions: Grader scores differ by 3.1% (absolute value) on any individual question, on average, and the average standard deviation of that difference is 2.1%. Also, as seen in Fig. 4, graders were in "close" agreement for at least 85% of students on all questions. The two graders were in *exact* agreement for 45% or more of the students on all questions but one. These results are promising, particularly considering the variability of student responses and difficulty in interpreting open-ended answers.

We also computed Cohen's kappa for each question on the CUE. For individual questions on the CUE, Cohen's kappa ranged from 0.41 to 1.0, indicating "moderate" to "excellent" agreement [22] (see Table I). The majority ($N = 11$) of questions showed "substantial" agreement (i.e., $k > 0.6$). Because Cohen's kappa is a conservative measure, we believe that agreement between raters was higher than Cohen's kappa might suggest (e.g., Fig. 4).

Thus inter-rater reliability is high for individual questions and the exam as a whole. We are able to accurately determine a student score on the exam as a whole within 5% (Fig. 3), and on individual questions within 20% (Fig. 4). These results indicate that it is meaningful to compare students according to their performance on individual questions as well as the exam as a whole.

IV. VALIDITY

Another key aspect of a conceptual survey is, Are the results a true measure of the construct(s) that we intend to measure? In order to answer this question, a variety of subquestions must be addressed: Does the instrument give similar results to other approaches of measuring the same construct? Was there a sound theoretical basis underlying the construct(s)? Do experts agree with the operationalization of those constructs into questions? Do item scores correlate with the test as a whole? Do students interpret the questions as they were intended? Do test items discriminate between students of different abilities? In this section, we show that the CUE is, overall, a valid instrument for this population of students in our institutional context.

A. Expert validation

The CUE was constructed to address the learning goals developed by the faculty working group. Thus, the learning goals can be taken to form the underlying construct, or theoretical basis, for the exam. We take these learning goals to be expert validated due to their generation by a working group of experts in both physics content and education.

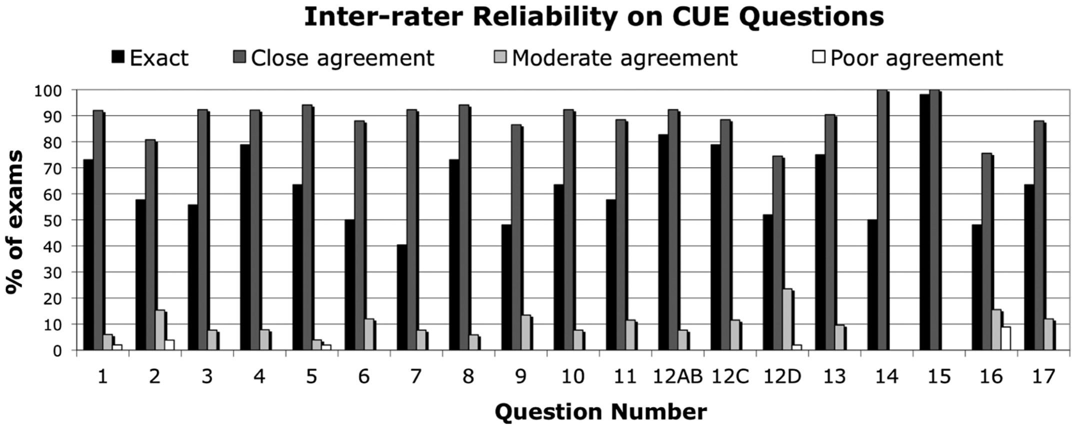


FIG. 4. Inter-rater differences on CUE questions for 52 exams. "Close agreement" is agreement within ±20% (± 1 point on a 5 point question), "moderate" is within ±20%–50% (± 1–2.5 points on a 5 point question), and "poor" is off by 50%–100%. Blanks omitted (thus, exam N varies slightly per question, with a maximum of 52 exams per question).

Do experts agree with the operationalization of these learning goals into CUE questions? The faculty working group ($N \sim 14$) was deeply involved in the generation and refinement of the first draft of the CUE, and PER faculty were involved in the continual refinement of the CUE over the additional years of development. This is one indication that the questions are considered valid by content experts.

Additionally, five faculty (four from external institutions) responded to a survey[35] about the instrument. The survey asked them to provide their best answer to each question and to indicate whether the question tested the learning goals (which were given), whether the information was scientifically accurate, and if the question was clear and precise. Three faculty were from liberal arts colleges and one from a state university; all had taught junior E&M and three had used the CUE in their course. Because of the low number of faculty responses, and the fact that not all faculty completed the same version of the survey, we offer qualitative rather than quantitative analysis.

Faculty were able to accurately answer questions on the CUE, indicating that they are able to correctly interpret the questions. However, we have not explicitly administered and graded the CUE for experts (such as faculty and graduate students), which would be a useful benchmark. Faculty overall rated questions as scientifically accurate and clear. Faculty did not mention a lack of alignment of CUE questions with the consensus, course-scale learning goals. Faculty conjectured some aspects of wording that students might miss, but these passed the student validation tests, Sec. IV B.

