


Classical test theory and item response theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism

Philip Eaton,^{*} Keith Johnson, Barrett Frank, and Shannon Willoughby
Department of Physics, Montana State University Bozeman, Montana 59715, USA

 (Received 25 July 2018; published 7 January 2019)

For proper assessment selection understanding the statistical similarities amongst assessments that measure the same, or very similar, topics is imperative. This study seeks to extend the comparative analysis between the brief electricity and magnetism assessment (BEMA) and the conceptual survey of electricity and magnetism (CSEM) presented by Pollock. This is accomplished by using large samples ($N_{\text{BEMA}} = 5368$ and $N_{\text{CSEM}} = 9905$) within classical test theory (CTT) and item response theory (IRT) frameworks. For the IRT comparison, after consideration of the conceptual content addressed in each assessment, it was assumed that each of these assessments are measuring the same student latent ability (θ), specifically a student's ability to do introductory electricity and magnetism. Via a CTT and IRT analysis it was found that both assessments are essentially equal in overall difficulty. Classical item analysis applied to 7 questions used by both assessments revealed that each assessment functions slightly differently internally. The test information curves found from IRT show that the CSEM has superior information compared to the BEMA in estimating student latent abilities for the entire range of typical latent abilities achieved by students on each assessment, $\theta \approx -2$ to $\theta \approx 3$. Information in this case is interpreted as how well a student's latent ability was estimated by an assessment as a function of latent ability. When the circuits questions are removed from the BEMA the majority of the information is lost in the $\theta \approx 0$ to $\theta \approx 2$ range. This means the circuits questions on the BEMA are information heavy for higher ability scores. So, special considerations should be made as to which assessment a study uses depending on the specific questions a researcher is attempting to answer.

DOI: [10.1103/PhysRevPhysEducRes.15.010102](https://doi.org/10.1103/PhysRevPhysEducRes.15.010102)

I. INTRODUCTION

For proper assessment selection understanding the statistical similarities among assessments that measure the same, or very similar, topics is imperative, since one instrument may be more suited to answering a specific research question over another. For example, numerous studies have been conducted to investigate the understanding of students introductory electricity and magnetism; see Refs. [1–3]. Currently, two of the most commonly used instruments for this subject are the conceptual survey of electricity and magnetism (CSEM) and the brief electricity and magnetism assessment (BEMA) [1,2]. Besides the validation statistics (like item difficulty, item discrimination, Cronbach's alpha, etc.) presented in Refs. [1,2] the authors have found no further investigations into the statistical structure of these assessments, with the exception

of Ref. [4]. This study seeks to present a large sample ($N_{\text{BEMA}} = 5368$ and $N_{\text{CSEM}} = 9905$) classical test theory (CTT) and item response theory (IRT) comparative analysis of these assessments.

In 2008, Pollock did a classical test theory comparison of the BEMA and CSEM with relatively small samples ($N < 200$) [5]. It was found that the BEMA was slightly more difficult than the CSEM with about a 5% difference in mean scores. Ultimately, it was concluded that both assessments were approximately equally useful for measuring the understanding of students within introductory electricity and magnetism courses. This study aims to corroborate the results of Pollock and to extend them by utilizing IRT. To achieve these goals this study seeks to answer the following four research questions:

RQ 1: How do the students' total scores on the BEMA and CSEM compare to one another?

RQ 2: How do nearly identical items on the BEMA and CSEM compare in their behavior from a classical item analysis perspective?

RQ 3: How do the latent ability scales from item response theory on the BEMA and CSEM compare for these samples and what is the linear translation between them, assuming one exists?

^{*} philip.eaton@montana.edu

Published by the American Physical Society under the terms of the *Creative Commons Attribution 4.0 International* license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

RQ 4: How do the assessments compare in their ability to estimate student latent abilities from item response theory as revealed by the test information curves?

Pollock found that both assessments have similar average total score percentages, as presented in Ref. [5]. This study offers support for these results by performing the same analysis using a much larger data set. To our knowledge, the other three research questions have never been addressed in the literature, and this study seeks to expand the current body of knowledge within this domain.

Before the results of the analysis are presented and the research questions answered, the conceptual content of both assessments will be discussed, as well as other limitations of this study in Sec. II. Next, the statistical methods used in this study will be detailed in Sec. III, and then the results of the analysis paired with discussion are presented in Sec. IV. Lastly, the conclusions and implications of this study are presented in Sec. V.

II. CONTENT SIMILARITY CONCERNS AND OTHER LIMITATIONS

Before the statistical comparative analysis can be performed, the conceptual content of each assessment should be compared. The BEMA contains conceptual questions pertaining to Coulomb's law, electric fields, electric potentials, electric potential energy, magnetic fields, circuits, and induction. This is similar for the CSEM with the exception of circuits. Since the BEMA has questions about circuits (questions 8, 9, 10, 11, 12, 13, and 17) that the CSEM does not, a researcher may question whether these assessments are measuring student understanding of the same content.

The topics discussed in a typical introductory electricity and magnetism course are the same ones contained in the BEMA. The BEMA can then be assumed to be measuring a more complete student understanding of their introductory electricity and magnetism course. The CSEM, since it is missing circuits, is only measuring a subset of the understanding that the BEMA measures. In a traditional introductory electricity and magnetism textbook the topic of circuits makes up roughly 1 to 2 chapters out of approximately 15 chapters [6]. From this it can be inferred that circuits in a typical introductory electricity and magnetism course makes up between 6.67% and 13.33% of the course. If we say that the BEMA measures 100% of a student's conceptual understanding from introductory courses, then the CSEM could be said to measure at minimum 86.67% of the same course. Thus, the lack of circuits on the CSEM is not an overwhelming loss of course content. As a result this study will assume that the effect of the CSEM not having circuits will be minimal on the statistical interpretation of the instruments and that both assessments are attempting to measure the same understanding of the students, otherwise known as student proficiency. Specifically, the ability of students to conceptually answer questions about electricity and magnetism from a second semester introductory

physics course taken in the first or second year of a postsecondary educational institute. Thus for IRT, both assessments will be assumed to measure the same, or a similar latent ability and that a unidimensional model can be assumed for each assessment.

To supply evidence that the CSEM's lack of circuits does not affect the statistical interpretations for this study a reduced BEMA [BEMA(R)], one without the circuits questions, was created from the full BEMA. Using the full BEMA data from Physport, the BEMA(R) was created by removing the circuits questions and recalculating the total score for the students. This assumes that the correlations between the circuits questions and the rest of the assessment is small and does not significantly affect the statistical conclusions made by this study. Since some of the statistics calculated and discussed in this study required the use of the total score students earned on the assessments (like classical item discrimination, item point biserial, and Cronbach's alpha), the BEMA(R)'s results were calculated and compared to the CSEM's when needed. This allows for a comparison of the assessments with and without the circuits questions' influence. Future studies using exploratory factor analysis or confirmatory factor analysis and multitrait item response theory are intended to test how interrelated the circuits questions on the BEMA are with the other questions on the assessment.

There are other limitations to this study that should be mentioned for the results to be interpreted in the correct frame of mind. First, the samples used for the BEMA and CSEM are assumed to be independent from each other, but are sampled from the same population; this population being second semester algebra or calculus-based introductory physics students attending all types of postsecondary educational institutes from across the United States. Since the sample that took the BEMA is not identical to the one that took the CSEM, any differences found between the assessments in the CTT analysis will be a combination of sample differences and assessment differences. This is due to CTT's direct dependence on the sample being analyzed. Without having the same sample take both assessments, assuming one administration does not affect the other, identifying how much of the difference in the BEMA and CSEM is due to sample differences or assessment differences is not possible. This issue is smaller for the results of IRT since item parameters are assumed to be a product of the questions themselves and not of the sample. It is true that the scale the assessments are put on in IRT is a product of the sample being analyzed at the time. However, scale linking procedures can be used to put different samples for the same assessment onto the same scale, where they can then be compared. A similar linking of scale can be done for two assessments that are not identical, but do share some common items, which is the case for the BEMA and CSEM. This ability to link scales reduces the impact of having different samples for the two different assessments, and thus

makes any conclusions about similarities and differences more applicable to the assessments and less about the samples.

