Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment

Lin Ding, Ruth Chabay, Bruce Sherwood, and Robert Beichner
Department of Physics, North Carolina State University, Raleigh, North Carolina 27695, USA

(Received 23 June 2005; published 15 March 2006)

The Brief Electricity and Magnetism Assessment (BEMA), developed by Chabay and Sherwood, was designed to assess student understanding of basic electricity and magnetism concepts covered in college-level calculus-based introductory physics courses. To evaluate the reliability and discriminatory power of this assessment tool, we performed statistical tests focusing both on item analyses (item difficulty index, item discrimination index, and item point biserial coefficient) and on the entire test (test reliability and Ferguson's delta). The results indicate that BEMA is a reliable assessment tool.

DOI: 10.1103/PhysRevSTPER.2.010105 PACS number(s): 01.40.Fk, 01.40.G−

I. INTRODUCTION

Standardized, multiple-choice tests can be a useful tool in assessing what students learn in physics courses. A number of such tests have been developed; these tests, covering different physics domains including kinematics,[1] force,[2] motion,[3] dc circuits,[4,5] electricity and magnetism,[6] and other topics, have increasingly been used by a wide range of physics instructors to measure some aspects of what students learn in both traditional and reform physics courses. BEMA (Brief Electricity and Magnetism Assessment)[7] was developed in 1997 by Chabay and Sherwood, aided by Fred Reif, to measure students' qualitative understanding and retention of basic concepts in electricity and magnetism. We report elsewhere on the use of BEMA to compare student performance at the end of both traditional and reform introductory electricity and magnetism (E&M) courses and to compare retention of these concepts over a period of up to five semesters after the end of the courses.[8] The test itself is not included here because the utility of a standardized test decreases if its contents are widely known and the questions become very familiar to the population who will be tested. Any instructor may obtain a copy of the test at http://www.compadre.org.

In this paper we report on the reliability of BEMA, as measured by statistical tests focusing both on individual items and on the test as a whole. Test reliability has two aspects: consistency and discriminatory power. A test is reliable if it is consistent within itself and consistent across time. If a test is shown to be reliable, one can have confidence that the same students would get the same score if they took the test more than once. In addition, on a reliable test, a large fraction of the variance in scores is caused by systematic variation in the population of test takers; students whose levels of understanding or mastery are different will achieve different scores on the test. Both of these aspects of test reliability can be assessed statistically. If a test is to be used in comparing the performance of different groups, the reliability of the assessment instrument is particularly important.

To be useful, a test must also be "valid." A test is valid if the skills or knowledge it measures are directly relevant to the stated domain of the test. Validity cannot be assessed statistically and is usually determined by a consensus of expert opinions. Though the issue of validity—the question of whether BEMA in fact assesses knowledge of E&M —is not one of the main topics of this paper, it is involved in the overall evaluation of BEMA. We will briefly address the validity of BEMA in Sec. II.

Aubrecht and Aubrecht[9] were among the first to describe the use of statistical methods to evaluate objective physics tests. The measures they employed included the item difficulty index, the item discrimination index, and test reliability. Subsequently others introduced additional statistical tests, including item point biserial coefficient and Ferguson's delta. Although all of these statistical measures are available for some published assessment tools such as the TUG-K (Ref. 1) and DIRECT (Refs. 4 and 5), many authors have limited their focus to individual item analyses, such as the item difficulty index and test reliability. In Sec. III we will report on the results of applying all these statistical tests to BEMA and will explain briefly the nature and significance of each test.

II. BACKGROUND AND VALIDITY OF BEMA

BEMA is a 30-item multiple-choice test which covers the main topics discussed in both the traditional calculus-based E&M physics curriculum and the matter and interactions curriculum (*Matter & Interactions II: Electric and Magnetic Interactions*[10]). It was originally designed for a retention study measuring students' knowledge of E&M at times ranging from three months to five semesters after completing an introductory E&M course. Test items are mostly qualitative questions with a few semiquantitative questions, which require only simple calculations. All test items are intended to assess students' understanding of basic concepts in calculus-based introductory E&M courses. An example of a question from BEMA is shown in Fig. 1.