Some faculty feedback was used to revise the CUE for the final version. Other feedback could be incorporated for future versions, such as some concerns raised about the explicit directions regarding the point of interest in Q12; the problem asks for the value of the z component of the E field "near the origin," without explicitly asking the student to consider only points along the z axis. However, most responses were indicative of differences in philosophy rather than problems with the CUE: For example, two respondents remarked in questions 1–7 [where students are asked to indicate the easiest method for solving a particular problem for E field (E) or voltage (V)] that they questioned giving students the choice of solving for E or V (or B field or vector potential, A). However, this was a strategic decision since telling students to solve for one or the other would give a clue as to the best method (e.g., separation of variables versus Coulomb's law). Thus, overall, faculty saw the CUE as a valid instrument, but some additional feedback will be discussed in Sec. VI.

B. Student validation

The penultimate version was validated by interviewing eight students using a think-aloud protocol: the interviewer did not interject except to remind students to verbalize their thought processes. Students were asked to read the question aloud, so that we would know if they were skipping parts. Students were observed as to whether their work reflected the nature of the question. At the end of the interview, students were asked about questions that seemed problematic, in order to probe their understanding of the question prompt.

Particular attention was paid to whether students correctly read and interpreted the instructions (e.g., not attempting to solve the problems Q1–Q7, understanding the meaning of "off-axis," or noticing key aspects of the questions).

This final validation resulted in a few, minor wording changes. For example, Q16 was originally worded "Does this distribution have a non-zero dipole moment? Yes/No." In interviews, many students verbally described that they knew that there was no dipole moment, but circled "No" on the exam, due to the unintended double negative. Thus, this question was changed to "The dipole moment of this distribution is: Zero/Non-zero/Not Sure." Interviews also enabled us to fix confusing wording in Q8.

One common theme in the interviews was that students were more complete in verbalizing their explanations than in writing them—which could account for some of the poor quality of explanations on the CUE. Providing motivation for students to document their understanding more completely would provide additional validity for the test.

C. Criterion validity

How well does the CUE score predict other, related variables? We collected data where available to provide external measures of student success. We find that the CUE score correlates well with other measures of student performance, indicating that the assessment does appear to measure aspects of student ability as determined by other criteria. The overall score on the final version of the CUE correlates highly with students' grades in the junior E&M course ($r = 0.59$, $p < 0.001$, $N = 100$) and their cumulative grade point average prior to junior E&M ($r = 0.49$, $p < 0.001$, $N = 99$), as well as with the BEMA [6] score obtained after freshman E&M ($r = 0.57$, $p < 0.001$, $N = 73$). These correlations match previous definitions [36] as "medium" (0.3–0.5) to "strong" (0.5–1.0), suggesting that the constructs measured on the CUE are highly related to other aspects of student performance typically valued by faculty.

D. Discrimination

Below we offer several measures of whether test items discriminate in a meaningful way between students of different abilities.

1. Item-test correlation

In order to determine whether test items are well-coordinated with the test as a whole, we examined the

Pearson linear correlation coefficient for each test item with the overall test score. Pearson correlation ranges from −1 to 1, with higher numbers indicating that the two variables share a large amount of variance. There is no widely accepted cutoff for item-test correlation coefficients for continuous variables, such as those on the CUE, so we refer to the criterion accepted for item-test correlation for dichotomous variables (the point-biserial coefficient), which is $r \geq 0.2$ [37]. On the CUE, all items were correlated with the overall score with at least $r = 0.5$: Nine items were correlated with at least $r = 0.6$. This item-test correlation is similar to another well-accepted instrument for E&M (at the introductory level): the BEMA, which has an average point-biserial coefficient of 0.43 [6]. Thus, no single item stands out as being poorly discriminatory, and all items are reasonably correlated with the overall test score.

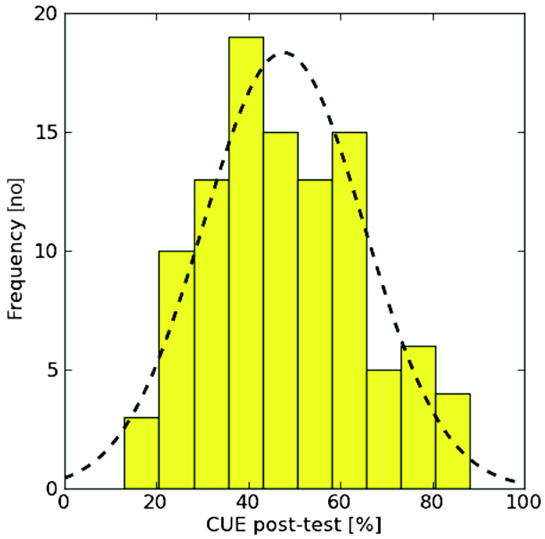


FIG. 5 (color online). Histogram of student scores on the CUE based on $N = 103$ students in three courses.

