Validation and analysis of the coupled multiple response Colorado upper-division electrostatics diagnostic

Bethany R. Wilcox and Steven J. Pollock

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

(Received 29 June 2015; published 2 November 2015)

Standardized conceptual assessment represents a widely used tool for educational researchers interested in student learning within the standard undergraduate physics curriculum. For example, these assessments are often used to measure student learning across educational contexts and instructional strategies. However, to support the large-scale implementation often required for cross-institutional testing, it is necessary for these instruments to have question formats that facilitate easy grading. Previously, we created a multiple-response version of an existing, validated, upper-division electrostatics diagnostic with the goal of increasing the instrument's potential for large-scale implementation. Here, we report on the validity and reliability of this new version as an independent instrument. These findings establish the validity of the multiple-response version as measured by multiple test statistics including item difficulty, item discrimination, and internal consistency. Moreover, we demonstrate that the majority of student responses to the new version are internally consistent even when they are incorrect and provide an example of how the new format can be used to gain insight into student difficulties with specific content in electrostatics.

DOI: 10.1103/PhysRevSTPER.11.020130 PACS numbers: 01.40.Fk

I. INTRODUCTION AND BACKGROUND

One natural focus of the physics education research community is on understanding and improving student learning in our physics courses. Often, a critical component of this research is achieving valid measures of student learning, both before and after instruction. Moreover, it is often important that these measures be standardized so that they can assess student learning across populations, learning environments, and instructional strategies. Research-based conceptual assessments, such as the Force Concept Inventory (FCI) [1] and Brief Electricity and Magnetism Assessment (BEMA) [2], are often used for this purpose. After careful validation, these instruments provide a standardized measure of student understanding of specific physics content. Previously, student performance on these assessments has been used to help motivate educational transformation efforts aimed at supporting increased student learning [3], as well as by individual physics instructors to provide formative feedback on the effectiveness of their own instructional practices.

A wide variety of conceptual assessments that target introductory physics content have been developed (see Ref. [4] for a list), and recently a smaller number of upper-division assessments have also been created [5]. Upper-division conceptual assessments are rarer in part because the more advanced physics content of these courses presents unique challenges, including the necessary use of specialized language and sophisticated mathematics. One example of an existing upper-division assessment is the Colorado upper-division electrostatics (CUE) diagnostic [6]. The CUE was designed to target junior-level electrostatics content (i.e., Chaps. 1–6 of Griffiths [7]). Unlike its introductory counterparts, the questions on the CUE primarily have a free response, rather than multiple-choice, format in order to more effectively target students' ability to synthesize and generate responses.

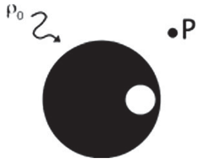
To date, the CUE has been used productively to assess student learning for a number of semesters at multiple institutions [5,6]. The CUE and its associated scoring rubric have been shown to be both valid and reliable for use with this population of upper-division physics students. The assessment has also demonstrated a sensitivity to different types of instruction (e.g., interactive engagement versus lecture only). Recently, in response to the challenges inherent in scoring this type of free-response instrument on a large scale, we crafted a new version of the assessment known as the Coupled Multiple-Response (CMR) CUE. The CMR format utilizes a tiered multiple-response format in which students can select multiple response options and receive credit based on both the accuracy and consistency of their responses. An example of this multiple-response format is given in Fig. 1. The development of the CMR CUE is described in detail in a companion publication [8]. Briefly, the CMR version was created using student responses to the free-response version to construct distractors. As part of its development, the new version was reviewed by nine physics faculty and postdoctoral researchers to ensure that the physics content was clearly and

Q5 - A charged, insulating solid sphere of radius R with a uniform volume charge density ρ_o, with an off-center spherical cavity of radius r carved out of it (see Figure).

Find $\vec{E}$ or V at point P, a distance $4R$ from the sphere.

Select only one: **The easiest method would be ...**

A. Direct Integration
B. Gauss's Law
C. Separation of Variables
D. Multipole Expansion
E. Ampere's Law
F. Method of Images
G. Superposition
H. None of these



because ...(select **ALL** that support your choice)

a. □ you can calculate $\vec{E}$ or V using the integral form of Coulomb's Law
b. □ the sphere will look like a dipole; approximate with $\vec{E}$ or V for an ideal dipole
c. □ $\vec{E}$ or V outside a uniform sphere is the same as from a point charge at the center
d. □ the location of the cavity doesn't matter, you just need $Q_{enclosed}$ to calculate $\vec{E}$
e. □ you can treat this as two uniform spheres, one with charge density ρ_o and one with charge density $-\rho_o$
f. □ this will be the same as a uniform sphere with total charge $\frac{4}{3}\pi(R^3 - r^3)\rho_o$
g. □ electric fields from multiple sources can be combined through a vector sum
h. □ $\nabla^2 V = 0$ outside the cube and you can solve for V using Fourier Series

FIG. 1. A sample item from the CMR CUE. The prompt has been paraphrased; see Ref. [8] for the full prompt and a detailed discussion of question development and scoring.

correctly expressed. The instrument was also administered to 13 students in a think-aloud interview setting to ensure that students were interpreting the questions and distractors as intended.