A. An example of a BEMA question

The test was designed to incorporate broad coverage of elementary E&M, rather than to probe any particular concept in great detail. Since the population of interest included students who had taken both traditional and reform introductory physics courses, only questions on topics common to both

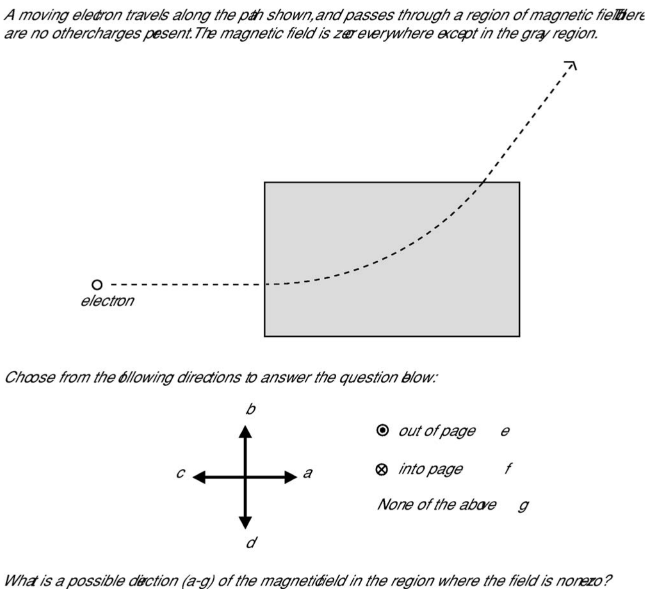
A moving electron travels along the path shown, and passes through a region of magnetic field. There are no other charges present. The magnetic field is zero everywhere except in the gray region.
electron
Choose from the following directions to answer the question below:
b
c
a
d
out of page e
into page f
None of the above g
What is a possible direction (a-g) of the magnetic field in the region where the field is nonzero?

FIG. 1. Post-test results (30 items).

courses were included in the test. To establish the validity of the test, initial drafts of the test were critiqued by all eight faculty members at Carnegie Mellon University (CMU) who had taught undergraduate E&M at any level (introductory or intermediate E&M) within the past five years. If an instructor reported that a proposed question dealt with a topic not covered in the version of the introductory course he or she had taught, the question was eliminated; the final set of questions was approved by all professors consulted, who agreed that the test did deal with important aspects of E&M .

Soliciting expert opinions is a standard method of assessing the validity of a test. The term “validity,” which is not a statistical construct, refers to the extent to which a test actually measures what it purports to measure. Validity can have several aspects.[11] “Face validity” can be determined by a surface level, common sense reading of an instrument; a test would lack face validity if it tested concepts not related to the subject matter. “Content validity” reflects the coverage of the subject matter—does a test cover enough aspects of a specific topic? Both of these aspects of validity are typically assessed by expert consensus, as was done with BEMA. (Other aspects of validity, not relevant here, are “construct validity”—the extent to which the test is demonstrated to measure a theoretical construct or trait such as creativity, honesty, or intelligence—and “criterion-related validity”—evidence that performance on one assessment instrument can be used to make inferences about performance in a different domain.)

Pilot testing was done with a small group of volunteers including senior physics majors who had recently completed the junior-level intermediate E&M course. The initial version of the test contained both multiple choice questions, whose alternatives were based on common errors made by students on written tests, and a small number of short-answer semi-quantitative questions, which were later converted to multiple-choice questions by including common incorrect responses as alternative answers. (We thank Tom Foster for converting the short-answer questions to multiple-choice questions.)

TABLE I. Post-test results.