2. Ferguson's delta

In order to determine the discriminatory power of the test as a whole (i.e., how well distributed are CUE scores over the possible range of scores?), we calculated Ferguson's delta, as described by Ding *et al.* [6]. To take into account the fact that test items are worth different numbers of points, we take the total number of test items (K) as 118 (the number of points on the test), and calculate the frequency (f_i) of the number of points earned (not the normalized CUE score out of 100). Ferguson's delta can take on values in the range [0, 1], with a value greater than 0.9 indicating good discrimination. Ferguson's delta for the CUE is 0.99, indicating excellent discrimination on a per-point basis. The discriminatory power of the CUE can also be seen visually from the histogram of student responses in Fig. 5: Scores are distributed normally around the mean. Normality was verified using the Shapiro-Wilk test, $p < 0.05$.

3. Item difficulty

Different questions on the CUE pose different levels of challenge for students, as can be seen by the variable mean and median of each question (Fig. 6). The mean and median often do not match due to the non-normality of the distribution of student responses. We are unable to use the item difficulty index (P, as described by Ding *et al.* [6]) due to the fact that our question scoring is continuous rather than dichotomous. What counts as a "correct" score

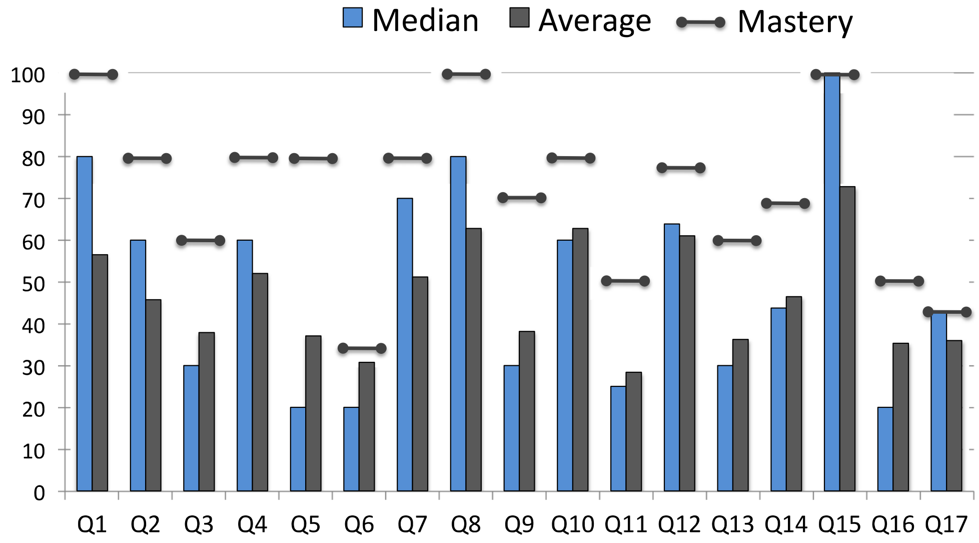


FIG. 6 (color online). Median, mean, and mastery performances on each question on the final version of the CUE for $N = 103$ students. "Mastery" level represents the upper 70th percentile on each question; that is, 30% of students scored at or above this level, and 70% scored below.

on any item on the CUE, especially given that the skills and ideas tested by this assessment are complex? We choose to define the 70th percentile as a "mastery" level, corresponding to the assignment of a grade of 3.0 or higher (A or B) on a standard grading curve, though one could certainly choose other methods of defining mastery. Figure 6 shows the results of this calculation, and Table I, as well as the Appendix, gives descriptions of each question. For three questions (Q1, Q8, and Q15), the 70th percentile score is 100%, indicating that they may have less value in differentiating student performance. On the other end of the spectrum, certain questions (e.g., Q3, Q6, Q11, Q16, Q17) are particularly difficult for the majority of students, but we find that our rubric makes it difficult for students to achieve high scores on most problems and results in many students receiving zero points for some problems. Thus, when considering student performance on any particular problem, it is important to keep these different metrics of success in mind: For example, a student score of 60% on Q3 can be taken as good performance, relative to the student population as a whole. These results are only valid for the population at hand, however—junior physics majors at CU-Boulder. Caution should be used when extrapolating to other student populations.

V. RESULTS

What are typical student results on the CUE? Because students have taken the CUE in its many different versions, we answer this question in terms of both the final, validated assessment and a "comparison" CUE score made of the questions that remained constant over several different administered versions of the exam (versions 9, 13, 16, 18, and 22). We remind the reader that we may make only ordinal, not interval, claims regarding student performance on the CUE.

A. Post-test

A total of 103 students in three courses at CU took the final form of the CUE. These courses (A, B, C) are described in Sec. II. Because of the variety of courses involved, these results can be taken as indicative of the results of a variety of types of instruction, but tending towards traditional lecture format. The timing of the test also varies—in particular, course C administered the CUE at the beginning of the following semester, introducing a variety of confounding variables (self-selection effects, studying for the final exam, and a period of forgetting).

