Multiple-response assessment for upper-division electrodynamics

Qing X. Ryan[1], Charles Baily[2] and Steven J. Pollock[3]

[1]*Department of Physics and Astronomy, California State Polytechnic University Pomona, CA, 91768*
[2]*School of Physics and Astronomy, University of St. Andrews, St Andrews, Fife KY16 9AJ, United Kingdom*
[3]*Department of Physics, University of Colorado Boulder, CO, 80309*

The CURrENT (Colorado UppeR-division ElectrodyNamics Test) was designed as an open-ended assessment to investigate student reasoning and learning, as well as assessing course transformations in upper-division electrodynamics. The assessment has been given at multiple universities over the past five years, but hand-grading the open-ended questions limits the scalability and usability of the instrument. For this reason, we are creating a multiple-response version of the assessment, using the database that consists of many student responses to the free-response CURrENT along with research on student difficulties. Our goal is to explore the logistical advantages of this objectively gradable format while preserving insights about student reasoning provided by the free-response format. Here we discuss development of the multiple-response CURrENT and present a comparison study between the multiple-response version and the free-response version. Some preliminary measures of the multiple-response CURrENT such as the test's validity, reliability and discrimination using classical test theory are also included.

I. INTRODUCTION

Research-based conceptual assessments play an important role in physics education research. Instruments for introductory physics such as the FCI [1] and BEMA [2] are used to characterize common and persistent student difficulties, as well as to support curricular transformation [3]. Upper-division assessments are developed for similar purposes. One example is the Colorado UppeR-division ElectrodyNamics Test (CURrENT) [4,5,6]. It is a free-response (FR) instrument that is designed to measure a representative sampling of skills and conceptual understanding in junior-level electrodynamics. This assessment has been given for 16 semesters at 9 universities to over 500 students. Validation studies of the CURrENT have been conducted previously and it shows considerable promise for research and assessment in upper-division electrodynamics [6].

Open-ended assessments such as the CURrENT have the advantage of providing rich responses supporting investigations of student reasoning and thought process. However, grading effort (and subjectivity) and time required of the faculty or researcher can still be a barrier for the wide adoption of research-based assessments. Previous research [7,8] has shown great promise using an objectively gradable format to increase the scoring efficiency while providing similar scores as the FR version. Previous work by Lin and Singh showed that carefully crafted research-based multiple-choice (MC) questions can reasonably reflect the relative performance of students on the FR questions [7]. More recent work on the upper-division multiple-response (MR) CUE also showed it is possible to improve the logistics while gaining meaningful insight into the details of common student difficulties [8]. MR format differs from regular MC format in that MC format has one single unambiguous correct answer with several distractors, while MR format allows students to select multiple responses and receive partial credit depending on the accuracy and consistency of their choices.

In order to utilize the logistical advantage of this easy-to-grade format, we have constructed a multiple-response version of the CURrENT, using student responses from previous semesters to help craft distractors. This paper describes the development and scoring of the MR-CURrENT, as well as providing a comparison of scores between the MR and FR versions, along with some preliminary (N=75) quantitative reliability and validity measures for the new version.

II. DEVELOPMENT

Adapting the Questions: The FR-CURrENT was written in such a way that scoring of most questions includes two parts: correctness and reasoning. Since our goal is to capture the same information as the FR version, we followed the same format when adapting these questions into a MR format. For the correctness part, we provide choices with a single, unambiguous correct answer. For the reasoning part, we list a number of common possible reasoning elements that support either a correct or wrong answer. Students are asked to choose ALL elements that support their answer of choice. These reasoning elements were created after analyzing student responses from the FR-CURrENT database, as well as utilizing our knowledge about common student difficulties in the content area [9]. A sample question of the MR-CURrENT is given in Fig.1. The full instrument is available online [4].