The CMR CUE is scored with an electronic scoring spreadsheet, thus preserving the fast, objective grading of a standard multiple-choice instrument. This spreadsheet utilizes a nuanced grading scheme designed to match the scoring of the original free-response version [8]. In this scheme, students responding to items like that in Fig. 1 are awarded points for selecting the easiest method and the correct reasoning elements. They can also receive points for selecting methods that are possible but not easy, or for selecting reasoning elements that are consistent with their choice of method. Students can also lose reasoning points if they select reasoning elements that are inconsistent with their choice of method. We also explored simpler grading schemes that, for example, did not offer partial credit or subtract points for inconsistent answers; however, consistency between the free-response and CMR versions was greatest when using the more nuanced rubric [8].

After initial development of the CMR version of the CUE, we performed a direct comparison of student performance on the CMR and free-response versions with students at University of Colorado Boulder (CU) [8]. No statistically significant differences were observed in students' average score on the two versions. Moreover, student performance on individual questions was largely consistent between the two versions, and a qualitative analysis of student responses to the free-response version showed a high degree of consistency between the nature of student responses in the two different formats. Overall, this study found that, for this population of students, the CMR CUE represented an easily graded assessment that produced scores consistent with that of a free-response instrument [8].

The goal of this paper is to report on the broader statistical validation of the CMR CUE for independent implementation across a range of student populations. After a description of the student population, data sources, and analysis (Sec. II), we report multiple test statistics relating to the validity and reliability of the CMR CUE (Sec. III), including item difficulty, item and whole-test discrimination, internal consistency, and overall consistency of the CMR CUE with other measures of student performance. We also present a more detailed analysis of student responses to individual questions to investigate consistency across an individual student's responses (Sec. III), as well as how student responses can be used to gain insight into underlying student difficulties (Sec. IV). Finally, we end with a discussion of limitations and implications (Sec. V).

II. CONTEXT AND METHODS

Following the initial comparison of the CMR and free-response versions of the CUE reported previously [8], we set out to more robustly establish the validity and reliability of the new version as an independent instrument. To do this, we expanded our data collection with an emphasis on including additional students and instructors at multiple institutions beyond the developing institution. We recruited instructors to pilot the CMR CUE in several ways, including soliciting participants during talks and posters presenting the results of the initial comparison study at professional meetings and workshops (e.g., the American Association of Physic Teachers summer meetings). The new version was also uploaded to our password-protected online materials repository (see Ref. [9]), where it can be accessed by any physics instructor interested in using our transformed course materials. We also contacted a number of colleagues working in physics education research who facilitated putting us in contact with the instructor in their department who was teaching electrostatics.

Ultimately, we collected post-test CUE data from 15 courses spanning 9 institutions and 13 instructors. Institution and course characteristics are shown in Table I. We also have pretest data from 13 of these courses.

TABLE I. General characteristics of each institution where we collected post-test CMR CUE data. N indicates the number of responses rather than the total number of students enrolled.

Institution code	Institution type	Highest degree[a]	Size (undergrads)	N courses	N students
R1-A	Public	Ph.D.	25 000	5[b]	193
R1-C	Public	Ph.D.	37 000	1	40
R1-D	Public	Ph.D.	40 000	1	30
R1-E	Public	Ph.D.	29 000	1[c]	67
R2-A	Public	Ph.D.	19 000	2	33
BG-B	Private	B.S.	4000	2	23
BG-C	Private	B.A.	3000	1	8
BG-E	Private	B.S.	2000	1	8
BG-F	Private	B.S.	3000	1	19

[a] Highest degree offered directly by the Physics Department.

[b] These courses include the two semesters in which we conducted the comparison study described in Ref. [8]. Only data from students who took the CMR version are included in the total N from these courses.

[c] The post-test was taken online at this institution.

The pretest version of the CUE is composed of the subset of the post-test questions that can, in theory, be answered using only introductory electrostatics. The response options for the pretest have also been modified to remove all jargon and techniques that students have not seen previously (e.g., multipole expansion, separation of variables). Pretests were administered in the first week of class as either 20-min in-class activities ($N = 7$) or as an out-of-class online survey ($N = 6$). For all courses but one, post-tests were administered in the last week of class as a 50-minute in-class activity. In one of the included courses, the post-test was given as an out-of-class online survey.