Students, on average, scored 47.8 ± 1.9% on the post-test. Figure 5 above shows the spread of student performance ranging from a low of 13 to a high of 88, following a rough Gaussian curve. Student performance on each question is shown in Fig. 6 (overall) and the Appendix (by class).

CUE results differed across courses, such that students in course A (using the most interactive techniques) scored higher on average (52.9 ± 3.0%) than did students in either course B (45.4 ± 4.3%) or course C (45.1 ± 2.3%). This suggests that CUE scores might be able to differentiate between different types of instruction—a point that will be supported later in this section.

CUE averages on individual questions vary widely. Question 2, for example, varies from a low of 25 (in course B) to a high of 56 (in course C). See the Appendix for average student response for individual questions in each of the three courses, which may be used as points of comparison for future CUE administrations. On several questions a large fraction of students scored 0 points, resulting in a large standard deviation for many questions. Comparing student performance on individual questions across courses can guide faculty as to which pedagogical approaches are most effectively addressing particular topical areas. By using the CUE repeatedly in their own class, faculty can use the individual question results to help them assess whether a change in their own approach to a topic improved student outcomes on that topic.

B. Pretest

Seven of the CUE questions—those which a student with a strong grasp of introductory E&M could be expected to do – are bundled into an optional pretest. Student pretest scores are consistently low. On the final version of the pretest (course A, $N = 51$), students score an average of 29.4 ± 2.4%. This is similar to results from the $N = 156$ students who took previous versions of the pretest, scoring 33 ± 1.2% on average. These junior students score very similarly on this pretest to a set of $N = 26$ freshmen who have just completed introductory E&M: 30.1 ± 2.9%. Overall, student responses on the pretest are low quality, reflecting a high level of forgetting and/or confusion. However, pretest scores [38] correlate well with a variety of measures of preparation, such as the BEMA after introductory physics ($r = 0.29$, $p < 0.05$, $N = 73$) and GPA in prior physics courses ($r = 0.35$, $p < 0.001$, $N = 138$) indicating that the pretest score appears to be, nonetheless, a meaningful measure of student preparation. Pretest scores are also well correlated with post-test scores ($r = 0.48$, $N = 138$).

C. Gain

There are several ways that we can choose to calculate CUE gain: (A) *post-pre*, the full post-test score minus the pretest score; (B) *normalized post-pre*, the normalized gain of the full post-test score [i.e., post-pre/(100-pre)]; (C) *7Q post-pre*, the post-test score on *only* the seven questions which match the questions on the pretest, minus the pretest score; or (D) *normalized 7Q post-pre*, the normalized gain of the seven questions on the post-test which match the questions on the pretest, minus the pretest. We choose to examine (C) and (D), so that we can compare student performance on the same set of questions pre and post.

Because we can make only ordinal, not interval, claims for the pretest and post-test scores, the subtraction of pre and post is problematic regardless of whether we use normalized or non-normalized gain. See Wallace and Bailey for a detailed discussion [27]. If we ignore the problematic nature of subtracting two ordinal numbers, and assume that the magnitude of the gain itself is meaningful, then post-pre could give a crude measurement of how much students improve on those questions over time. This is of particular interest when administering the CUE to different populations with differing background preparation. For example, the pretest scores of other institutions were 41.0% and 17.3%, respectively (see Sec. V D).

Average non-normalized gain is 24% (34% normalized) for the 20 students in a single course (total $N = 43$ students) who took the final version of the test and also took both the pre- and post-tests, and 24% (37% normalized) for the population who took V18 and V21 versions.

D. Common score

To provide a larger sample size, we include all courses that administered the CUE at a similar point in instruction (i.e., all courses but course C, which administered the CUE at the start of the following semester), including several traditionally taught courses at CU and elsewhere that did not use our materials. Because many courses used previous versions of the CUE, we include only those questions that did not change substantially over time on an artificial "common" test score. Out of the total 118 points on the final version of the test, these common questions represent 88 points, or about 3/4 of the exam. The common CUE score correlates well ($r = 0.97$, $p < 0.001$) with the full exam score for the $N = 103$ students who took the final version of the CUE, suggesting that this portion of the exam is a valid proxy for the exam score as a whole.

Results for all $N = 488$ students are given in Fig. 7. Courses using the materials developed by our PER efforts (PER courses) have higher CUE scores on average (60.6 ± 4.3%, $N = 189$) compared to the more standard lecture-based courses (STND; 40.8 ± 3.9%, $N = 299$).