Sample	Number of students	Average (percentage)	Standard deviation	Standard error
CMU	189	13.6 (45%)	5.9	0.43
NCSU	245	11.6 (39%)	6.0	0.38
Overall	434	12.5 (42%)	6.0	0.29

III. STATISTICAL EVALUATION OF BEMA

BEMA was first administered to 189 paid volunteers at CMU in the spring of 1997.[7] All of these students had completed either the traditional calculus-based introductory E&M course or the matter and interactions (M&I) version of (E&M) at some time before they took the test; most were science, computer science, or engineering majors. This was not an end-of-course assessment; the elapsed time between completion of the E&M course and the BEMA assessment varied from three months up to five semesters. Since this was a longitudinal study, BEMA was also administered to a control group of students who had just completed the first-semester physics course (classical mechanics and thermal physics) and were ready to take the second-semester E&M course. In the fall 2003 semester, BEMA was administered as both a pre- and post-test to a large number of students at North Carolina State University (NCSU) via WebAssign, a computer-based homework system. (WebAssign is an online homework system. It is a centrally hosted subscription service with users from many different institutions. For more information see http://www.webassign.net) All students were taking either a traditional calculus-based E&M course or an M&I course in that semester. Two hundred and forty-five students took the post-test, and 191 students took both pre- and post-tests. Students were asked to take the tests seriously with no penalty for wrong answers.

Pretest performance on BEMA does not vary much among different populations, and pretest scores average around 23%. In this paper we use only post-instructional data for test statistics, since we are focusing on evaluation of BEMA, and not on a comparison of student pre- and post-instructional performance. Post-instruction averages, standard deviations, and standard error for students at CMU, NCSU, and the combined groups from both CMU and NCSU are given in Table I. For comparison, the average score of senior physics majors at CMU was 80%.

Using the data from this combined sample, we performed five statistical tests: three measures focusing on individual test items (item difficulty index, item discrimination index, item point biserial coefficient) and two measures focusing on the test as a whole (test reliability and test Ferguson’s δ). In the following sections, each test is briefly explained and the results discussed. Sections III A–III C discuss statistical measures focusing on individual test items, while Secs. III D and III E discuss statistical measures focusing on the test as a whole.

A. Item difficulty index

The item difficulty index (P) is a measure of the difficulty of a single test question. It is calculated by taking the ratio of

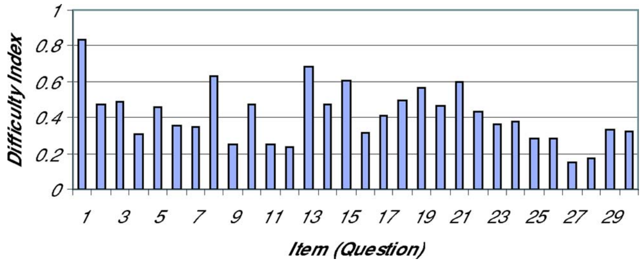

FIG. 2. (Color) BEMA item difficulty indices for each question, based on a combined sample of 434 students.

the number N_1 of correct responses on the question to the total number N of students who attempted the question:

$$P = \frac{N_1}{N}.$$

This difficulty index P might more meaningfully have been called the "easiness index," since it is simply the proportion of correct responses on a particular question. The greater the P value is, the higher the percentage of respondents giving the correct answer and the easier this item is for this population. The range for the difficulty index P value is $[0, 1]$ If the P value is 0, then no one can answer the question correctly; on the other hand, if the P value is 1, then every one can correctly answer this question. Under most circumstances, such extremes should be avoided in a test.

A noteworthy aspect of the difficulty index is that the P value depends on the particular population taking the test. As an example, consider the first question in BEMA. Of 189 CMU students, 168 students answered correctly, so the difficulty index for the first question is 0.89. Among 245 NCSU students who took the post-test, 194 students chose the correct answer, so the difficulty index for the first question is only 0.79 for this population of students.