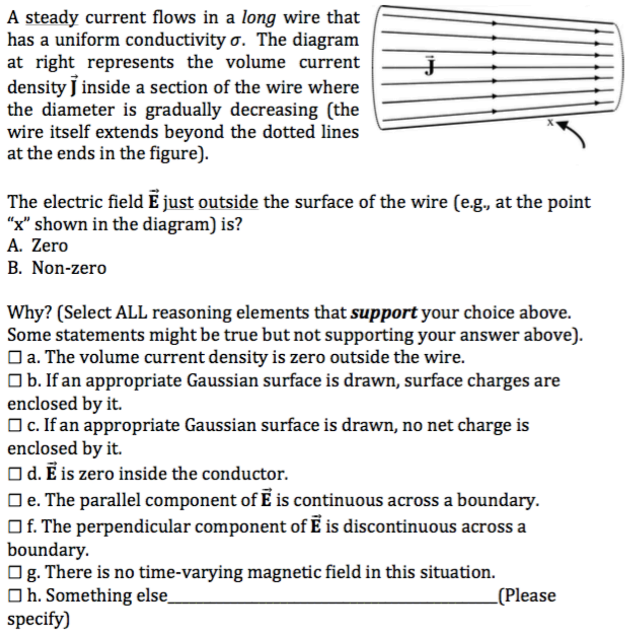
Scoring: The scoring of the MR-CURrENT is done by entering students' choices into a spreadsheet. The results are given instantly[1]. When answering a FR question,

[1] Currently the researchers enter students' answers by hand, but it is possible to further speed up the process by using bubble sheet formats and/or creating an electronic online version.

2016 PERC Proceedings, edited by Jones, Ding, and Traxler; Peer-reviewed, doi:10.1119/perc.2016.pr.066

students may provide a complete justification that captures all the reasoning elements. Students may also give partially correct justifications that are missing one or more key elements. Some students may write conflicting reasoning elements, which shows an inconsistency in their logic. Likewise, we have maintained the possibility to encounter a variety of answer combinations in this MR format [8]. In this way, we can give partial credit for incomplete but correct reasoning, or remove some credit for inconsistent answers. Alternatively, one can choose to give full credit only when the key reasoning elements (and only those elements) are present. One advantage offered by the MR format is flexibility in choosing different grading schemes. We will explore more grading schemes in future work.

The grading scheme currently employed is intended to replicate as closely as possible the grading scheme used in the FR format [4,5]. All data reported in this paper were based on this grading scheme. This scheme only looks for the most important key reasoning element(s), generally ignoring wrong or irrelevant choices, with a few exceptions where partial credit is possible. For the sample question below, if students choose the correct answer (B), then they need to choose either statement b or e (key reasoning element) to get full credit for the reasoning part. Other irrelevant or wrong statements are ignored except statement d, which is directly linked to a wrong conclusion.



A steady current flows in a *long* wire that has a uniform conductivity σ. The diagram at right represents the volume current density $\vec{\mathbf{J}}$ inside a section of the wire where the diameter is gradually decreasing (the wire itself extends beyond the dotted lines at the ends in the figure).

$\vec{\mathbf{J}}$
x

The electric field $\vec{\mathbf{E}}$ just outside the surface of the wire (e.g., at the point "x" shown in the diagram) is?
A. Zero
B. Non-zero

Why? (Select ALL reasoning elements that ***support*** your choice above. Some statements might be true but not supporting your answer above).
□ a. The volume current density is zero outside the wire.
□ b. If an appropriate Gaussian surface is drawn, surface charges are enclosed by it.
□ c. If an appropriate Gaussian surface is drawn, no net charge is enclosed by it.
□ d. $\vec{\mathbf{E}}$ is zero inside the conductor.
□ e. The parallel component of $\vec{\mathbf{E}}$ is continuous across a boundary.
□ f. The perpendicular component of $\vec{\mathbf{E}}$ is discontinuous across a boundary.
□ g. There is no time-varying magnetic field in this situation.
□ h. Something else____________________________(Please specify)

FIG 1. Sample question from the MR-CURrENT.