Lastly, recall that the data used in this study were a mixture of algebra and calculus-based classes, as well as a mixture of teaching styles. The results of this study may be affected due to the heterogeneity of the data used. A more homogeneous data set, say one that comes from a calculus-based flipped-classroom course only, may have different results from those presented in this study. Because of the novelty of the content discussed in electricity and magnetism for students learning it for the first time, it is assumed that the results of this study will change very little between algebra and calculus-based classes, as well as across different teaching styles. However, studies to look into these possible differences are encouraged by the authors.

III. METHODOLOGY

The methodology section will begin with a discussion of how the data were obtained. Following that is an explanation of all the statistical tools used for this study. If the reader has a strong understanding of classical test theory and/or item response theory they are encouraged to begin at Sec. IV. It should be made clear that the CTT analysis done in this study is motivated by the one done in Ref. [5]. This will help supply evidence for or against the conclusions made in that analysis. The IRT analysis was done to supply more evidence for any conclusions derived from a CTT analysis.

A. Data

The data used in this study were received from Physport and contained both pre- and postinstruction student responses to the BEMA and CSEM [7]. All of the analyses within this study were done on postinstruction student responses due to the pretest results giving Cronbach's alpha values of 0.60 (± 0.01) for the BEMA and 0.68 (± 0.01) for the CSEM. These values for Cronbach's alpha, as is explained in Sec. III B, indicate that these assessments do not reliably (i.e., precisely) give students a total score for preinstruction data. As such, any statements made about the pretest data will be accompanied with a large amount of error. To avoid this, the pretest data will not be analyzed for this comparative study.

Originally the postinstruction BEMA and CSEM data sets contained 5918 and 10410 student responses, respectively. For each of these data sets any students whose responses contained a blank entry were removed from the data set. This left the BEMA data set with 5368 student responses (a loss of 9.23%) and the CSEM set with 9905 student responses (a loss of 4.85%).

When grading the BEMA for the CTT calculation some questions require careful consideration; solutions for the BEMA and CSEM can be requested from Physport.org.

For example, questions 28 and 29 on the BEMA are graded together as a single question. In order for a student to get this paired question correct they must answer both individual questions correctly. So that the assessments could be compared in a fair manner when performing the CTT analysis, the grading scheme for the BEMA was adopted for any shared items when grading the CSEM. For example, questions 1, 2, and 3 on the BEMA are identical to 3, 4, and 5 on the CSEM. So, these questions were graded using the criteria demanded by the BEMA's grading rubric and not the CSEM's. This grading of both assessments is in line with the way the assessments were graded in Ref. [5].

Grading the BEMA in the manner indicated by its creators introduces linked questions in the grading scheme, which is something not allowed in the assumptions of IRT. This grading criteria causes the probability of answering one of the linked questions correctly to depend on how a student responded to another linked question(s). This nonindependence of items on the BEMA breaks one of the assumptions of IRT discussed in Sec. III. Thus, when using IRT, the questions were graded as strictly correct or incorrect (coded as 1 or 0, respectively) to avoid dependencies between items on the BEMA. This preserves the independence of item assumption that is needed to perform IRT on the data.

So, when CTT was performed the data were graded following the BEMA's grading criteria. When IRT was performed the BEMA's grading criteria were abandoned and questions were graded strictly as right or wrong. Correct answers in either grading criteria were coded as a 1 and incorrect responses were coded as a 0. Further, for IRT a tetrachoric correlation was used since the data was graded dichotomously.

B. Classical test theory

Classical test theory is a psychometric theory that assumes the observed score a student earns on an assessment is a combination of their true score and error from the assessment: $S_{\text{obs}} = S_{\text{true}} + \epsilon$. CTT is concerned with the relationship between these three variables and how they relate to test reliability. However, this kind of analysis offers no insight on how the items (the questions on an assessment) are functioning individually. Classical item analysis is concerned with a number of item specific statistics such as classical item difficulty, classical item discrimination, and the item point biserial, all of which are explained in detail in the following subsections.

To statistically compare some of these statistics, specifically the classical item difficulty and discrimination, the spread of the statistic needs to be calculated. Since these two item statistics are found through counting student responses of the entire data set, they do not come with any associated errors or measures of variance. To find a standard deviation for the statistics, bootstrapping methods were used. Generating 10 000 uniformly sampled classes

TABLE I. Test statistics for the BEMA and CSEM post-test samples. BEMA(R) is the BEMA with the circuits questions removed from consideration, and α stands for Cronbach's alpha reliability statistic.

Test	No. students	Mean	St. Dev.	Skew	Kurt.	α
BEMA	5368	0.487	0.184	0.071	-0.686	0.818
BEMA(R)	5368	0.500	0.199	0.005	-0.787	0.800
CSEM	9905	0.446	0.185	0.453	-0.422	0.825

of 250 students from the ranked data for each assessment, all of the item level statistics' standard deviations were able to be determined. The distribution of these statistics was normal, so a Cohen's d could be used to calculate the effect size between measures. Cohen's d is a measure of effect size that is equal to the difference in the measurement means divided by a pooled standard deviation. Specific information on how this statistic is calculated can be found in Ref. [8].

1. Overall test statistics

To look at the overall performance of each of the assessments the mean, standard deviation, skew, and kurtosis were calculated. The test statistics for the BEMA, BEMA(R), and CSEM can be found in Table I. To support the assumption that the students' total scores could be well represented by a normal univariate distribution, the skew and kurtosis should take values between -2 and 2 [9]. Since the values of the assessments' skew and kurtosis are well within the recommended values, this study will assume that the data come from a normal univariate distribution, which agrees with what is seen visually in the histograms in Fig. 1. Although the CSEM does have a relatively large skew, it does fall into the recommended range of skew values for a normal distribution.

Another important overall test statistic for verifying that an assessment is operating properly is the reliability

of the assessment. Statistics that measure the reliability can be thought of as measuring the consistency that an assessment has for giving a student a particular score. Meaning, if a student could take a test multiple times while retaining no memory of their previous attempts, then an assessment with perfect reliability will give that student exactly the same score for every attempt the student makes. This study used the commonly calculated Cronbach's alpha to measure the internal reliability of the assessment; there are numerous other statistics that measure the internal reliability, such as the KR20 or KR21 [10]. Internal reliability is a measure of how well the test performed in one administration and not necessarily for test or retest administrations.

To calculate the Cronbach's alpha the following expression was used:

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^K \sigma_{Y_i}^2}{\sigma_X^2} \right),$$

where K is the total number of items on the assessment, $\sigma_{Y_i}^2$ is the variance of the students' performance on item i , and σ_X^2 is the variance of the students' total scores. Acceptable values for Cronbach's alpha range from 0.7 to greater than 0.9, however a value less than 0.8 may indicate that the assessment is suitable for group measurements but not for individual students [11].

2. Classical item difficulty

Item difficulty (called classical difficulty so as not to be confused with IRT item difficulty) measures the proportion of students who answered the item being considered in a correct manner. To calculate the classical item difficulty the following equation was used:

$$P_i = \frac{N_i}{N},$$

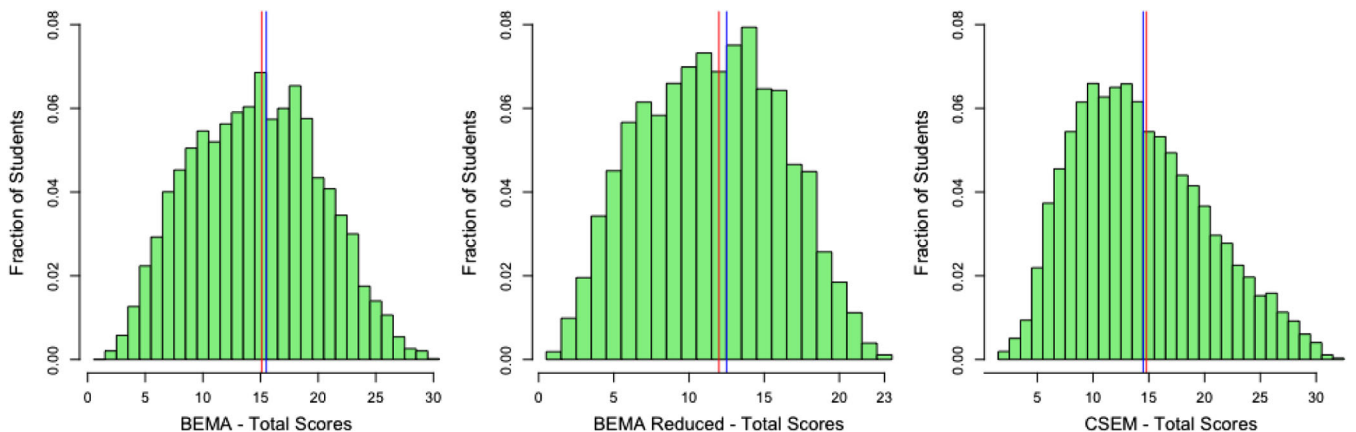


FIG. 1. Histograms of the fraction of the total number of students for all possible total scores. For the BEMA there were no scores below 3 and for the CSEM there were no scores below 2. The vertical red line indicates the sample mean and blue is for the median.

where P_i is the classical difficulty for item i , N_i is the number of students who got item i correct, and N is the total number of students who attempted the item.