There are a number of different possible criteria for acceptable values of the difficulty index for a test.[12] In evaluating BEMA, we choose a widely adopted criterion that requires the difficulty index value to be between 0.3 and 0.9,[13] a range which includes the optimum value of 0.5. A difficulty level of 0.5 on each question would lead to the highest values of the statistics discussed in the following sections. However, it is difficult to control every item in one test, especially when the number of items (K) in one test becomes large. An averaged difficulty index value ($\bar{P}$) of all the items (P_i) in a test is often used as an indication of the test difficulty:

$$\bar{P} = \frac{1}{K}\sum_{i=1}^{K} P_i.$$

We can compare the $\bar{P}$ value with the criterion chosen to check if it meets a certain standard.

Figure 2 plots the difficulty index P values of each item in BEMA from the combined sample of 434 students. BEMA item difficulty index values range from slightly below 0.2 to slightly above 0.8, with most items being around 0.4–0.5, within the desired range. The averaged difficulty index $\bar{P}$ is 0.42, which also falls into the criterion range $[0.3, 0.9]$.

B. Item discrimination index

The item discrimination index (D) is a measure of the discriminatory power of each item in a test. In other words, it measures the extent to which a single test item distinguishes students who know the material well from those who do not. On a test item with a high discrimination index, students with more robust knowledge will usually answer correctly, while students whose understanding is weaker will usually get the item wrong. (In contrast, a flawed test question might lead more thoughtful students to give answers that are judged wrong, while students who think less deeply give a correct answer.) If a test contains many items with high discrimination indices, the test itself can be useful in separating strong students from weak students in a specific test domain.

In order to calculate the item discrimination index (D), we divide the whole sample of students into two different groups of equal size, a high group (H) and a low group (L), based on whether an individual total score is higher or lower than the median total score of the entire sample. For a specific test item, one counts the number of correct responses in both H and L groups: namely, N_H and N_L. If the total number of students who take the test is N, then the discrimination index D of this item can be calculated as

$$D = \frac{N_H - N_L}{N/2}.$$

In educational and psychological studies, there are several different calculations of discrimination index often employed by researchers. The calculation described above (50%–50%) is the one which we adopted to calculate discrimination indices for BEMA items. Other researchers may use the top 25% as the high group and the bottom 25% as the low group (25%–25%), in which case the discrimination index D can be expressed as

$$D = \frac{N_H(\text{top } 25\,\%\,) - N_L(\text{bottom } 25\,\%\,)}{N/4}.$$

The 50%–50% calculation can underestimate the discriminatory power of test items, since it takes all the students, especially the relatively unstable middle 50%, into account. The 25%–25% calculation uses only the most consistent individuals, reducing the probability of underestimating the discrimination index due to unstable performance, but necessarily discarding half of the available data.

The possible range for the item discrimination index D is $[-1, +1]$, where +1 is the best value and −1 is the worst value. In the extreme ideal case all students in the high group get the item correct and all students in the low group get it wrong, giving a discrimination index D of +1. In the worst case the situation is reversed: everyone in the low group answers correctly, and everyone in the high group gets it wrong. In this case the discrimination index D will be −1. These extreme cases are unlikely, but it is important to eliminate any items with negative discrimination indices. An item is typically considered to provide good discrimination if $D \geqslant 0.3$.[14] Items with a discrimination index lower than 0.3 (but greater than 0) are not necessarily bad, but a majority of the items in a test should have relatively high discrimination

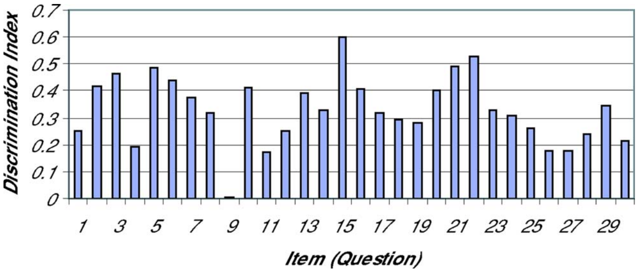

FIG. 3. (Color) BEMA item discrimination index from a combined sample of 434 students. The average discrimination index is 0.33 (50% method).

index values to ensure that the test is capable of distinguishing between strong and weak mastery of the material.