Feedback from Experts: The design of the FR-CURrENT was guided by course-scale faculty consensus learning goals [4,6]. The instrument was also reviewed by physics experts to establish that the questions aligned with their learning goals, that question language was clear and appropriate, and that the questions were perceived by faculty as interesting and useful measures of student learning. Since the MR version asks the same questions, evaluating the validity of the new instrument is focused on whether or not the new presentation of the same content elicits similar student responses. The development of the new MR-CURrENT was done collaboratively between three physics experts from three different institutions, two of whom have taught the Electrodynamics course for multiple years. We also solicited feedback from two other content experts at different institutions. Small modifications were made to the phrasing of several items as a result of this feedback. The expert reviewers offered no critiques that questioned the overall validity of the new format.

Student Interviews: When changing question format, student interviews need to be conducted to ensure that the questions are written clearly and interpreted correctly by students. During the development stage, we validated the new MR version by conducting think-aloud interviews with six undergraduate students and four graduate students and post-docs. The undergraduate students were physics majors who had taken the second semester of E&M (E&M II: Griffiths Ch.7-12) [10] at the University of Colorado Boulder (CU). The graduate students and post-docs were volunteers from the physics department at CU. During the interviews, interviewees were asked to verbalize their thinking process and the interviewer did not interject except to remind them to verbalize their thinking. At the end of the interview, interviewees were asked about questions in more detail to probe their reasoning, in particular why they chose certain distractors. The interviews were recorded and later analyzed to determine whether student work reflected the intended nature of the question, as well as whether their written work reflected their verbal interpretation of the question. As a result of these interviews, we made some changes in the wording and formatting of the questions, as well as adding or removing several distractors. For example, in the sample question shown in Figure 1, we originally had two separate statements about having no net charge enclosed by the Gaussian surface (statement c in Fig.1), differing on the exact location of the Gaussian surface (spanning the edge of the wire vs. outside). After the interviews, we consolidated the two statements into one because no students chose the distractor where the Gaussian surface is drawn outside.

III. PRELIMINARY RESULTS

A. MR vs. FR comparison

In order to evaluate whether the new MR format can produce a meaningful level of agreement with the FR version, we conducted a direct comparison study. Data were collected from three different institutions over four semesters. Roughly a quarter (27%) were undergraduate physics students at CU taking the second semester of E&M, covering electrodynamics. Another group (28%) was undergraduate students from a public institution (classified as an R2 university [11]). The rest of the students (all physics majors) were from a public research university in the UK. All populations are predominately white and male. In each class, we randomly gave half of the students the MR version and half of the students the FR version. The total number of students who took the MR and FR version was 75 (28 CU students) and 81 respectively (28 CU students). Due to logistical constraints, it was difficult to assign a particular test to a particular student, which means we could not match students based on their exam scores before giving the diagnostic test. So we randomly assigned the two different versions to two halves of the class (that is why there was a slight difference in the number of tests for each version) and matched students of each group afterwards. For each student in the FR group (i.e. who took the FR-CURrENT), we tried to find the best matched student in the MR group (who took the MR-CURrENT) using course grade as the first matching variable. We allowed for a maximum of ±5 points (out of 100) when pair-matching the two groups. If there was more than one potential match, we used their final exam score as a secondary matching variable and picked the best match. The final matched data set consists of 122 students from both FR and MR groups, with 61 in each group.

Fig.2 shows that the total score distributions for the MR and FR versions are similar, with no statistically significant differences (Two-sample t-test p=0.2). Both distributions are nearly normal (Anderson-darling test, p=0.54 and 0.52 for FR and MR respectively), and have similar variances (Brown-Forsythe test, p=0.08).

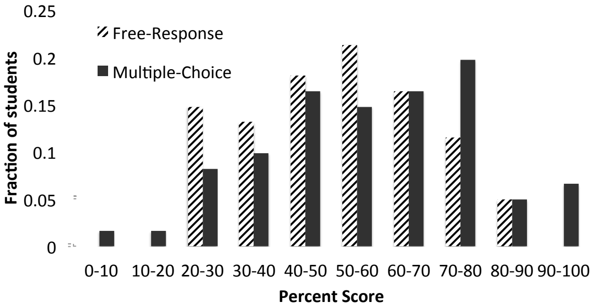


FIG 2. Distributions of scores on the MR and FR CURrENT.