To what extent in-class implementations of the pre- and post-tests can be compared to out-of-class, online implementations is still an open question that we will not attempt to robustly answer here. However, for students at CU who took both the pre- and post-test, pretest data show an average score of $31.8 \pm 1.5\%$ when the pretest was taken online ($N = 102$) compared to $30.9 \pm 1.5\%$ when taken in class ($N = 126$). This indicates that, for the pretest, in-class and online implementations are likely comparable. The post-test, however, is a considerably longer and harder instrument, and it may be that scores on a 50-min assessment administered online and in class are not directly comparable. However, for the single course where the post-test was given online, the average score and standard deviation were consistent with that from in-class implementations. Moreover, the inclusion of this course does not significantly change any of the statistics or conclusions reported in the rest of this section. As such, we have opted to include these data in the following analysis in order to realize greater statistical power.

Consistent with the majority of conceptual assessments in physics, our analysis of the validity and reliability of the CMR CUE will be guided by classical test theory (CTT) [10]. CTT posits a number of characteristics of a high-quality assessment and provides various test statistics that quantify how well an instrument matches these characteristics. For polytomously scored assessments like the CMR CUE, these statistics include [10] item difficulty as measured by the average score on each individual item, item discrimination as measured by Pearson correlation coefficients [11] of item scores with the rest of the test, internal consistency as measured by Cronbach's alpha [12], and whole-test discrimination as measured by Ferguson's delta [13]. Each of these test statistics will be discussed in greater detail in Sec. III.

One significant drawback of CTT is that all test statistics are population dependent. As a consequence, there is no guarantee that test statistics calculated for one student population (e.g., physics students at a community college) will hold for another population (e.g., physics students at a university). For this reason, scores on assessments validated through the use of CTT can only be clearly interpreted inasmuch as the student population matches the population with which the assessment was validated. For additional discussion of the limitations of CTT, see Ref. [14]. To address the shortcomings of CTT, psychometricians later developed item response theory (IRT). To explicate our decision not to utilize IRT, the remainder of this section will discuss some of its advantages and disadvantages. In the simplest IRT model (i.e., the Rasch model [15]), a student's performance on individual items is assumed to depend only on their latent ability and the item difficulty. More complex IRT models also include parameters to account for item discrimination and student guessing. For test items that fit this model, all item and student parameters can be determined in such a way as they are independent of both population and test form [13,16].

Despite the appeal of generating population-independent parameters, there are several significant drawbacks to IRT as a potential tool to develop upper-division physics assessments. Even the simplest dichotomous IRT models require large N (> 100) to produce estimates of item and student parameters that are reliable enough for low-stakes testing [13,17]. This number increases for more complex models that, for example, include item discrimination parameters, or for instruments with polytomous scoring [17]. The small class sizes typical of upper-division physics would necessitate classroom testing at multiple institutions, possibly over multiple semesters, to collect this volume of data. Additionally, in order for the parameters generated by IRT to be truly population independent, they must fit the appropriate IRT model. Crafting a large number of items that fit these models often requires multiple iterations of preliminary testing, further increasing the number of students necessary to develop and validate an assessment. Because, in large part, of the logistical barriers to IRT, this analysis will exclusively utilize CTT.

III. RESULTS: STATISTICAL VALIDATION

This section presents the statistical validation of the CMR CUE. Using the nuanced grading rubric described in Sec. I, the overall average on the CMR CUE is $52.6 \pm 0.9\%$ ($\sigma = 18.9\%$) when treating students as data points. The distribution of $N = 421$ scores is shown in Fig. 2. The distribution is slightly non-normal (Anderson-Darling test [18], $p = 0.03$), due in part to the slight positive skew. Averaging by students differentially weights the impact of large courses, which, in these data, come exclusively from large research institutions (Table I). This effect can be reduced by considering performance by course, rather than by students. Taking the mean of the average scores for each course, the overall performance on the CMR CUE is $50.3 \pm 2.5\%$ ($\sigma = 9.6\%$). With only $N = 15$ courses, the difference between the by-course and by-student averages is not statistically significant; however, we argue that a difference of 2% is also not of practical significance. Thus, we will treat student scores as data points for the remainder of this analysis.