Figure 3 shows the discrimination index for each BEMA item. As one can see, most of the discrimination index D values for BEMA items vary from 0.2 to 0.6, with a majority number of items (18 items) around 0.3–0.4. This shows that most BEMA items have quite satisfactory discriminatory power. We also calculated the averaged discrimination index $\bar{D}$ for all K items (D_i) in BEMA, which can be expressed as

$$\bar{D} = \frac{1}{K}\sum_{i=1}^{K} D_i .$$

We found the average discrimination index $\bar{D}$ for BEMA to be 0.33. This satisfies the criterion that $\bar{D} \geqslant 0.3$. In order to illustrate the underestimation of the 50%–50% calculation, we also computed BEMA item discrimination indices using 25%–25% method. The index values for all 30 items were increased, and the averaged discrimination index $\bar{D}$ for BEMA using 25%–25% calculation is 0.52.

Question 9 has the lowest D value and clearly stands out as different from all the other questions. This question asks students to select an algebraic expression for the conventional current in a pipe containing ionized salt water, given the drift speeds of sodium ions and chloride ions, the cross-sectional area of the pipe, and the number density of the ions. Almost no students get this question correct, probably because systems with more than one kind of mobile charge are not emphasized in introductory E&M courses.

C. Point biserial coefficient

The point biserial coefficient (sometimes referred to as the reliability index for each item) is a measure of consistency of a single test item with the whole test. It reflects the correlation between students' scores on an individual item and their scores on the entire test, and is basically a form of the correlation coefficient. The point biserial coefficient has a possible range of $[-1,+1]$. If an item is highly positively correlated with the whole test, then students with high total scores are more likely to answer the item correctly than are students with low total scores. A negative value indicates that students with low overall scores were the most likely to get a particular item correct and is an indication that the particular test item is probably defective.

To calculate the point biserial coefficient for an item, one needs to calculate the correlation coefficient between the item scores and total scores. A student's score on one item is a dichotomous variable which can have only two values: 1 (correct) or 0 (wrong). Scores for the whole test usually can be viewed as continuous (if the test has a relatively large number of items—say, $\geqslant 20$). The correlation coefficient between a set of dichotomous variables (score for an item) and a set of continuous variables (total scores for the whole test)[15]

$$r_{pbs} = \frac{\bar{X}_1 - \bar{X}}{\sigma_X}\sqrt{\frac{P}{1-P}} .$$

Here $\bar{X}_1$ is the average total score for those students who score 1 for the test item (correctly answer this item), $\bar{X}$ is the average total score for a whole sample, σ_X is the standard deviation of the total score for the whole sample, and P is the difficulty index for this item.

As an example, consider item 1 in BEMA. Among the combined sample of 434 students from CMU and NCSU, 362 students answered the question correctly, so $P=0.83$. For those 362 students, the average total score ($\bar{X}_1$) is 13.52. For all 434 students in the combined sample, the average total score ($\bar{X}$) is 12.50. Together with the standard deviation ($\sigma_X=6.04$) of the total score for the whole combined sample, we can calculate the point biserial coefficient for BEMA item 1 to be around 0.37.

Ideally all items in a test should be highly correlated with the total score, but that is somewhat unrealistic for a test with a large number of items. The criterion widely adopted[16] for measuring the "consistency" or "reliability" of a test item is $r_{pbs} \geqslant 0.2$. Items with point biserial coefficient lower than 0.2 can still remain in a test, but there should be few such items. One way to check whether there are a majority number of items satisfying $r_{pbs} \geqslant 0.2$ is to calculate the average point biserial coefficient ($\bar{r}_{pbs}$) of all items (K) in a test:

$$\bar{r}_{pbs} = \frac{1}{K}\sum_{i=1}^{K} (r_{pbs})_i ,$$

where K is the number of items and $(r_{pbs})_i$ is the point biserial coefficient for the ith item. The average point biserial coefficient for BEMA is 0.43, which is greater than the criterion value 0.2, so BEMA items overall have fairly high correlations with the whole test.