The total average score on the MR-CURrENT 56.9%± 2.7% does not differ statistically from the total average on the FR-CURrENT 52.6%±2.2% (Two-sample t-test p=0.2)[2]. We also compared the average score per question between the FR and MR format (Fig.3). The two formats have comparable average scores for each question, none of the differences are statistically significant (Mann-Whitney test, p values: 0.12(Q2)-0.70 (Q6)).

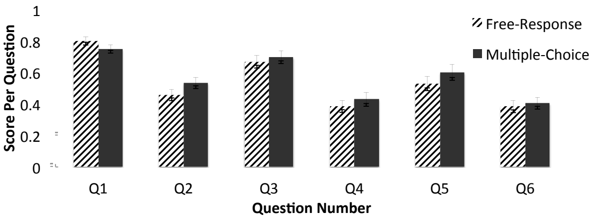


FIG 3. Score per question on the MR and FR CURrENT.

B. MR-CURrENT Statistics

Statistical validation measures of the original FR-CURrENT have already been conducted with a much larger data set [6]. For the remainder of this paper, we present some preliminary statistical measures of the new MR-CURrENT. The entire data set of MR-CURrENT (N=75) was used for this data analysis. The statistical measures reported here were based on classical test theory [12] and consist of the following: Internal consistency (measure of reliability), Criterion validity (measure of validity), Item difficulty, Item-test correlation, and Coefficient of test discrimination (the latter three are all measures of discrimination).

Internal Consistency: Internal consistency is a measure of reliability which is defined as the overall consistency and stability of a test measure. Since the MR format eliminates subjective grading bias, there is no need to investigate inter-rater reliability. Therefore we concentrate on internal consistency which investigates if students' performance on any given test item correlate with the remaining items on the test. Cronbach's alpha (α) is a statistical measure of internal consistency. We treat each sub-question as a single test item (15 sub questions total), and obtain α=0.80 (N=75), where α-values between 0.7-0.9 are traditionally considered adequate [13]. We also computed α more conservatively by treating each question (6 total, including all sub-parts) as one test item and obtained α=0.79. This suggests we have achieved an acceptable level of consistency with the new MR format.

Criterion Validity: Validity is defined as the extent to which test scores measure the intended concept or construct. In order to investigate if the MR-CURrENT gives similar results to other approaches that measure the same construct, we looked at how well the results given by the assessment correlate with students' final exam scores as well as their course grade. Given that the data were collected from different courses, the correlation was computed using z-scores for both final exam and course grade. Scores on the MR-CURrENT correlate with students' final exams in their junior E&M course (Pearson correlation coefficient r=0.53, p<0.001, N=75) and their course grades (r=0.43, p<0.001, N=75). These correlations are considered ''medium'' (0.3–0.5) to ''strong'' (0.5–1.0) [14], suggesting that the constructs measured on the MR-

[2] Similar results were obtained for the two groups even without matching. The total average score on the MR-CURrENT for the entire data set (N=75) was 55.5% ± 2.5% and the total average score on the FR-CURrENT for the entire data set (N=81) was 55.2% ± 2.0%. We will continue to explore the differences in score distributions of the two formats as we gather more data in the future.

CURrENT are related to other aspects of student performance typically valued by faculty.

Item Difficulty: For a test to have good discrimination power between high and low performing students, we expect to see a reasonable level of difficulty for the test items. As shown in figure 3, there are no statistically significant differences between the MR and FR version in terms of the average score on individual questions. This suggests the new MR format offers a comparable difficulty level to the FR format. The overall pattern of item-difficulty is consistent with what was obtained before with the FR format [6] (e.g: students tend to score lower on certain questions such as Q4&6), indicating the MR format is likely to give similar insights about common student difficulties as the FR format. With the automatic grading spreadsheet, we can also compute the percentage of each answer choice easily, which provides the potential to give more detailed insights into common student difficulties.