A. Criterion validity

To establish the extent to which scores on the CMR CUE are consistent with other, related learning outcomes, we would ideally correlate these scores with final course grades and/or aggregate exam scores for all students in our sample. Unfortunately, we only have access to final course and exam scores for a subset of the students at CU ($N = 154$, four courses) and for none of the external institutions. For this subset of our sample, we find a high correlation of CMR CUE scores with aggregate exam score (Pearson's correlation coefficient [11], $r = 0.71$), as well as final course score ($r = 0.64$). To account for differences between the average exam, course, and CUE scores between the semesters, the correlations above are based on standardized scores (z scores) calculated separately for each class using the class mean and standard deviation. This finding establishes the criterion validity of the CUE for the student population at CU; however, we are not able to extend this conclusion to external institutions with the available data.

B. Item difficulty

In addition to looking at the overall performance of students on the CMR CUE, we characterize the difficulty of each item by looking at the average score by question (Fig. 3). Item difficulties for all questions fall between 30% and 75%. We are not aware of a well-established range of acceptable values for item difficulty on polytomously scored items. However, for dichotomously scored items, where item difficulty is measured as the percent of students who answer each item correctly [2], it is typically argued that ideal values should fall halfway between 100% and the percent expected by random guessing [19]. This maximizes the potential discriminatory power of each item. Since not all items will achieve this ideal, one standard range for acceptable values is 30%–90% [2]. Extending this same logic of maximizing the potential discriminatory power of each item as well as the test as a whole, we argue that item difficulties for our polytomously scored items fall within an acceptable range, with no single item being too easy or too hard to contribute to the overall discrimination of the test.

Scores on each individual item are rarely normally distributed. This is in part an artifact of the grading scheme in which there are a finite number of potential point combinations (typically between 0 and 5 points in 0.5–1 point intervals). For this reason, the median score on each item is often different from the average score.

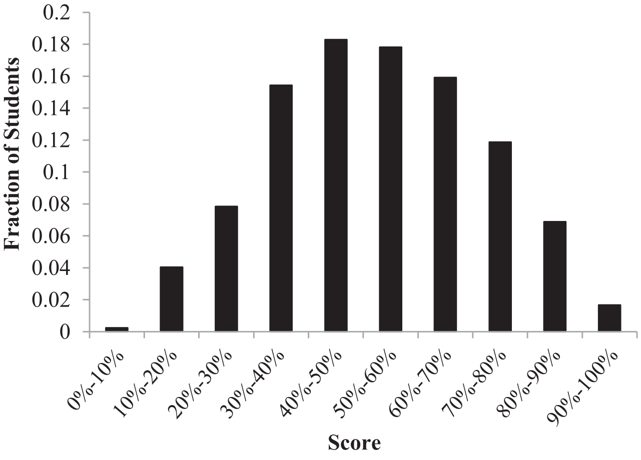


FIG. 2. Distributions of scores ($N = 421$) on the CMR CUE from 15 courses at the institutions described in Table I. These data did not pass a statistical test for normality (Anderson-Darling test, $p = 0.03$).

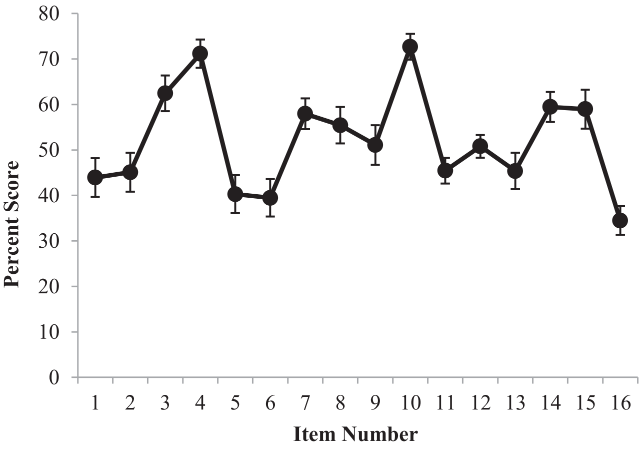


FIG. 3. Average scores for each item on the CMR CUE ($N = 421$). Error bars represent a 95% confidence interval (double the standard error on the mean). Score distributions for each individual item are not necessarily normally distributed.

C. Discrimination

One preliminary indication of the whole-test discrimination of the CMR CUE comes from the overall spread in the distribution of students' scores (Fig. 2). These scores span nearly the full range of possible scores (from 0% to 100%) with a minimum score of 4.3% and a maximum score of 92.5%. Thus, the students are well distributed across the range of possible scores. As another measure of the whole-test discrimination of the CMR CUE, we use Ferguson's delta [13]. Ferguson's delta is a measure of how well scores are distributed over the full range of possible point values (0–93 points). For the full population of students, Ferguson's delta is 0.99. Delta can take on values in the range [0, 1], and anything above 0.9 indicates good discriminator power [2].