Figure 4 provides the point biserial coefficient values for each BEMA item. As one can see, almost all items have satisfactory r_{pbs} values, indicating that almost all BEMA items are reliable and consistent. We again see that item 9 on the current in salt water is problematic.

Note that Fig. 4, plotting the point biserial coefficient, and Fig. 3, plotting the discrimination index, are quite similar. It is worth asking whether or not these two statistics actually measure the same property of an item. The answer is no; theoretically, these two statistics are different measures of an item and could in principle give different results. The item discrimination index is a measure of how powerful an item is

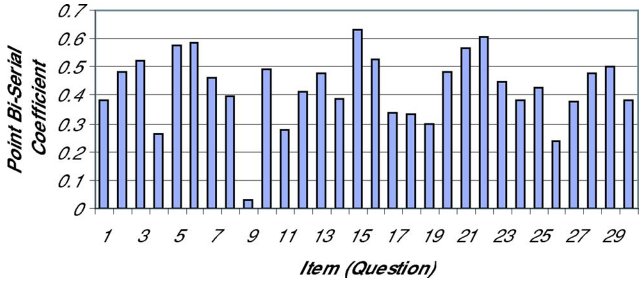

FIG. 4. (Color) BEMA item point biserial coefficient from a combined sample of 434 students.

in separating strong and weak students, while the point biserial coefficient is a measure of whether an item is consistent with the whole test. An item could have a fairly high discrimination index value, but could also show little consistency with the test as a whole. If this were the case, the item might actually be testing some topic that is different from the main subject matter of the rest of the test. On the other hand, an item could be consistent with the test as a whole (high point bi-serial correlation coefficient) but could offer little discriminatory information.

For example, suppose half of the students in a sample answer a question correctly, giving it an item difficulty index (P) of 0.5. If half of those who answer correctly (25%) have total scores in the top 25% (quartile) and the other half of them (25%) have total scores in the lower mid-25% (quartile), this item would have a fairly high point biserial coefficient ($\overline{r}_{pbs}$), but zero discrimination index (D) according to the 50%–50% method. This zero discrimination index could be avoided by switching to the 25%–25% discrimination calculation, but this method has its own extreme cases. Suppose only the top 8% of test takers get a particular item correct. Then the point bi-serial coefficient ($\overline{r}_{pbs}$) will still be fairly high, but the discrimination index (D) will be lower than 0.3. There are many other possible situations in which the two statistics may be different.

D. Kuder-Richardson reliability index

In contrast to the point biserial coefficient, which is a measure of single-item consistency or reliability, the Kuder-Richardson reliability index is a measure of the self-consistency of a whole test. If a test is administered twice (at different times) to the same sample of students, then we would expect a highly significant correlation between the two test scores, assuming that the students' performance is stable and that the test environmental conditions are the same on each occasion. The correlation coefficient between the two sets of scores is defined as the reliability index of the test. However, this approach does not actually provide a practical way of determining the reliability index of a test, since students may remember the test questions and study for the test, test conditions at different times may not be identical, etc. Another method of measuring the reliability index of a certain test is to calculate the correlation coefficient of students' scores on two parallel tests that have the same content, structure, number of items, etc., but with different question contexts. As we know, designing two truly parallel tests is very difficult, so this does not seem to be a feasible way of measuring the reliability index.