This statistic is traditionally presented as a decimal value that ranges from 0 to 1. A classical difficulty of 0 indicates that none of the students who attempted the item answered in the correct manner, and 1 means that all of the students who attempted the item got the item correct. Items are said to be “very difficult” if their difficulties are below 0.35, and ones with values above 0.85 are referred to as “very easy” [11]. Since a value closer to 1 indicates an item which is more frequently answered correctly, this statistic could be thought of as the “easiness index,” however the current nomenclature will be upheld. It has been suggested that “ideal” classical difficulties should be between 0.40 and 0.60, or 0.20 and 0.80 [11]. This study used cutoff values of 0.20 and 0.80 to identify acceptable classical item difficulties.

3. Classical item discrimination

Item discrimination (called classical discrimination so as to not be confused with IRT item discrimination), otherwise known as the validity index or item power index, is a comparative index between how the “high” students did on an item versus “low” students [11]. This is done by calculating the classical difficulty using only the upper p percentile students and the lower p percentile students for an item. Then the difference in classical difficulties is taken for the upper and lower p percentile students:

$$D_i(p\%) = P_i(\text{upper } p\%) - P_i(\text{lower } p\%),$$

where $D_i(p\%)$ is the classical item discrimination for item i using the upper and lower $p\%$ of the students based on their total scores on the assessment, $P_i(\text{upper } p\%)$ is the item difficulty for the students in the $(100 - p)$ th percentile and above, and similarly for $P_i(\text{lower } p\%)$. This study used $p = 27$, so the high students were in the 73rd–100th percentile and the low students were from the 0th to the 27th percentile. The p has been seen in the literature to take on a few different values: $p = 50$ in Ref. [11], and $p = 25$ in Ref. [12]. This study used 27% since it appears to maximize the differences between the high and low students [13]. The bar plots for both exams’ item discriminations can be found in Fig. 2.

Classical discrimination can take values from -1 to 1 . Negative values indicated that the low scoring students did better on the item than the high scoring students, which should not be the case. This is generally an indication that there is a negative statement in the problem statement not reflected in the response options, or that the key is not identifying the correct response for the solution. There are no maximum cutoff values for this statistic, but there have been suggested minimum cutoffs at 0.2, 0.3, or 0.4 [11].

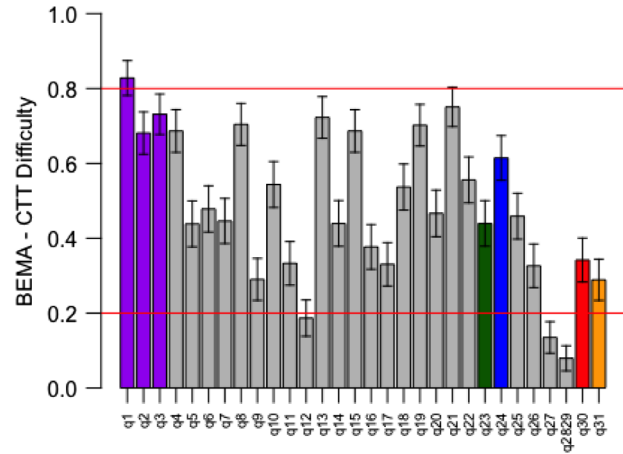


FIG. 2. The error bars are ± 2 standard deviations from 10 000 randomly sampled classes of 250 students from the full sample for the BEMA. The horizontal red lines indicate an ideal range of difficulties for well-functioning questions. The colored bars are the 7 identical items between the BEMA and CSEM.

This study used a minimum cutoff of 0.3, the mean of the suggested cutoff values.

4. Item point biserial

The item point biserial is the correlation between the responses given by the students on a given item (1’s and 0’s) and the total scores achieved by the students. The point biserial can be calculated for each item using

$$r_{pb,i} = \frac{\bar{X}_{1,i} - \bar{X}_{0,i}}{\sigma_X} \sqrt{\frac{N_{1,i}N_{0,i}}{(N_{1,i} + N_{0,i})^2}},$$

where $\bar{X}_{1,i}$ and $\bar{X}_{0,i}$ are the test averages for the students who got item i correct and incorrect, respectively, σ_X is the standard deviation of all of the total scores for the assessment, and $N_{1,i}$ and $N_{0,i}$ are the numbers of students who got item i correct and incorrect, respectively.

Because this statistic is a correlation, it is restricted to values between -1 and 1 . However, one would expect that the better a student does on the assessment the greater the likelihood that they will have answered any given question correct. As a result any items with negative point biserial should be put under scrutiny, similar to the classical discrimination. The generally accepted values for this statistic are above 0.2 [14].

C. Item response theory

Item response theory (IRT) is a model based theory that attempts to model the probability that a student of a certain latent ability will get an item correct given the parameters of the item. Developed beginning in the 1950s, IRT is built on three fundamental assumptions:

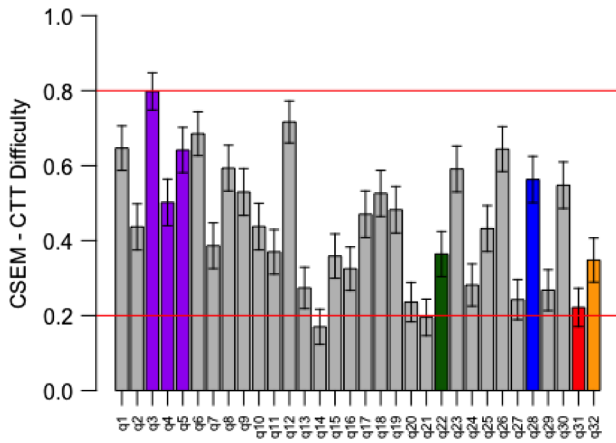


FIG. 3. The error bars are ± 2 standard deviations from 10 000 randomly sampled classes of 250 students from the full sample for the CSEM. The horizontal red lines indicate the range of ideal difficulties for well-functioning questions. The colored bars are the 7 identical items between the BEMA and CSEM.

- (1) There exists a single latent ability, θ , of a student for the subject being tested by an assessment.
- (2) Items on an assessment are locally independent, meaning that the probability of answering item i and j both correctly can be given as $P(i \& j) = P(i)P(j)$, where P is a probability.
- (3) The probability for a correct response from a student on an item can be estimated using the item response function and the student’s latent ability score.

There is a multitrait extension IRT called multitrait item response theory (MIRT) that changes the first assumption to multiple traits, rather than only using one. The distinction of IRT as unidimensional as opposed to MIRT being multi-dimensional is common throughout IRT and MIRT literature.

To verify the unidimensionality of an assessment the correlations amongst the items need to be analyzed. If it is found upon an eigenvector reduction of the item correlation matrix that one eigenvector is dominant (i.e., has an eigenvalue an order of magnitude larger than the next largest one) then the unidimensional assumption can be used [15]. One method to check unidimensionality is to examine the Scree plots for both assessments, see Fig. 3. This method is preferred over other methods due to its simplicity. A Scree plot displays the eigenvalues associated with the eigenvectors in decreasing order. Since these are the eigenvalues of a correlation matrix, they describe how much of the assessment’s variance is explained by each of the factors independently. From the Scree plots it can be seen that both assessments have a single dominant factor and as a result can be treated as being unidimensional. As explained in Ref. [16], if the unidimensional assumption holds then local item independence also holds as a result. Further, it should be noted that the way in which the BEMA, and, consequently, the CSEM, is graded appears to

remove any conflicts with the item independence assumption of item response theory.