We also examine the discrimination of each individual item by comparing a student's score on that item to their performance on the rest of the test. Item-test correlations for all 16 items are shown in Fig. 4, and all correlation coefficients fall between 0.28 and 0.55 and are statistically significant given $N = 421$ [20]. As has been done before [6], we adopt the standard cutoff of $r = 0.2$ used for dichotomously scored items [2] to argue that all items on the CMR CUE demonstrate acceptable discriminatory power.

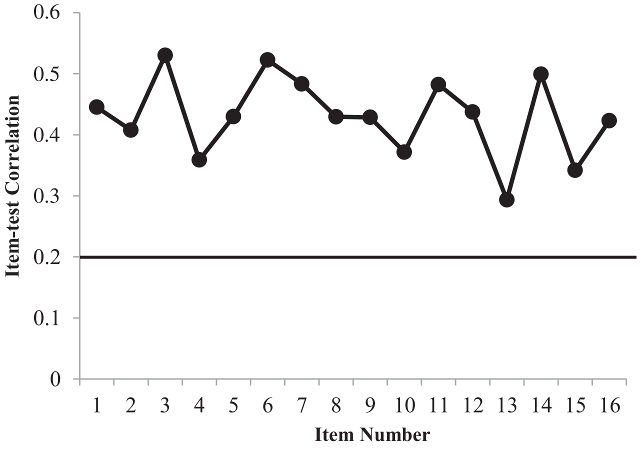


FIG. 4. Item-test correlations (as measured by Pearson's r) for each of the items on the CMR CUE. For $N = 421$, any correlation greater than 0.09 is significant at the $p < 0.05$ level [20]; thus, item-test correlations are statistically significant for all items. The conventional cutoff for an acceptable correlation (0.2) is marked as a bold line.

D. Consistency

The consistency of scores on individual items or subsets of items is another important property of the CMR CUE. We utilize Cronbach's alpha as a conservative measure of the internal consistency of the CMR CUE as a whole. Cronbach's alpha can be interpreted as the average correlation of all possible split-half exams [12]. Alpha is a conservative measure because it assumes a unidimensional assessment, and, while we have no *a priori* reason to assume that the CUE measures a single construct, multidimensionality will tend to drive alpha downwards [12]. For our population of students, we calculate a value of $\alpha = 0.82$. For a test used for low-stakes testing of individuals rather than just groups, the commonly accepted threshold is $\alpha > 0.8$ [10]. Thus, the CMR CUE demonstrates an acceptable level of internal consistency.

In terms of the new CMR format, there is another aspect of consistency that is important to consider. As the name implies, the majority of the questions on the *coupled* multiple-response CUE have several subparts whose scores and/or content are coupled, either explicitly (as with the method and reasoning items, Fig. 1) or implicitly (i.e., there is an opportunity for a student to be consistent or inconsistent in their responses to consecutive subparts). For example, the distribution of method selections for the item shown in Fig. 1 are given in Fig. 5(a). The two most common methods are Gauss's law and superposition, with superposition being the correct response. Figure 5(b) breaks down the reasoning choices for students who selected each of these methods. There is a clear qualitative difference between the reasoning elements selected by these two sets of students. Students who chose superposition were more likely to select reasoning elements "e" and "g," which represent the two elements required to fully justify superposition as the easiest method. Alternatively, students who selected Gauss were more likely to select reasoning elements "d" and "f." Both of these elements are consistent with the use of Gauss's law and represent the commonly expressed justifications for using Gauss to solve this problem.

While Fig. 5 qualitatively suggests a certain degree of consistency between students' method and reasoning selections, we also wanted to get a more quantitative sense of students' consistency. To do this, we assigned a consistency code to students' response to each question (excluding Q8, Q11, Q14, and Q15 which have no consistency check). Students were coded as "consistent" if they selected at least one of the reasoning elements that supported their specific choice of method or answer and no inconsistent elements. Alternatively, if they selected any reasoning elements that were directly inconsistent with their choice of method, they were coded as "inconsistent" regardless of whether they also selected some consistent reasoning elements. The remaining subset of students were coded as "neither," meaning they left one of the two parts blank, chose the "None of These" method option, or selected only reasoning elements that were neither directly consistent nor inconsistent with their choice of method. For example, on Q5 (Fig. 1), the combinations $(\text{Method}, \text{Reasoning}) = (B, df)$ or (A, a) would both be coded as "consistent," whereas the combinations (B, bd)