E. Broader conclusions

We can draw several conclusions from these data. One is that the CUE is able to robustly and repeatedly differentiate between different types of instruction. This is a finding that is suggested by the overall difference in scores between research-based and lecture-based courses across institutions in Fig. 7, and confirmed by the scores in those courses

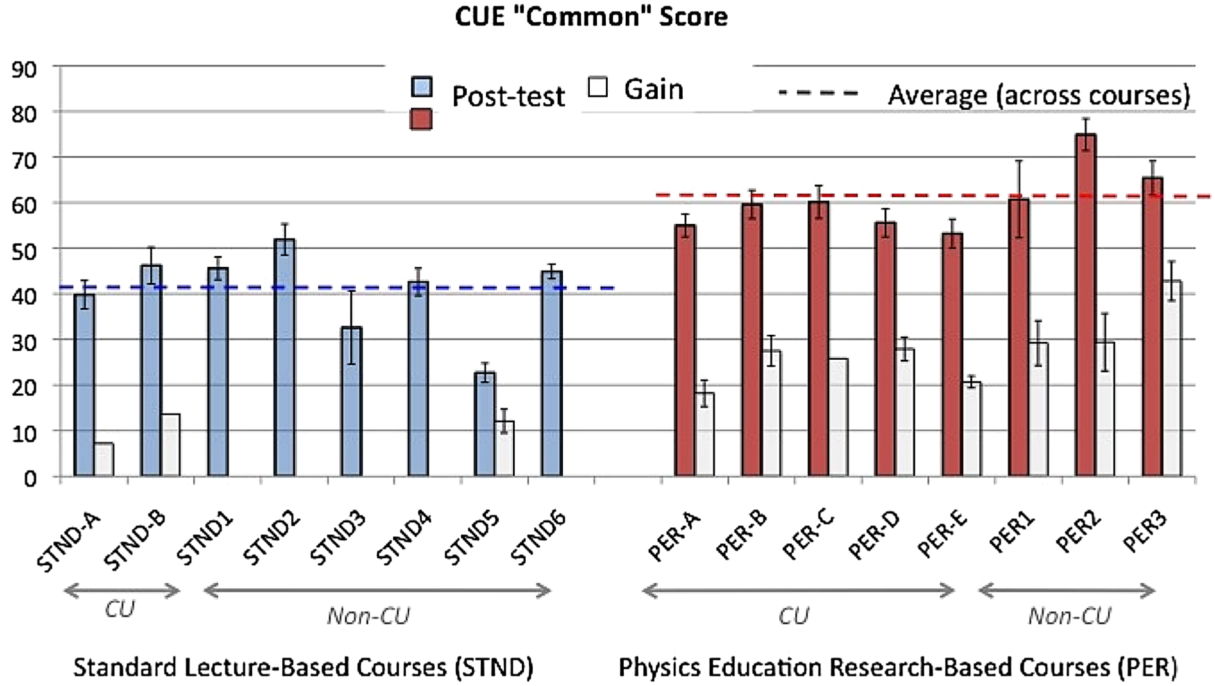


FIG. 7 (color online). "Common" CUE scores across institutions for $N = 488$ students. "Post-test" represents course average score (% correct) for the subset of CUE questions given in common across all exams (88 out of 118 possible points). "Gain" represents the course average (out of 100%) for the difference between the pretest (60 points) and the matched subset of the post-test (i.e., 60 points). PER courses used the research-based materials developed at CU. STND courses used a standard lecture-based format. Because of the lack of pretests for CU-RES1 and CU STND1 and STND2, pretest scores are estimated (and thus gain scores are effective and lack error bars) based on the stable pretest scores for other semesters. Courses are not listed chronologically. Courses A and B (from final CUE administration) are labeled. Error bars represent 1 standard error of the mean.

at CU, which included students with similar background preparation (see Sec. II).

Another finding is that CUE scores are able to differentiate between the caliber of students at different institutions: The CUE post-test scores of PER courses at non-CU institutions are slightly higher, particularly non-CU2—confirmed by a Kruskal-Wallis test and a *post hoc* pairwise comparison. All non-CU PER-based courses were at small liberal arts institutions with competitive entrance requirements, and the pretest scores of these students tended to be somewhat higher than those at CU (35%–40%, with the exception of non-CU-PER3 at 17%), and post-test scores are substantially higher than at CU.

For instructors and researchers interested in the exact student scores on the ultimate version of the CUE exam (used by courses A, B, and C), we refer them to Sec. V A (and Fig. 6) which lists student scores on the final exam as a whole, as well as the Appendix, which includes student scores on individual exam items.

VI. DISCUSSION AND CONCLUSIONS

In summary, we have developed an open-ended test that probes a subset of the expert-developed learning goals in electrostatics and magnetostatics in junior E&MI. We developed a detailed grading rubric to provide comparison of scores from student to student, and inter-rater reliability studies show that the responses to the CUE are graded with a high degree of reliability. The test shows high internal consistency, measured by Cronbach's alpha. Students and experts correctly interpret the questions on the exam, and expert validation results were overall positive. The CUE is well correlated with other variables of interest (such as grades and GPA), indicating that it measures things that faculty care about. The CUE exam scores can be used to find measurable differences between students of different abilities (as evidenced by item-test correlations, Ferguson's delta, and item difficulty). The CUE exam scores can also be used to find measurable differences between different populations of students—those with different backgrounds (i.e., small liberal arts institutions versus large state universities), or who took different kinds of E&M courses (i.e., standard lecture-based courses or research-based interactive-engagement courses).