To support the unidimensional assumption of the assessments analyzed in this study the goodness-of-fit indexes were calculated for the unidimensional models found using a 2-parameter logistic (2PL) model (discussed in detail below). The fit statistics used to support the fit of the unidimensional models will be the comparative fit index (CFI), the Tucker-Lewis index (TLI), the upper 95% confidence interval root mean square error of approximation (RMSEA95%), and the standardized root mean square residual (SRMR). Detailed explanations for these fit indexes can be found in Ref. [17], as well as further resources for the acceptable fit values used in this study. The following are the “good fit” criteria for each of the fit indexes: for the CFI and TLI values above 0.9 is said to be acceptable, and both the RMSEA95% and the SRMR values should be below 0.08 for an acceptable fit [18]. The CSEM had the following fit indexes for its unidimensional model: CFI = 0.921, TLI = 0.915, RMSEA95% = 0.042, and SRMR = 0.034; all of which are within the acceptable ranges. When all of the questions on the BEMA were included it was found to not fit well with the unidimensional model, and after investigation it was found that the sources of this misfit were questions 14 and 29. Once these items were removed the BEMA had the following fit indexes: CFI = 0.921, TLI = 0.914, RMSEA95% = 0.047, and SRMR = 0.038; all of which are now within the acceptable ranges of fit. The misfit of the BEMA, when all of its questions are retained, could be an indication that a multidimensional model would be better suited for this assessment, or it could be a result of the sample itself. The authors plan on a future study in which MIRT will be used to model each assessment, and comparisons made between these models. This study, however, will assume the assessments are unidimensional and will be carried out as such.

There are many models used for the item response function that relate the probability of answering an item correctly to the latent ability of the student taking the assessment. Some of the most commonly used models are the Rasch model and the closely related 1-parameter logistic model (1PL), the 2-parameter logistic model (2PL), and the 3-parameter logistics model (3PL). For this study the 2PL model was used to analyze both assessments, and can be written in the following manner:

$$\pi_i(\theta) = \pi_i(\theta|\alpha_i, \delta_i) = \frac{1}{1 + e^{-D\alpha_i(\theta - \delta_i)}},$$

where $\pi_i(\theta)$ is the probability that a student with a latent ability of θ will get item i correct, and α_i , and δ_i are the item discrimination and difficulty for item i , respectively. Lastly, D is a number, generally 1.702, which adjusts the scale of the student ability axis to more closely correspond to a normal ogive scale. The item parameters are estimated using parameter estimation techniques that are described

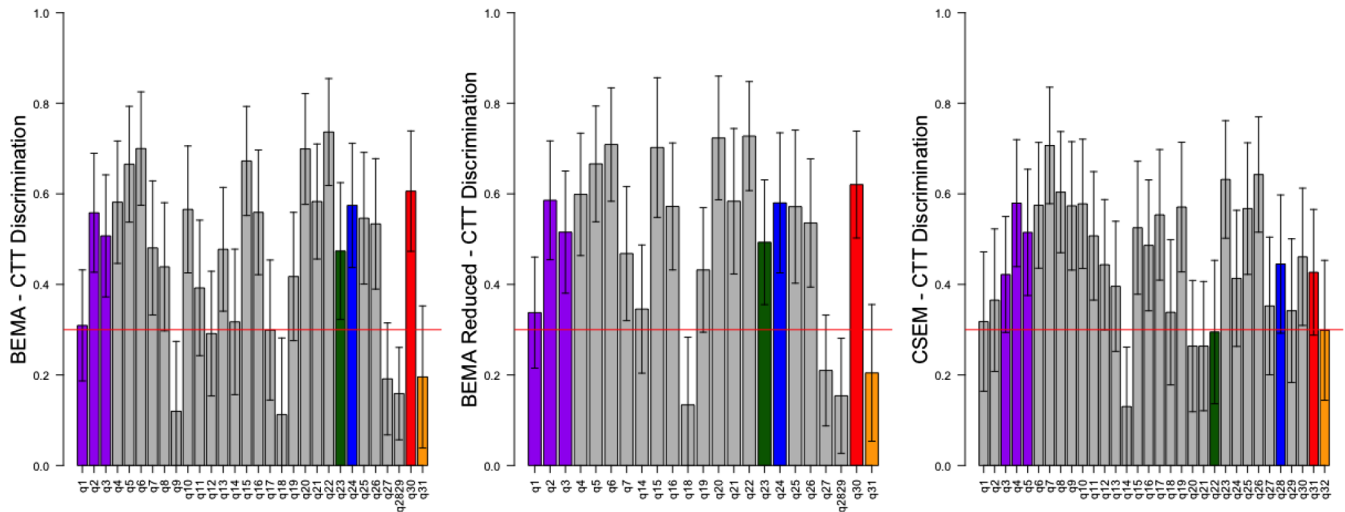


FIG. 4. The error bars are ± 2 standard deviations from 10 000 randomly sampled classes of 250 students from the full sample for both the BEMA and the CSEM. The horizontal red line indicate the minimum appropriate discrimination for well-functioning questions. The upper and lower 27% was used to calculate the classical discrimination. The colored bars are the 7 identical items between the BEMA and CSEM.

in part in Refs. [19,20]; this study used the open source R based software “mirt” for all item parameter estimations for the dichotomously graded assessments [19]. This study used the 2PL model rather than the 3PL model since instabilities in the parameter estimations of the 3PL model were detected. These instabilities are known to exist and suggest the data are insufficient or not large enough to estimate the 3PL parameters adequately [21].

1. Item difficulty

Item difficulty, similar to classical difficulty, relates to the probability students will answer the item correctly. The “harder” an item is the larger the item difficulty will be, and vice versa. Thus, an item with a difficulty of 2 is harder than an item with a difficulty of -2 . In the 2PL model the item difficulty parameter δ_i indicates the latent ability location of the inflection point for the item response function and is also where the item response function has a probability of 50%. Through some manipulations it can be shown that the inflection point occurs at $(\theta, \pi_i(\theta)) = (\delta_i, \frac{1}{2})$, as indicated in Fig. 4.

2. Item discrimination

An item’s ability to discriminate amongst students refers to its ability to identify students with latent abilities above and below the item’s difficulty parameter. The larger an item’s discrimination the better it is at placing a student above or below a specific latent ability score. Functionally, the larger the discrimination, the more like a step function the item response curve will become. For items that are meant to be used for ranking students (e.g., the GRE), item discriminations should be relatively large for all of the items on the assessment. However, even with this being the case there are no cutoff values for IRT item discrimination.

Graphically, item discrimination α_i is related to the slope of the item response function at its inflection point. It can be shown by direct calculation that the slope at the inflection point comes out to be $D\alpha_i/4$, which can be seen in Fig. 4.

3. Information functions and scale comparison

From the probability function the likelihood can be developed as well as the uncertainty in the estimation of the latent ability. Taking the reciprocal of this uncertainty gives the item information function. For the 2PL model the item information function is given mathematically as

$$I_i(\theta) = D^2\alpha_i^2\pi_i(\theta)[1 - \pi_i(\theta)],$$

and an example of a plot for an item information curve can be found in Fig. 4. Because of the local item invariance assumption, the item information functions can be added together to give test information function [20]. The item information function describes how well a single item can estimate a student’s latent ability and the test information function describes the how well the entire test (i.e., assessment) can estimate student ability scores. Information in this case is interpreted as how well a student’s latent ability was estimated by an assessment as a function of latent ability. Latent abilities in a region with high information are thus better estimated than those in regions of lower information.

Further, it can be shown that most of the item information curves for these assessments will have their maxima centered at $\theta = \delta_i$. If many of the items on these assessments have the same difficulty values, then the test information curve will also have a peak close to $\theta = \delta_i$. Assessments that do a “good job” of estimating all students’ abilities will, more often than not, be ones that have their item difficulties evenly spread out over the ability space. As an analogy, think of identifying one’s location on a ladder via

unevenly spaced rungs. The more evenly spaced the rungs are the better locations can be identified along the entire length of the ladder. Whereas, if the ladder’s rungs were grouped up towards the bottom then locations at the bottom of the ladder could be very accurately described, whereas the top of the ladder would be left quite ambiguous. The rungs of this ladder are similar to the location of the items along the latent ability axis. The more grouped up the questions are the better the student abilities in that region will be estimated compared to regions where very few items exist.