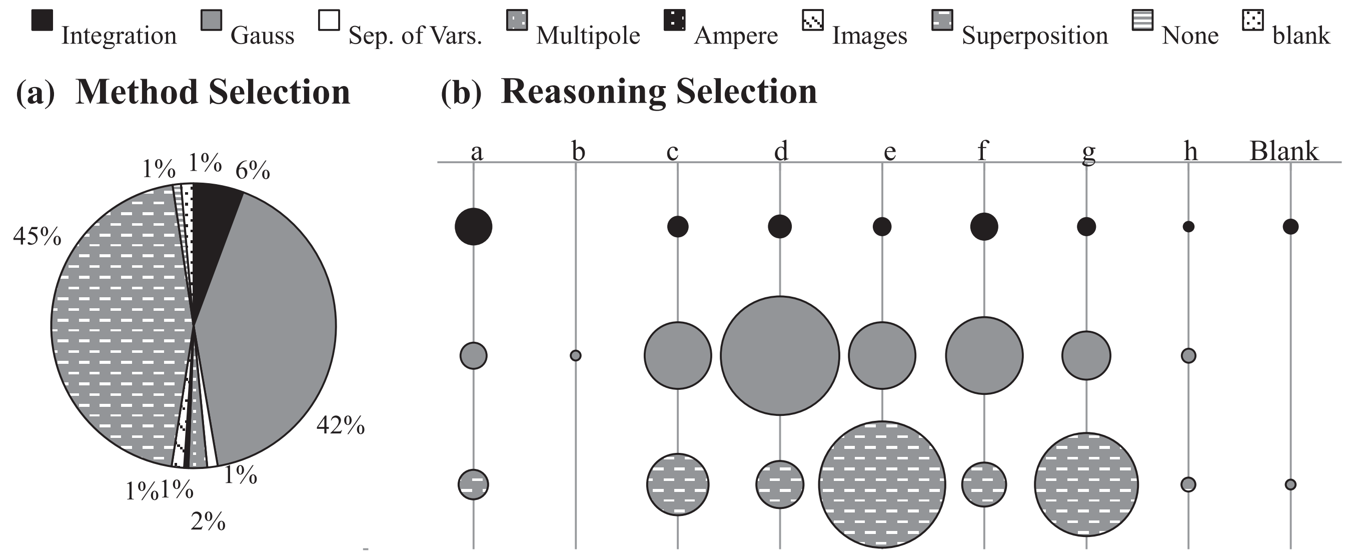


FIG. 5. (a) Method selections for $N = 421$ students on Q5 (Fig. 1). (b) Reasoning selections for the subset of students who selected each of the three most common methods: direct integration ($N = 24$), Gauss's law ($N = 176$), and superposition ($N = 191$). Number of responses for each reasoning element is proportional to the area of the circle.

or (A, af) would be coded as "inconsistent," and the combinations (B, g) or (A, c) would be coded as "neither".

The breakdown of the fraction of students receiving each consistency code is given in Fig. 6. On all questions but one, the fraction of consistent students is ≥ 0.5, and the fraction of inconsistent students is ≤ 0.32. Consistency between Q12 subparts iii and iv is noticeably lower than on other questions. These two subparts ask for qualitative graphs of E_z and V from a finite disk of charge and, for any given response to subpart iii, there is, at most, one consistent response to subpart iv. The relatively small number of potential consistent response patterns and the fact that consistency between these subparts is not explicit in the problem statement both contribute to the greater degree of inconsistency on this question. For all questions, consistent responses do not come exclusively from correct responses. In other words, many students are consistent even when they are incorrect. We take this as an indication that the majority of students are connecting their answers and reasoning selections in reasonable and meaningful ways rather than randomly guessing. This finding further supports the overall validity of the multiple-response format.

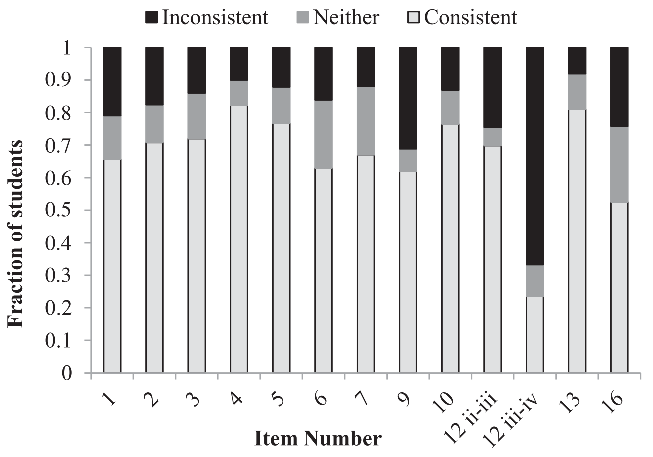


FIG. 6. Fraction of students coded as "consistent" (at least one consistent reasoning element and no inconsistent ones), "inconsistent" (any inconsistent reasoning elements), or "neither" (neither consistent nor inconsistent reasoning elements) on each of the questions that have consistency checks. Note that there are no consistency checks on questions 8, 11, 14, and 15, but there are two possible consistency checks in Q12: between subparts ii and iii, and between subparts iii and iv.