Thus, the CUE has achieved the goals that we set for it—to be able to provide reliable and valid information about the achievement of students in junior E&M instruction.

These positive results are somewhat remarkable considering that the CUE is an *open-ended* test, which is based on knowledge of student difficulties that are still emerging and poorly defined. In fact, the CUE does not even match some definitions of "concept inventory" [26] because it is not multiple choice. Thus, we have also demonstrated more broadly that validity and reliability can be achieved—with some effort—with a non-multiple-choice inventory. Typically, one would take common student answers on an open-ended exam as we have created and generate multiple-choice distractors from common student responses. This is certainly still possible and may be a fruitful direction of future test development. However, we have hesitated to change the nature of the exam this dramatically—many of the learning goals of this course require that students *generate* particular representations or arguments rather than *recognize* them. We expect this higher level of cognitive function [39] from our junior students and want to be able to assess it.

We have already found that the CUE is a valuable tool in identifying common student difficulties [11], and future publications are planned on this rich data source. For example, Q4 (identifying the method of images) was answered correctly by 83% of students, which could be seen as confirming our expert validators' opinions that this problem was "too easy," or "a touchstone problem." However, the overall score on the problem was only 64%, showing that *explanations* for the correct answer are often challenging for students—a common theme on the exam. Similarly, faculty reviewers were concerned that Q6 (the magnetic field of a current loop off axis, far away) did not test students' ability to determine when to use Biot-Savart since Biot-Savart is "complicated" for this problem and this is the "most basic example of magnetic dipole." We concur; however, many students mistakenly indicate that they would use Biot-Savart as the *easiest* method to solve this problem (41% out of $N = 103$ students taking the final version of the CUE post-test), not recognizing this basic dipole.

These findings indicate how the CUE may be useful to illuminate student thinking. To provide a sense of insights available through CUE data, we summarize a few other common themes here.

- In both lecture-based and research-based courses, there is a marked difference between students' ability to choose the right answer and their ability to justify or explain that answer—students are more proficient at choosing the right answer than justifying it. Such analysis of CUE data is possible due to separate coding of "correctness" and "explanation quality."
- Choosing the easiest method for solving a given problem (Q1–Q7) proves difficult for students, possibly highlighting difficulties in extracting the central features of problems out of the context of recent material.
- Students often fail to recognize when approximations are useful or, indeed, provide the only tractable solution for a problem. For example, many students do not recognize that an observation point that is far away and off axis calls for the use of the dipole approximation (Q3 and Q6). Instead, they often fall back on the "fail-safe" of direct integration.

See previous publications [11,12] for more examples of common trends on the exam. The CUE provides a rich source of fruitful data for ongoing and future research of this nature.

Because of the large sample dependency of many of these validity and reliability measures [27], our conclusions are only strictly valid for the student population with which it was tested—primarily junior E&M students at the University of Colorado. Population dependence is one of the main drawbacks of classical test theory, whereas a more detailed study using item response theory could enable us to disentangle student ability from the quality of the test items. This analysis was beyond the scope of the current study.

We can also make only ordinal, not interval, claims regarding student scores on the test. While it is clear that the CUE differentiates between students of different abilities, it is not clear just what the magnitude of CUE scores represents.

The CUE is also necessarily restricted in scope. Many learning goals are not addressed, which was a conscious choice. For example, as pointed out by expert validation, the CUE does not test students' ability to connect Maxwell's equations to boundary conditions, is more focused on abstract principles than phenomena, does not explicitly ask students to address the difference between exact and approximate solutions, and has only a few questions on magnetostatics or electric fields in matter (i.e., dielectrics). Indeed, the CUE is heavily focused on problem-solving methods, strategies, and skills in electrostatics—while these match the learning goals described by our faculty, an expanded focus could be considered. Additional formal construct validation, as well as gathering expert scores on the exam, would be appropriate next steps in the development of the CUE.

For instructors or researchers whose learning goals for their students match those learning goals developed by our faculty, the CUE can be a tool for testing broad achievement of learning goals (by comparing scores on the CUE exam overall, as graded by trained graders). Such administration can also allow for insight into student difficulties by reviewing student responses, comparing student performance on individual items to the mastery level in Fig. 6, comparing student scores to those in the Appendix, or looking at common errors using the documented ‘‘mistake codes’’ assigned during grading.

We hope that the CUE will serve as a tool to highlight such areas where there is a mismatch between what faculty believe students will be able to do and what they are actually able to do, as well as to provide a measure of the success of various instructional strategies in junior E&M.