Estimation of the item parameters allows one to create the test information curves for each of the assessments. When the 2PL item parameters are estimated, as well as the other models mentioned, they are put onto a scale that is built based on the latent ability scores estimated for the students who took the assessment. For instance, the scale used along the horizontal axis in Fig. 4 uses a spacing roughly equal to 1 standard deviation in the estimated students’ abilities. The model used in this study is translationally invariant along the latent ability axis. This can be seen by looking at the relationship between θ and δ_i : $\theta - \delta_i$. If both parameters are shifted by the same amount, then the difference will remain the same. So, the definition of “0” along the latent ability axis is usually taken to be the average of the estimated student abilities (student centered) or the average of the item difficulties (test centered). For this study, since the purpose is to compare the IRT results, the scales of each assessment will need to be linked through a linking procedure.

There are many linking procedures that place both exams onto the same scale. For example, the fixed parameter with score transformation procedure, and the mean-mean or mean-sigma procedure are commonly used linking procedures [22–24]. For these assessments it was found that the mean-mean procedure produced the most acceptable

transformation between the two scales. The details of the mean-mean transformation, as well as many other linking procedures, and the tests for identifying the “best” transformation can be found in Ref. [22].

Now that the item parameters for the BEMA are known in the CSEM scale, using the mean-mean transformation, these item parameters can be used to estimate the student abilities of the BEMA on the CSEM’s scale. These scales are related through the linear transformation: $\theta^* = S\theta + C$ where for mean-mean linking

$$S = \mu(\text{BEMA's } \alpha) / \mu(\text{CSEM's } \alpha),$$

$$C = \mu(\text{CSEM's } \delta) - S \cdot \mu(\text{BEMA's } \delta),$$

from Ref. [22]. After following this procedure the values for S and C were found to be $S = 1.166$ and $C = 0.224$. It should be noted that this linking is data driven and thus will change when repeated for different data sets. Further, it must be assumed that the data sets are pulled from the same population when performing this scale linking. Considering these numbers and the linear transformation, it can be seen that the original CSEM and BEMA latent ability scales are quite close in spacing ($S \approx 1$). This should not be surprising since the standard deviations for the total test scores in Table I are almost identical. The scale linking transforms the estimated parameters for the BEMA to the parameter space of the CSEM. Thus, the goodness of fit of the new model against the original data must be recalculated to ensure the model is still appropriate. The goodness-of-fit indexes for the BEMA’s fit to the unidimensional model using this scale transformation comes out to be CFI = 0.921, TLI = 0.927, RMSEA95% = 0.044, SRMR = 0.48; all of which are within the accepted ranges of values. The test information curves for the BEMA, BEMA(R) and CSEM, plotted in the CSEM scale, can be found in Fig. 5.

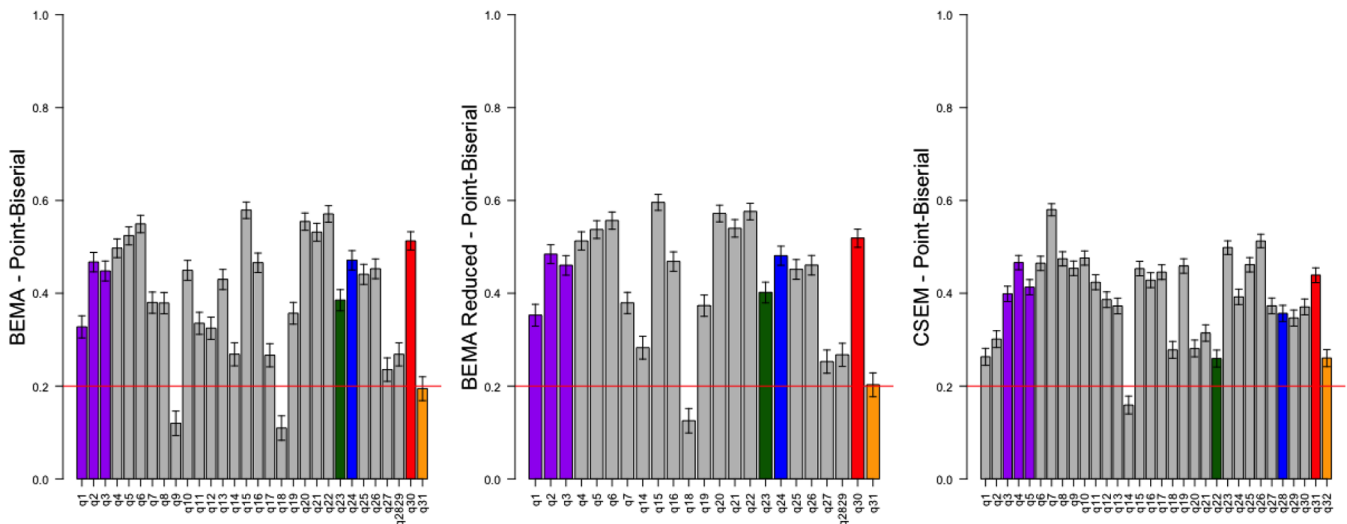


FIG. 5. The point biserial for each of the items in the BEMA and CSEM. The red line is at 0.2, which is an accepted minimum for item point-biserial values. There error bars represents the 95% confidence interval for each of the bins, respectively. The colored bars are the 7 identical items between the BEMA and CSEM.

Now that both assessments can be placed onto the same scale, the test information curves and item response functions can be plotted and compared, which will be presented and discussed in Sec. IV B.

IV. RESULTS AND DISCUSSION

The results of the analysis performed on the BEMA and CSEM will be discussed below following a similar section outline as in Sec. III.

A. Classical test theory

1. Overall test statistics

The distributions of the total scores on all of the assessments have means of about 50% and standard deviations of about 20%. Considering the suggestion from Ref. [11], that the ideal classical item difficulty is 0.50, implies the ideal average test score for an assessment is 50%. This suggestion comes from the desire to avoid ceiling and floor effects due to the rigid boundaries placed on the available total scores. If the average on an assessment were too high, say 75%, then the group of students above that score have a good chance of bunching up at the 100% mark, and vice versa for an average that is too low. By having the test mean close to 50% the assessment will use the full range of total scores in the most effective manner possible to distinguish students from one another. So, all of these assessments are functioning ideally from this consideration.

When comparing the BEMA and CSEM using a two-tailed *t* test it was found that the difference between the means was statistically significant with a *p* value of 0.001 05. However, when the effect size was calculated the difference was found to be negligible with a Cohen's *d* of 0.054. Thus, the detection of significant difference was due to the large samples used, and not because the difference in the means is large. As a whole, when considering the mean and standard deviations of these assessments, these two tests were found to be the same in overall difficulty.

All forms of the assessments were found to have satisfactory reliabilities with Cronbach's alphas all greater than 0.8; the specific values can be found in Table I.

2. Classical item difficulty

Bar plots of the classical difficulties for the BEMA and CSEM can be found in Figs. 6 and 7, and values can be found in Tables IV and V located in Appendix. Using cutoff values of 0.20 and 0.80 it can be seen that the CSEM and BEMA contain a few questions that fall outside of this range. Comparing these results to Refs. [1,2], it is found that the classical item difficulties found in this study for the BEMA and CSEM are similar to the previously found classical difficulties.

Of the shared questions only question 1 on the BEMA (question 3 on the CSEM) falls above the 0.8 cutoff value, which was also seen in Refs. [1,2]. This question asks about the charge dependence of Coulomb's law and what happens

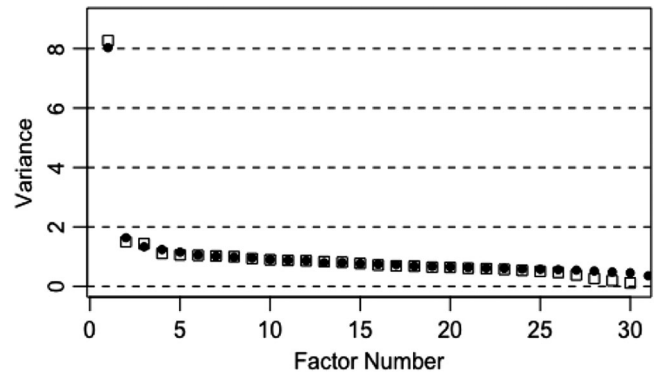


FIG. 6. Scree plots for each of the assessments. The hollow boxes represent the BEMA's eigenvalues and the solid dots represent the CSEM. These plots support the assumption that these assessments can be treated as unidimensional for an IRT model.

to the magnitude of the electric force when the magnitude of a charge is changed. Considering the content of the question, it is unsurprising that this question was found to be easy for the student responses that make up the data.