IV. ACCESSING STUDENT DIFFICULTIES WITH THE CMR CUE

The previous sections have supported the validity and reliability of the CMR CUE according to classical test theory. However, in addition to providing a quantitative measure of student outcomes, the CUE also presents an opportunity to gain insight into student difficulties. For example, we have used student responses to several CMR CUE questions in our investigations of student difficulties with the Dirac delta function [21] and separation of variables [22]. In this section, we focus on the question in Fig. 1 (Q5) as an example of using the CMR CUE to think about student difficulties. The distributions of student responses to the remaining questions are available from Ref. [22] but will not be discussed in further detail here.

Q5 (Fig. 1) presents students with a solid sphere with an off-center, spherical cavity carved out of it and asks for the easiest method to find $\vec{E}$ or V outside the sphere. The correct response is superposition (“G”) because you can treat this situation as two oppositely charged spheres (“e”) and superpose the electric fields (“g”) from each uniform sphere (“c”) individually to determine the total electric field at point P. It is also possible, though *much* more difficult, to solve this problem through Direct Integration (“A”) via Coulomb’s law (“a”). A student selecting the former pattern of responses (G,ceg) would receive full credit (5 points), while a student selecting the latter pattern (A,a) would receive only 1.5 points. The distribution of method selections from this population of students is shown in Fig. 5(a). Almost half of the students (45%, $N = 191$ of 421) correctly selected superposition as the easiest method, and only a small number (6%, $N = 24$ of 421) selected the more difficult method, integration. Of the remaining students, the overwhelming majority (42%, $N = 176$ of 421) selected Gauss’s law.

There are at least two possible reasoning paths that could lead a student to select Gauss’s law as the method for this question [23]. First, they are imagining using a single large Gaussian sphere centered on the origin of the solid sphere (not the cavity) to calculate $\vec{E}$ from Q_{enclosed} (consistent with reasoning elements “d” and “f”). Alternatively, they could be imagining using two Gaussian spheres, one centered on a solid, uniform sphere and one on a solid, negatively charged sphere in place of the cavity (consistent with reasoning elements “e” and “g”). The latter strategy is correct and is awarded full credit in the nuanced grading scheme described in Sec. I, while the former strategy is fundamentally incorrect and is awarded 0 points.

To distinguish between these two lines of reasoning, we must examine the reasoning selections of those students who selected Gauss’s law [Fig. 5(b)]. The two most common reasoning elements selected by these students are “d” and “f,” which supports the conclusion that the majority of these students were following the first (incorrect) line of reasoning. Indeed, of the students who selected Gauss’s law, only a tenth (11%, $N = 20$ of 176) did not select one or both of reasoning elements “d” or “f.” Only 10 of the remaining students selected both reasoning elements “e” and “g,” suggesting that they were using the second (correct) line of reasoning. This finding is consistent with previous research [24] and our own findings [25] that suggest students often misapply Gauss’s law. In this case, the majority of students have argued that the location of the cavity does not matter, suggesting either that they have not recognized that the asymmetrical location of the cavity breaks the symmetry of the electric field or that they have not recognized that the asymmetry of the electric field eliminates Gauss’s law as a potential solution method. However, it is not possible to decide which of these two issues is at play for a particular student given only their response to this question.

As superposition is the correct response to this question, it is tempting to assume that any student selecting method “G” understands the correct solution method. This conclusion is generally supported by the observation that the most common reasoning elements selected by these students are “e” and “g.” However, just under a fifth of students who selected superposition (18%, $N = 34$ of 191) also selected one or both of reasoning elements “d” and “f,” suggesting that these students were thinking only about superposition of charges (i.e., $Q_{\mathrm{large}} - Q_{\mathrm{small}}$) rather than fields (i.e., $\vec{E}_{\mathrm{large}} - \vec{E}_{\mathrm{small}}$). This distinction between superposition of charges rather than fields was also observed in previous research examining student responses to the free-response version of the CUE [23]. Both this result and the finding that a small number of students ($N = 10$) selected Gauss’s law along with reasoning elements that suggest a correct strategy underscore the importance of asking students to express their reasoning to avoid misinterpreting student responses.

V. SUMMARY AND DISCUSSION

We previously created a multiple-response version of an existing upper-division conceptual assessment, the CUE. This new version utilizes a novel approach to multiple-choice questions that allows students to select multiple reasoning elements in order to construct a complete justification for their answers. By awarding points based on the accuracy and consistency of students’ selections, this assessment has the potential to produce scores that represent a more fine-grained measure of students’ understanding of electrostatics than a standard multiple-choice test. Previous research demonstrated that the multiple-response and free-response versions of the CUE resulted in similar student performance and showed a high degree of consistency on multiple measures of test validity and reliability.