ACKNOWLEDGMENTS

We thank the members of our faculty working group who contributed to the course learning goals and the development of the CUE, including S. DeAlwis, P. Beale, M. Betterton, T. DeGrand, O. DeWolfe, M. Dubson, N. Finkelstein, W. Ford, A. Hasenfratz, T. Munsat, S. Parker, K. Perkins, S. Pollock, C. Rogers, and C. Wieman. We also gratefully acknowledge the faculty at CU-Boulder and elsewhere who administered the CUE in their course and provided us with feedback on the instrument. Many thanks to Bethany Wilcox and Amy Bloom, who served as graders on many versions of the CUE and assisted with inter-rater reliability studies, as well as Colin Wallace, who provided feedback and recommendations on statistical analyses in this paper. This work is funded by the SEI, CU-Boulder, and NSF-CCLI Grant No. 0737118. R. E. P. is currently supported by the Miller Institute for Basic Research in Science at the University of California Berkeley.

SUPPLEMENTAL MATERIAL

See separate supplemental material for student performance on CUE questions, the CUE assessment as a whole, and the CUE grading rubric.

[1] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999), and references therein; R. Chabay and B. Sherwood, Restructuring the introductory electricity and magnetism course, Am. J. Phys. **74**, 329 (2006).

[2] C. Singh, Student understanding of symmetry and Gauss's law of electricity, Am. J. Phys. **74**, 923 (2006).

[3] A wide variety of conceptual instruments can be seen at the NCSU Assessment Instrument Information Page, North Carolina State University, http://www.ncsu.edu/per/TestInfo.html accessed August 30, 2011.

[4] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992); The (updated) FCI instrument is available online at http://modeling.asu.edu/R&E/Research.html; R. R. Hake, Interactive engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[5] D. Maloney, T. O'Kuma, C. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[6] 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).

[7] L. C. McDermott, P. S. Shaffer, and the Physics Education Group at the University of Washington, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).

[8] M. Dancy and C. Henderson, Pedagogical practices and instructional change of physics faculty, Am. J. Phys. **78**, 1056 (2010).

[9] *How People Learn*, edited by J. Bransford, A. Brown, and R. Cocking (National Academy Press, Washington, DC, 2000).

[10] C. A. Manogue, K. Brown, T. Dray, and B. Edwards, Why is Ampère's law so hard? A look at middle division physics, Am. J. Phys. **74**, 344 (2006); T. J. Bing and E. F. Redish, Analyzing problem solving using math in physics: Epistemological framing via warrants, Phys. Rev. ST Phys. Educ. Res. **5**, 020108 (2009).

[11] R. E. Pepper, S. V. Chasteen, S. J. Pollock, and K. K. Perkins, Our best juniors still struggle with Gauss's law: Characterizing their difficulties, AIP Conf. Proc. **1289**, 245 (2010); C. Wallace and S. V. Chasteen, Upper-division students' difficulties with Ampère's law, Phys. Rev. ST Phys. Educ. Res. **6**, 1 (2010); R. Pepper, S. V. Chasteen, S. J. Pollock, and K. K. Perkins, Observations on student difficulties with mathematics in upper-division electricity and magnetism, Phys. Rev. ST Phys. Educ. Res. **8**, 010111 (2012).

[12] B. S. Ambrose, Investigating student understanding in intermediate mechanics: Identifying the need for a tutorial approach to instruction, Am. J. Phys. **72**, 453 (2004), and the first six references therein; A. Mason and C. Singh, Do advanced physics students learn from their mistakes without explicit intervention?, Am. J. Phys. **78**, 760 (2010); C. M. Sorensen, D. L. McBride, and N. S. Rebello, Studio optics: Adapting interactive engagement pedagogy, Am. J. Phys. **79**, 320 (2011); C. A. Manogue, P. J. Siemens, J. Tate, K. Browne, M. L. Niess, and A. J. Wolfer, Paradigms in physics: A new upper-division curriculum, Am. J. Phys. **69**, 978 (2001); C. A. Manogue, L. Cerny, E. Gire, D. B. Mountcastle, E. Price, and E. H. van Zee, Upper-division activities that foster "thinking like a physicist," AIP Conf. Proc. **1289**, 37 (2010); S. V. Chasteen and S. J. Pollock, Tapping into juniors' understanding of E&M: The Colorado Upper-Division Electrostatics (CUE) Diagnostic, AIP Conf. Proc. **1179**, 109 (2009); K. K. Perkins and C. Turpen, Student perspectives on using clickers in upper-division courses, AIP Conf. Proc. **1179**, 225 (2009); S. J. Pollock, Longitudinal study of student conceptual understanding in electricity and magnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020110 (2009); S. J. Pollock, S. V. Chasteen, M. Dubson, and K. K. Perkins, The use of concept tests and peer instruction in upper-division physics, AIP Conf. Proc. **1289**, 261 (2010); S. V. Chasteen, R. E. Pepper, S. J. Pollock, and K. K. Perkins, But does it last? Sustaining a research-based curriculum in upper-division electricity & magnetism, AIP Conf. Proc. **1413**, 139 (2012).