The question is whether there is any method one can employ to calculate the reliability index without administering one test twice or designing two tests, by using the information from just one test administered just once. For tests that are designed specifically for a certain knowledge domain with all items being parallel measures, the Spearman-Brown formula can be invoked to calculate the reliability index. This equation connects the reliability index with the correlation between any two parallel composites of a test. The parallel composites are subsets of the test containing the same number of components (test items). For an example, a 100-item test can have two 50-item composites, or four 25-item composites, or five 20-item composites, and so on. Based on the stipulation that the means, variance, and standard deviation of parallel measures be the same, the Spearman-Brown formula can be expressed[17]

$$r_{test} = \frac{K r_{xx}}{1 + (K-1) r_{xx}},$$

where K is now the number of parallel composites and r_{xx} is the correlation between any two parallel composites.

Kuder and Richardson further developed this idea and proposed to divide a test into its smallest components—items. Simply put, each item is regarded as a single parallel test and is assumed to have the same means, variance, and standard deviation. Two theoretical perspectives, "true and error theory" and "domain theory," can be used independently to derive the Kuder-Richardson formula from the Spearman-Brown formula. Though the two theories focus on different features of a test ("true and error theory" deals with the performance of students and "domain theory" deals with sample tests formed from a test pool), they yield the same final expression (KR-20) for calculating the reliability index of a test[18–20]:

$$r_{test} = \frac{K}{K-1}\left(1 - \frac{\sum \sigma_{xi}^2}{\sigma_x^2}\right).$$

K is once again the number of the test items, σ_{xi} is the standard deviation of the ith item score, and σ_x is the standard deviation of the total score.

This calculation takes into account the different variances of the different items, relaxing the strict assumption that all items have the same means, variance, and standard deviations. One does not have to have perfectly parallel items in a test to be able to use this formula.

For a multiple-choice test where each item is only scored as "correct" or "wrong," the above formula can be written as [19–21]

$$r_{test} = \frac{K}{K-1}\left(1 - \frac{\sum P(1-P)}{\sigma_x^2}\right).$$

P is the difficulty index of an item. This is the so-called Kuder-Richardson reliability formula KR-21. The two formulas are referred to as KR-20 and KR-21 because they

TABLE II. Summary of statistical test results.

Test statistic	Possible values	Desired values	BEMA value for CMU (N=189)	BEMA value for NCSU (N=245)
Item difficulty index P	[0, 1]	$\geqslant 0.3$	Average 0.47	Average 0.37
Item discrimination index D	[−1, 1]	$\geqslant 0.3$	Average 0.34	Average 0.32
Point biserial coefficient r_{pbs}	[−1, 1]	$\geqslant 0.2$	Average 0.45	Average 0.42
KR-21 test reliability index r_{test}	[0, 1]	$\geqslant$0.7 or $\geqslant$0.8	0.85	0.85
Ferguson's delta	[0, 1]	0.9	0.98	0.98

appeared for the first time in Kuder and Richardson's paper as the 20th and 21st formulas.

Possible values for the KR-21 reliability index fall into the range [0,1]. Different tests for various purposes have different criteria. A widely accepted criterion[22] is that tests with reliability index higher than 0.7 are reliable for group measurement and tests with reliability index higher than 0.8 are reliable for individual measurement. Under most circumstances in physics education, evaluation instruments are designed to be used to measure a large group of students, so if a certain physics test has a reliability index greater than 0.7, one can safely claim it is a reliable test.

In the BEMA analysis, we adopted Kuder-Richardson formula KR-21 to calculate the reliability index. We find the reliability index for BEMA to be 0.85, which is satisfactorily high for both group measurement and individual measurement.

E. Ferguson's delta

Ferguson's delta is another whole-test statistic. It measures the discriminatory power of an entire test by investigating how broadly the total scores of a sample are distributed over the possible range. If a test is designed and employed to discriminate among students, one would like to see a broad distribution of total scores.