We collected scores on the CMR CUE from multiple courses at multiple institutions. These data support the validity and reliability of the instrument as measured by various test statistics including item difficulty, item discrimination, and internal consistency. We also examined the consistency of students’ responses on consecutive subparts of individual questions. These data showed that the majority of students selected responses that were internally consistent even when the overall response was incorrect. Additionally, as an example of using the CMR CUE to gain insight into student difficulties, we demonstrated that student responses to one question support the findings from previous research that students tend to misapply Gauss’s law in nonsymmetric situations. These finding support the overall validity of the CMR CUE as a research-based assessment that can be used to reliably measure some aspects of student learning in upper-division electrostatics.

Some potential limitations of the CMR CUE include its content coverage and presentation. Both versions of the

CUE were designed to be consistent with Griffiths's text [7] in terms of both scope and wording. Instructors not using Griffiths should carefully examine the content of the CUE to ensure that it matches their content learning goals and coverage. For example, feedback from some external institutions has indicated that some students may be less likely to use or interpret the term "superposition" in the same way as it is used in the CUE [23]. Moreover, the CMR CUE was validated using classical test theory, and thus all test statistics are population dependent. If used to assess a significantly different population of students, care should be taken to ensure that students' scores can still be reliably and accurately interpreted.

ACKNOWLEDGMENTS

We give particular thanks to Stephanie Chasteen for all her work developing and validating the original, free-response version of the CUE, and to the members of the PER@C group for their feedback. This work was funded by NSF-CCLI Grant No. DUE-1023028 and a NSF Graduate Research Fellowship under Grant No. DGE 1144083.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[2] 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).

[3] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in introductory physics, Phys. Rev. ST Phys. Educ. Res. **1**, 010101 (2005).

[4] http://www.ncsu.edu/per/TestInfo.html.

[5] B. R. Wilcox, M. D. Caballero, C. Baily, H. Sadaghiani, S. V. Chasteen, Q. X. Ryan, and S. J. Pollock, Development and uses of upper-division conceptual assessment, Phys. Rev. ST Phys. Educ. Res. **11**, 020115 (2015).

[6] S. V. Chasteen, R. E. Pepper, M. D. Caballero, S. J. Pollock, and K. K. Perkins, Colorado upper-division electrostatics diagnostic: A conceptual assessment for the junior level, Phys. Rev. ST Phys. Educ. Res. **8**, 020108 (2012).

[7] D. J. Griffiths, *Introduction to Electrodynamics* (Prentice-Hall, Englewood Cliffs, NJ, 1999), ISBN: 9780138053260.

[8] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020124 (2014).

[9] http://per.colorado.edu/Electrostatics.

[10] P. Engelhardt, *Getting Started in PER* (2009), Vol. 2.

[11] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic press, New York, 1977).

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

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

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

[15] L. Ding, in *Proceedings of the Physics Education Research Conference, Omaha, 2011* (American Institute of Physics, Omaha, Nebraska, 2012), Vol. 1413, pp. 175–178.

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

[17] R. J. De Ayala, *Theory and Practice of Item Response Theory* (Guilford Publications, New York, 2009).

[18] M. A. Stephens, EDF statistics for goodness of fit and some comparisons, J. Am. Stat. Assoc. **69**, 730 (1974).

[19] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction* (ERIC Clearinghouse on Assessment and Evaluation, College Park, Maryland, 1980).

[20] B. L. Weathington, C. J. Cunningham, and D. J. Pittenger, *Understanding Business Research* (John Wiley & Sons, New York, 2012).

[21] B. R. Wilcox and S. J. Pollock, Upper-division student difficulties with the dirac delta function, Phys. Rev. ST Phys. Educ. Res. **11**, 010108 (2015).

[22] B. Wilcox, Ph. D. thesis, University of Colorado Boulder, 2015.

[23] J. Zwolak, M. B. Kustusch, and C. Manogue, in *Proceedings of the Physics Education Research Conference, Portland, 2013* (American Institute of Physics, Portland, Oregon, 2014), pp. 385–388.

[24] R. Pepper, S. Chasteen, S. Pollock, and K. Perkins, in *Proceedings of the Physics Education Research Conference, Portland, 2010* (American Institute of Physics, Portland, Oregon, 2011), Vol. 1289, pp. 245–248.

[25] B. R. Wilcox, M. D. Caballero, D. A. Rehn, and S. J. Pollock, Analytic framework for students' use of mathematics in upper-division physics, Phys. Rev. ST Phys. Educ. Res. **9**, 020119 (2013).