[13] S. V. Chasteen, S. J. Pollock, R. E. Pepper, and K. K. Perkins, "Thinking like a physicist": A multi-semester case study of junior E&M, Am. J. Phys. (to be published); Transforming the junior level: Outcomes from instruction and research in E&M, Phys. Rev. ST Phys. Educ. Res. **8**, 020107 (2012).

[14] G. Zhu and C. Singh, Surveying students' understanding of quantum mechanics in one spatial dimension, Am. J. Phys. **80**, 252 (2012); S. Goldhaber, S. Pollock, M. Dubson, P. Beale, and K. Perkins, Transforming upper-division quantum mechanics: Learning goals and assessment, AIP Conf. Proc. **1179**, 145 (2009).

[15] B. M. Notaros, Concept inventory assessment instruments for electromagnetics education, in *Proceedings of the Antennas and Propagation Society International Symposium, San Antonio, TX, 2002* (IEEE, Piscataway, NJ, 2002), Vol. 1, p. 684.

[16] S. V. Chasteen, K. K. Perkins, P. Beale, S. J. Pollock, and C. E. Wieman, A thoughtful approach to instruction: Course transformations for the rest of us, J. Coll. Sci. Teach. **40**, 70 (2011) [http://www.nsta.org/publications/browse_journals.aspx?action=issue&id=10.2505/3/jcst11_040_04].

[17] This effort was supported by funding from the CU Science Education Initiative.

[18] D. J. Griffiths, *Introduction to Electrodynamics* (Prentice-Hall, Upper Saddle River, NJ, 1999), 3rd ed.

[19] R. E. Pepper, S. V. Chasteen, S. J. Pollock, and K. K. Perkins, Facilitating faculty conversations: Development of consensus learning goals, AIP Conf. Proc. **1413**, 291 (2012).

[20] http://www.colorado.edu/sei/departments/physics_learning.htm.

[21] S. J. Pollock, R. E. Pepper, and A. D. Marino, Issues and progress in transforming a middle-division classical mechanics/math methods course, AIP Conf. Proc. **1413**, 303 (2012).

[22] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33**, 159 (1977).

[23] The questions on the existing EMCI (Ref. [15]) were not used because we were not aware of the exam at the time; this exam remains unvalidated to our knowledge.

[24] The finalized version of the CUE used in this paper is version 22.

[25] 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).

[26] R. S. Lindell, E. Peak, and T. M. Foster, Are they all created equal? A comparison of different concept inventory development methodologies, AIP Conf. Proc. **883**, 14 (2007).

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

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

[29] D. George and P. Mallery, *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 update* (Allyn & Bacon, Boston, 2003), 4th ed.

[30] J. M. Graham, Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them, Educ. Psychol. Meas. **66**, 930 (2006).

[31] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,

Psychol. Bull. **70**, 213 (1968). Cohen's kappa is computed as $\kappa = (q_o - q_e)/(1 - q_e)$, where q_o is the actual observed agreement between raters and q_e is the agreement expected by chance. The chance agreement q_e is calculated from the observed data, using the number of times each rater assigns a given score category, by multiplying the proportion of cases one rater assigns a score i by the proportion of cases the other rater assigns that same score i, summed over all possible score categories.

[32] Bins were assigned as 0–5, 6–10, 11–15, etc. for 5-point bins and 0–9, 10–19, 20–29, etc. for 10-point bins. Grader agreement within these bins was defined as ''agreement'' for purposes of calculating the actual agreement. Number of exams scored within each of these bins was used to calculate the agreement expected by chance, as defined above. For purposes of computing Cohen's kappa on individual items, agreement was defined as giving the same point value; that is, rater scores of 3.0 and 3.5 were counted as agreement on the same point value of ''3,'' whereas rater scores of 3.5 and 4.0 would not be counted as agreement.

[33] H. De Vries, M. N. Elliott, D. E. Kanouse, and S. S. Teleki, Using pooled kappa to summarize interrater agreement across many items, Field Methods **20**, 272 (2008).

[34] R. L. Brennan and D. J. Prediger, Coefficient kappa: Some uses, misuses, and alternatives, Educ. Psychol. Meas. **41**, 687 (1981).

[35] http://www.colorado.edu/sei/surveys/Faculty/CUE/Sp09_CUE.html.

[36] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge, New York, 1988), 2nd ed.

[37] See Ref. [16] in L. Ding *et al.* (Ref. [6]).

[38] To increase sample size, we included pretest scores from courses using the prior version of the CUE (V18), as most questions only changed in minor wording.

[39] Bloom's Taxonomy, http://projects.coe.uga.edu/epltt/index.php?title=Bloom%27s_Taxonomy, accessed August 30, 2011.