The calculation of Ferguson's delta is based on the relationship between total scores of any two subjects (students). These scores may either be different or equal. If a sample has N subjects, then the total number of pairs is $N(N-1)/2$, and the total number of pairs of equal scores is

$$\sum \frac{f_i(f_i-1)}{2} = \frac{\sum f_i^2 - \sum f_i}{2}.$$

Here f_i represents the frequency (number of occurrences) of each score. The total number of pairs of different scores is $[(\Sigma f_i)^2 - \Sigma f_i^2]/2$. The number of unequal pairs will be greatest if $f_i = N/(K+1)$, where K is the number of items. Using this frequency to replace f_i in the above expressions, the number of unequal pairs becomes $[N^2 - N^2/(K+1)]/2$, which is the maximum number of unequal pairs a test can provide. The ratio between the number of unequal pairs of scores produced by a test and the maximum number such a test can yield is defined as Ferguson's delta. Accordingly, the expression of Ferguson's delta can be written as

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - N^2/(K+1)},$$

where N is the number of students in a sample, K is the number of test items, and f_i is the frequency (number of occurrence) of cases at each score.

The possible range of Ferguson's delta values is [0,1]. If a test has Ferguson's delta greater than 0.9, the test is considered to offer good discrimination.[23] Ferguson's delta for BEMA is 0.98, which is greater than 0.9.

IV. SUMMARY

The reliability and discriminatory power of the Brief Electricity & Magnetism Assessment test was evaluated by five statistical tests, three of which focus on individual items and two of which focus on the test as a whole. The results, which are summarized in Table II, indicate that BEMA is a reliable test with adequate discriminatory power.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant Nos. DUE-0320608 and DUE-0237132.

[1] R. Beichner, “Testing student interpretation of kinematics graphs,” Am. J. Phys. **62**, 750 (1994).

[2] D. Hestenes, M. Wells, and G. Swackhamer, “Force concept inventory,” Phys. Teach. **30**, 141 (1992).

[3] R. Thornton and D. Sokoloff, “Assessing student learning of Newton’s laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula,” Am. J. Phys. **66**, 338 (1998).

[4] P. Engelhardt and R. Beichner, “Students’ understanding of direct current resistive electrical circuits,” Am. J. Phys. **72**, 98 (2004).

[5] P. Engelhardt, “Examining students’ understanding of electrical circuits through multiple-choice testing and interviews,” Ph.D thesis, North Carolina State University, 1977.

[6] 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).

[7] R. Chabay and B. Sherwood, “Qualitative understanding and retention,” AAPT Announcer **27**, 96 (1997).

[8] R. Chabay and B. Sherwood, “Matter & Interactions,” in *PER-based Reform in University Physics*, edited by E. F. Redish and P. Cooney (AAPT, College Park, MD, in press).

[9] G. Aubrecht II and J. Aubrecht, “Constructing objective tests,” Am. J. Phys. **51**, 613 (1983).

[10] R. Chabay and B. Sherwood, *Matter & Interactions*, 1st ed. (John Wiley & Sons, New York, 2002), Vol. 2.

[11] P. Kline, *A Handbook of Test Construction: Introduction to psychometric design* (Methuen, London, 1986).

[12] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980).

[13] D. Doran, Ref. 12, p. 97.

[14] D. Doran, Ref. 12, p. 99.

[15] E. Ghiselli, J. Campbell, and S. Zedeck, *Measurement Theory for the Behavioral Sciences* (Freeman, San Francisco, 1981).

[16] P. Kline, Ref. 11, p. 143.

[17] E. Ghiselli, J. Campbell, and S. Zedeck, Ref. 15, p. 232.

[18] E. Ghiselli, J. Campbell, and S. Zedeck, Ref. 15, p. 254–259.

[19] G. Kuder and M. Richardson, “The theory of the estimation of psychometrika test reliability,” Psychometrika **2**, 151 (1937).

[20] E. Ghiselli, J. Campbell, and S. Zedeck, Ref. 15, p. 255.

[21] J. Bruning and B. Kintz, *Computational Handbook of Statistics*, 3rd ed. (Scott, London, 1987).

[22] D. Doran Ref. 12, p. 104.

[23] P. Kline, Ref. 11, p. 144.