Item-test Correlation: We expect that students who score well on the test as a whole will tend to score well on individual items. One measure of the discriminatory power of a test is to examine how well performance on each item compares to performance on the rest of the test. Item-test correlations were between 0.49 and 0.60 for all questions on the MR-CURrENT. Minimum acceptable correlation coefficients are generally considered to be around 0.2 [2]. The MR format shows comparable item-test correlations to the FR format (0.4-0.49 [6]).

Coefficient of Test Discrimination: Ferguson's delta (δ), or the "coefficient of test discrimination" [15], measures the discriminatory power of a test by investigating how broadly the total scores of a sample are distributed over the possible range [2]. We obtained δ =0.98 for both the MR format (N=75) and the FR format (N=81), both are consistent with what was obtained before with the FR-CURrENT (0.98). The possible range of δ values is [0,1]. Traditionally, $\delta > 0.9$ is considered good discrimination and thus MC-CURrENT offers similar substantial discrimination power compared to the FR format in differentiating students with different abilities.

IV. DISCUSSION

We have created a multiple-response format for an existing upper-division diagnostic test (the CURrENT: Colorado UppeR-division ElectrodyNamics Test). Design of the distractors was guided by research on common student difficulties as well as student responses to the original free-response version of the instrument. The new MR-CURrENT has logistical advantages to make large-scale implementation much easier. It also allows us to probe student thinking more deeply than a standard MC format by awarding points based on the accuracy and consistency of students' selections of reasoning elements.

A quantitative comparison study was conducted to compare the overall score distribution between the MR and FR formats. For our sample, there was no statistically significant difference of the total score between the two versions. Scores on individual questions also have a high level of agreement between the two versions. We also conducted a direct analysis of the validity and reliability of the MR-CURrENT. The MR-CURrENT score correlates well with other variables, such as final exams and course grades, that are typically valued by faculty. The test also shows high internal consistency and good discriminatory power. Given the scoring efficiency, MR-CURrENT shows promise for large-scale testing implementation.

Future work includes giving MR-CURrENT at more diverse institutions and gathering more data to establish the robustness of the statistics reported here. As we gather more data, we can also examine the ways in which we lose some insights into students reasoning by switching to a multiple-response format [8], which is a common concern with non free-response assessments. We will also explore different grading schemes. For example, we can examine consistency between students' choice of the reasoning elements and their answer; as well as investigating if there are certain connections between different reasoning elements. All of these are advantages given by the new format and can add to our knowledge about students' reasoning and common difficulties.

ACKNOWLEDGEMENTS

We gratefully acknowledge contributions of many students and faculty including: A. Becker, P. Kohl, R. Pepper, and B. Wilcox. We especially thank C. Astolfi for some of the preliminary work she did on this project.

[1] D. Hestenes, et.al., *Phys. Teacher* 30, 141-158 (1992).
[2] L. Ding, et.al., *Phys. Rev. PER*. 2 (1), 7 (2006).
[3] D. Meltzer & R. Thornton. *AJP* 80, 478 (1992).
[4] http://per.colorado.edu/sei, or http://physport.org
[5] C. Baily, et.al., *PERC Proc*. 2012 p.54-57.
[6] Q. X. Ryan, et.al., *PERC Proc*. 2014 p.231-234.
[7] S.-Y.Lin, et.al., *PERC Proc*. 2011, p.47-50.
[8] B. Wilcox, et.al., *Phys. Rev. PER*.10, 020124 (2014)
[9] R. Pepper, et.al., *Phys. Rev. PER*.8, 010111 (2012)
[10] D. Griffiths, Introduction to Electrodynamics, 3rd Ed. (Prentice-Hall, Upper-Saddle River NJ, 1999).
[11] Carnegie Classification of Institutions of Higher Educ.
[12] P.V.Engelhardt, Getting Started in PER (2009), Vol.2.
[13] J. Nunnally, (1978). Psychometric theory (NY: McGraw-Hill, 1978) 2nd ed.
[14] J. Cohen, Statistical Power Analysis for the Behavioral Sciences (Routledge, NY, '88), 2nd ed.
[15] G. Goldstein, et.al., Handbook of Psychological Assessment, (Kidlington, Oxford, UK,2000), 3rd ed.