None of the shared questions had an item difficulty below the 0.20 cutoff. Questions 12, 27, and the paired 28 and 29 on the BEMA and questions 14 and 21 on the CSEM were found to be quite difficult for these students. Looking at the numerical values for these items in Tables IV and V, it can be seen that question 21 on the CSEM only just falls below the cutoff value. Thus, the majority of the questions that appear on both assessments fall within the accepted range of values for classical difficulty.

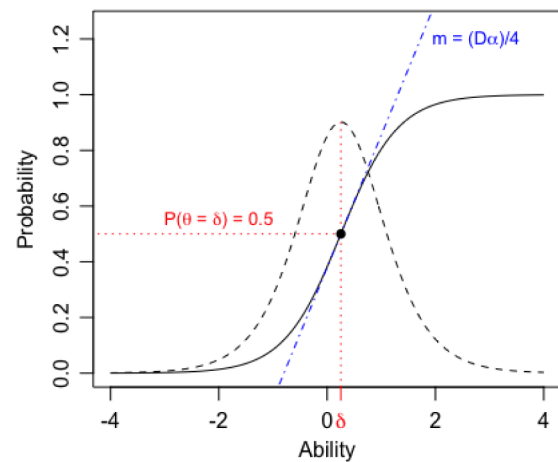


FIG. 7. An example of an item characteristic curve (solid, black) and its associated item information curve (dashed, black) using the 2PL model. When the student ability is equal to the item difficulty ($\theta = \delta$) the probability of the student answering the questions correctly is 50%; this is also the location of the maximum for the item information curve. The slope (*m*) of the tangent line (dot-dashed, blue) at the 50% mark is proportional to the item discrimination (α). *D* is a constant generally set so 1.702 so that the ability axes better lines up with spacings that are equal to a standard deviation of student ability.

TABLE II. Comparison between classical difficulties, discriminations, and point biserials for the questions on the BEMA and CSEM that are the same. The numbers to the left of the “/” are for the BEMA and to the right are for the CSEM. The entries marked with a * were found to be statistically different. All other entries were found to not be statistically different with p values < 0.001 . C.d stands for the absolute value of Cohen’s d .

BEMA/CSEM	Difficulty	C.d.	Discrimination	C.d.	r_{pb}	C.d.
q1/q3	0.83/0.80*	1.26*	0.30/0.41*	1.79*	0.33/0.40*	1.50*
q2/q4	0.68/0.50*	6.02*	0.55/0.58	0.359	0.47/0.47	0.03
q3/q5	0.73/0.64*	3.08*	0.50/0.51	0.108	0.45/0.41*	0.72*
q23/q22	0.44/0.36*	2.46*	0.48/0.29*	2.34*	0.38/0.26*	2.20*
q24/q28	0.61/0.56*	1.70*	0.57/0.44*	1.69*	0.47/0.36*	2.27*
q30/q31	0.34/0.22*	4.33*	0.61/0.43*	2.67*	0.51/0.44*	1.42*
q31/q32	0.29/0.35*	2.05*	0.19/0.29*	1.30*	0.20/0.26*	1.10*

Comparing the questions on the assessments that are the same (see Table II) reveals that many of these items have similar values. Using a z test of proportions, since item difficulties are by definition proportions of the number correct versus the number attempted, it was found that all of the item difficulties were found to be significantly different for the same items between both assessments. This is likely due to the large sample sizes used in this study. To understand how different these difficulties actually are the Cohen’s d was calculated for all 7 questions shared by both assessments. All of these Cohen’s d values were found to be large, even for q1/q3 when the difference in difficulties was observed to be only 0.03. This appears to be too harsh of a criteria, and a difference of 0.1 (10%) is suggested for a significant difference [25]. From this, only questions 2/4 and 30/31 have significant differences in difficulty. Since classical difficulty makes no reference to the total score on the assessment, there is no need to consider the BEMA(R) since the classical difficulties will not change.

3. Classical item discrimination

Many of the items on each of the assessments had classical discriminations that fell below the 0.3 cutoff value. From the BEMA these questions were 9, 12, 18, 27, 28 and 29, and 31, of which only question 9 was about circuits. For the CSEM questions 14, 20, 21, and 22 had classical discriminations below the cutoff value. Because of slight differences in how the classical discrimination was calculated between this study and Ref. [1] and the factor of 12 difference in sample size, this study did not get nearly as many malfunctioning items as was previously found. All of the questions on the BEMA that were found to have classical discriminations below 0.3 were found to have similar issues in previous examinations, with the exception of question 12 from this study’s results [1]. The item discriminations for the CSEM have not been reported as of the time of this study, so no comparisons to other results could be made.

Because of the relationship between classical difficulty and the maximum value possible for an item’s classical discrimination, some item’s poor discriminations can be explained due to their poor difficulty scores [11]. Questions 12, 27, and 28 and 29 on the BEMA and 14, 20, and 21 on

the CSEM all have values around, or below 0.2. Thus, the possible range of classical discriminations available to these items is severely limited, and may be the cause for their low discrimination values. This leaves items 9 and 18 on the BEMA and 22 on the CSEM without a reason for their poor discrimination performance. Reasons for these items’ performance may be revealed through student interviews, or distractor-level analysis.

Comparing the same items between the BEMA and the CSEM (see Table II) reveals that some of the question discriminations differ greatly. Since classical discrimination can be written as a proportion, the z test of proportions was used to compare the items’ results. It was found that all of the items, besides 2/4 and 3/5 on the BEMA/CSEM, had significantly different discriminations, and similar results were found when comparing the BEMA(R) and the CSEM; see Table III.

4. Item point biserial

Since item point biserial and classical discrimination are both measures of an item’s ability to distinguish between high and low students their results should be qualitatively similar, but not necessarily identical. This can be seen by considering Fig. 8. A fundamental difference between the two is that the item point biserial is more directly dependent

TABLE III. Comparison of classical discrimination and point-biserial values for questions on the BEMA(R) and CSEM that are the same. The numbers of the left of the “/” are for the BEMA(R) and to the right are for the CSEM. The entries marked with a * were found to be statistically different. All other entries were found to be statistically different with p values < 0.001 . C.d stands for the absolute value of Cohen’s d .

BEMA(R)/CSEM	Disc.	C.d.	r_{pb}	C.d.
q1/q3	0.33/0.41*	1.35*	0.35/0.40*	0.97*
q2/q4	0.58/0.58	0.04	0.48/0.47*	0.39*
q3/q5	0.51/0.51	0.01	0.46/0.41*	0.97*
q23/q22	0.50/0.29*	2.61*	0.40/0.26*	2.50*
q24/q28	0.57/0.44*	1.77*	0.48/0.36*	2.46*
q30/q31	0.62/0.43*	2.87*	0.52/0.44*	1.54*
q31/q32	0.20/0.29*	1.16*	0.20/0.26*	0.96*

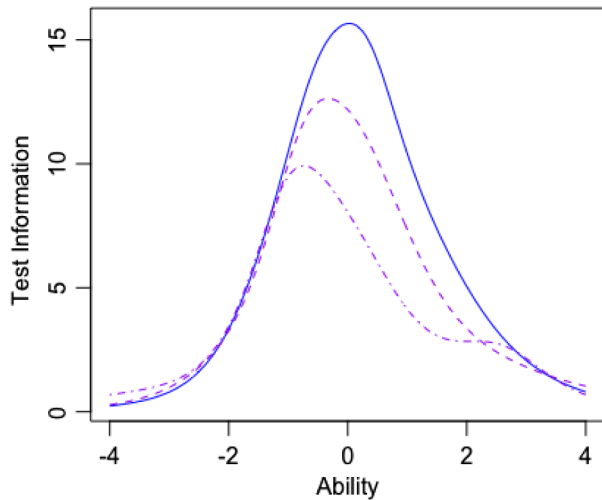


FIG. 8. Item response theory test information curves for the CSEM (blue, solid), BEMA (purple, dashed), and BEMA(R) (purple, dot-dashed).

on the classical difficulty as compared to the classical discrimination, whose maximum depends on the difficulty and not the value itself. This lends evidence that questions 12, 27, and 28 and 29 on the BEMA and 14 and 21 on the CSEM are malfunctioning due to their low classical difficulties, and that 9 and 18 on the BEMA are having issues for other reasons.

Comparing the results for the BEMA to previously published results the item point biserials presented here differ for many of the questions that are the same between the assessments [1]. This could be due to a number of differences between these two studies, such as (i) only calculus-based course data was used in the previous study and a mixture of calculus and algebra-based data was used in this study, or (ii) the difference in sample sizes. This does warrant a hint of caution when trying to generalize the results of this study to purely calculus or algebra-based classes, and does suggest further research is needed.

To perform significance testing between the CSEM, BEMA, and BEMA(R)'s point-biserial values for the shared questions, 10 000 classes of 250 uniformly sampled students from the ranked data were used. The item point biserials for each of the items for every class were calculated. The resulting distributions of the item point-biserial values were found to satisfy a univariate normal distribution assumption for all three assessments ($|\text{skew}| < 2$ and $|\text{kurtosis}| < 2$). From the assumed normal distributions for all of the items a t test was used to compare the items between the CSEM and BEMA as well as the CSEM and BEMA(R).

Between the BEMA and the CSEM it was found that all of the point-biserial values were significantly different for shared questions with the exception of question 2/4; see Table II. When removing the circuits questions from the BEMA and again comparing to the CSEM it was found that

all of the point-biserial values were significantly different. This suggests that the circuits questions effect other items' relations to the total score. So, if a factor analysis were to be done, correlations between the circuits questions and the others should be expected to be nonzero.

5. Conclusions of classical test theory

Since CTT is explicitly sample dependent any differences found between the assessments could be a result of differences in the samples. This means any conclusions about differences in the assessments made using CTT cannot be separated from potential sample differences, and thus doing assessment comparisons in a CTT framework becomes almost impossible and nongeneralizable.

Item response theory does not suffer from such issues. Since IRT is model driven, and the results of each assessment can be placed onto the same scale, comparisons can be made without significant sample interference. These conclusions, discussed below, give a more reliable comparison of the assessments compared to CTT. This is because the item parameters, item characteristic curves, and item or test information curves are assumed to be independent of the student results. The only influence the student responses have is on the ability scale, which can be adjusted using linear scale transformations. Thus, this study advises using IRT to compare assessments in the future over CTT.

B. Item response theory

The item characteristic curves for each assessment can be found in Figs. 11–15, located in the Appendix. The test information curves overlaid with histograms of the estimated student ability scores, each on the CSEM scale, for both assessments gives a visualization for how well each assessment estimates the student latent ability scores, see Figs. 9 and 10. From these plots it can be seen that the BEMA's and CSEM's test information curves adequately cover the majority of the students' latent abilities for the samples analyzed. To compare each instrument to one another the test information curves can be overplotted, as was done in Fig. 5. From this figure it can be seen that the CSEM has superior information compared to the BEMA for the majority of the latent ability space. This is partially due to the unidimensional treatment of the CSEM and BEMA where all 32 questions on the CSEM fit the model, but only 29 questions from the BEMA properly fit. Since each question on an assessment lends information to the test, the BEMA will have less information overall compared to the CSEM since it has fewer questions. In short, it appears that the CSEM is superior to the BEMA in estimating student abilities for the entire range of typical student scores, $\theta \approx -2$ to $\theta \approx 3$. When the circuits questions are removed from the BEMA the majority of the information is lost in the $\theta \approx 0$ to $\theta \approx 2$ range. This means the circuits questions on the BEMA are information heavy for higher ability scores.

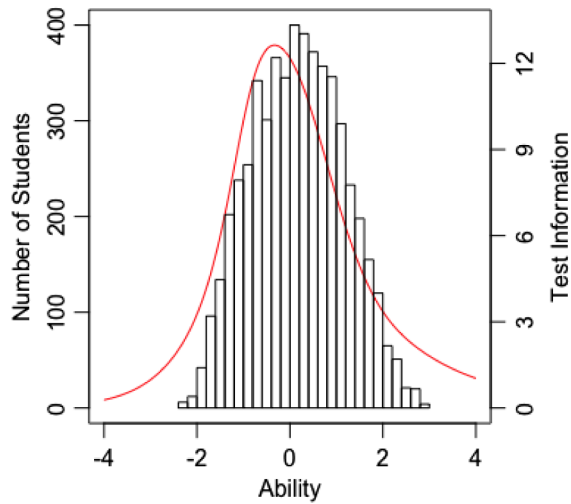


FIG. 9. Plotted in the CSEM scale, the histogram of estimated student abilities for the BEMA is overlaid with the BEMA’s test information curve. This shows that the bulk of the students are located in a region of ability space where their latent abilities will be well estimated.

The median student abilities for each instrument, on the CSEM scale, came out to be BEMA: 0.177, CSEM: -0.123 , and the median item difficulties were BEMA: 0.358, CSEM: -0.274 . The differences between these values for each assessment ($\theta_{med} - \delta_{med}$) comes out to be BEMA: -0.181 , CSEM: 0.151. This means the student abilities found by the BEMA tend to be located on the bottom half of the item abilities slightly more often than the top half. Similarly, the student abilities found by the CSEM tend to be located on the top half of the item abilities slightly more often than the bottom half. This may suggest

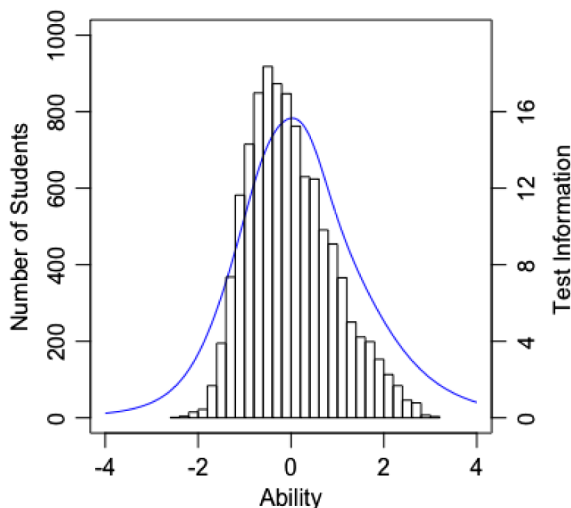


FIG. 10. Plotted in the CSEM scale, the histogram of estimated student abilities for the CSEM is overlaid with the CSEM’s test information curve. This shows that the bulk of the students are located in a region of ability space where their latent abilities will be well estimated.

that the BEMA is seen by the students as being slightly more difficult than the CSEM. However, since the BEMA also tended to give a larger student ability score compared to the CSEM it could be said that the BEMA is easier than the CSEM. Evidence for one side of this dispute over the other can be found in Fig. 5. As can be seen, the BEMA and CSEM have very similar test information on the range from $\theta \approx -4$ to $\theta \approx -1$. It is here that the CSEM’s information is superior to the BEMA until $\theta \approx 3$. This indicates that the CSEM better estimates student abilities in the higher ability range, and could be said to be the more difficult assessment. Ultimately, the differences in the difficulty of the assessments are minor, so both assessments can be said to be of equal difficulty, as was suggested by the CTT results. This result seems to support the classical results presented in Ref. [5], where it was concluded that although there were minor differences, each assessment was essentially equally difficult.

V. CONCLUSIONS

It was found through a two-tailed t test of the total scores that the assessments were significantly different from one another, but this difference was marginal and for practical use is not important. However, when a classical item level analysis was done most of the 7 shared questions between the BEMA and CSEM functioned in significantly different ways; albeit the differences were quite small for most of the items. This could be due to differences in the samples used for each of the assessments, and/or due to differences in the performance of the assessments themselves. However, since the majority of the differences were small this indicates that overall both assessments function in very similar manners, but the internal interactions between questions are slightly different. Ultimately, from a CTT perspective the BEMA and CSEM are essentially identical in their overall performance. This conclusion is in good agreement with the conclusion reached by Ref. [5]. However, since CTT is explicitly sample dependent, any differences found between the assessments could be a result of differences in the samples. Since the samples for the BEMA and CSEM are not identical, any conclusions about differences in the assessments made using CTT cannot be separated from potential sample differences. This is why IRT was used to offer support for the CTT conclusions. Since IRT is model driven, comparisons can be made without significant sample interference.

Through the use of latent ability scale linking, using the mean-mean method, it was found that the linear transformation of student abilities from the BEMA scale to the CSEM scale for these data sets was $\theta_{CSEM\ scale} = S\theta_{BEMA\ scale} + C$, where $S = 1.166$ and $C = 0.224$ [22]. This shows that the spacing between the two scales is essentially the same, $S \approx 1$, and that the means of each assessments’ item difficulties will be slightly different $C \neq 0$. It was found that the student abilities for the BEMA tended

to be higher than the ones found by the CSEM, and similarly with the item difficulties. But, due to information superiority of the CSEM over the BEMA in the higher ability ranges it is hard to identify which assessment is more difficult than the other. This potentially allows instructors to reduce the effects of test and retest on pre-post testing since the CSEM could be used as a pretest and the BEMA as a post-test, or vice versa. With both assessment being of similar difficulty, one assessment could be given as a pretest, and the other given as a post-test. Results can then be compared knowing that each assessment is of similar difficulty.

When looking at the assessments from an IRT perspective slight differences in performance were found. Although both assessments were observed to be quite similar, the CSEM was found to have superior test information in the typical student ability range of $\theta \approx -2$ to $\theta \approx 3$ for these assessments. Thus, if IRT is to be used to answer a research question in future studies, this result suggests that the CSEM is the instrument of choice.

Because of content comparison concerns, specifically that the BEMA contains circuits questions and the CSEM does not, a reduced BEMA was constructed using the data from the full BEMA. Upon redoing the analysis there were no large changes

in the CTT results of the study, which may indicate that circuits are not well correlated with the other concept areas on the BEMA. The IRT information function was heavily changed, but this was due to the direct removal of items (and thus information), and is not a result of item-item interactions which have been assumed to be zero due to item independence.

Suggested further studies would be to perform the analysis done here while attempting to fix some of the limitations of this study addressed in Sec. II: An extension of this study that exclusively explores the shared items for these assessments using IRT which seeks to understand the interaction of the items on each assessment; a comparative analysis between the BEMA and CSEM with the Electricity and Magnetism Conceptual Assessment would be of interest, as well as an IRT comparative analysis between the Force Concept Inventory and the Force and Motion Conceptual Evaluation that would extend the results of Ref. [26].

ACKNOWLEDGMENTS

The authors would like to thank Physport for allowing us to use their data. This project was funded by the Montana State University Physics Department.

APPENDIX: TABLES AND FIGURES

The following are the numerical and graphical results for the questions on the BEMA and CSEM found within this study. The CTT item statistics and IRT item parameters (with errors in parenthesis) are located in Tables IV–V. Figures 11–15 contain plots of the item response functions, also called item characteristic curves, which can be used to visually assess the goodness-of-fit of the 2PL model to the data.

TABLE IV. BEMA’s CTT and IRT item parameters in the BEMA scale with standard errors.

Item	CTT statistics			IRT statistics	
	Diff. (S.E.)	Disc. (S.E.)	r_{pb} (S.E.)	α (S.E.)	δ (S.E.)
q1	0.828 (0.023)	0.309 (0.061)	0.328 (0.024)	1.046 (0.054)	-1.805 (0.077)
q2	0.681 (0.028)	0.558 (0.066)	0.467 (0.021)	1.397 (0.055)	-0.75 (0.033)
q3	0.731 (0.027)	0.507 (0.067)	0.448 (0.022)	1.407 (0.053)	-0.369 (0.028)
q4	0.687 (0.029)	0.582 (0.068)	0.497 (0.02)	1.414 (0.055)	-0.77 (0.032)
q5	0.439 (0.031)	0.665 (0.064)	0.524 (0.02)	1.355 (0.05)	0.235 (0.028)
q6	0.478 (0.031)	0.7 (0.063)	0.549 (0.019)	1.472 (0.053)	0.069 (0.026)
q7	0.446 (0.03)	0.481 (0.074)	0.38 (0.023)	0.709 (0.036)	0.334 (0.045)
q8	0.704 (0.028)	0.439 (0.071)	0.379 (0.023)	0.847 (0.042)	-1.183 (0.06)
q9	0.29 (0.028)	0.12 (0.077)	0.12 (0.026)	0.084 (0.033)	10.671 (4.165)
q10	0.544 (0.031)	0.566 (0.07)	0.45 (0.022)	0.956 (0.041)	-0.227 (0.035)
q11	0.333 (0.029)	0.392 (0.075)	0.336 (0.024)	0.624 (0.036)	1.208 (0.078)
q12	0.187 (0.024)	0.291 (0.069)	0.325 (0.024)	0.801 (0.044)	2.068 (0.104)
q13	0.723 (0.028)	0.477 (0.068)	0.43 (0.022)	1.085 (0.047)	-1.093 (0.047)
q14	0.44 (0.031)	0.317 (0.08)	0.269 (0.025)
q15	0.687 (0.029)	0.673 (0.06)	0.579 (0.018)	1.741 (0.065)	-0.696 (0.028)
q16	0.377 (0.03)	0.559 (0.069)	0.466 (0.021)	0.93 (0.047)	1.85 (0.083)
q17	0.33 (0.029)	0.299 (0.077)	0.267 (0.025)	0.435 (0.034)	1.695 (0.14)
q18	0.537 (0.031)	0.113 (0.085)	0.11 (0.027)	0.043 (0.03)	-3.451 (2.485)
q19	0.702 (0.028)	0.417 (0.071)	0.357 (0.024)	0.754 (0.04)	-1.279 (0.07)

(Table continued)

TABLE IV. (*Continued*)

Item	CTT statistics			IRT statistics	
	Diff. (S.E.)	Disc. (S.E.)	r_{pb} (S.E.)	α (S.E.)	δ (S.E.)
q20	0.466 (0.031)	0.699 (0.061)	0.555 (0.019)	1.471 (0.053)	0.115 (0.026)
q21	0.751 (0.026)	0.583 (0.064)	0.532 (0.019)	1.905 (0.075)	-0.926 (0.03)
q22	0.556 (0.031)	0.737 (0.059)	0.571 (0.018)	1.664 (0.059)	-0.216 (0.025)
q23	0.44 (0.03)	0.474 (0.075)	0.385 (0.023)	0.747 (0.037)	0.36 (0.044)
q24	0.615 (0.03)	0.575 (0.069)	0.471 (0.021)	1.116 (0.045)	-0.533 (0.034)
q25	0.459 (0.031)	0.546 (0.073)	0.441 (0.022)	0.928 (0.04)	0.202 (0.036)
q26	0.326 (0.029)	0.534 (0.072)	0.453 (0.022)	1.043 (0.044)	0.842 (0.042)
q27	0.135 (0.021)	0.191 (0.062)	0.236 (0.025)	0.624 (0.046)	3.194 (0.217)
q2829	0.079 (0.017)	0.159 (0.051)	0.269 (0.025)
q28	0.533 (0.049)	4.066 (0.345)
q29
q30	0.342 (0.029)	0.606 (0.066)	0.513 (0.02)	1.329 (0.051)	0.649 (0.033)
q31	0.289 (0.028)	0.196 (0.078)	0.195 (0.026)	0.276 (0.033)	3.323 (0.405)

TABLE V. CSEM's CTT and IRT item parameters, in the CSEM scale, with standard errors.

Item	CTT statistics			IRT statistics	
	Diff. (S.E.)	Disc. (S.E.)	r_{pb} (S.E.)	α (S.E.)	δ (S.E.)
q1	0.647 (0.03)	0.318 (0.077)	0.263 (0.018)	0.439 (0.026)	-1.443 (0.091)
q2	0.437 (0.031)	0.365 (0.079)	0.301 (0.018)	0.511 (0.024)	0.52 (0.048)
q3	0.798 (0.025)	0.422 (0.064)	0.399 (0.017)	1.297 (0.047)	-1.376 (0.039)
q4	0.502 (0.031)	0.579 (0.07)	0.466 (0.016)	1.284 (0.037)	-0.042 (0.021)
q5	0.642 (0.03)	0.515 (0.07)	0.413 (0.016)	1.09 (0.033)	0.053 (0.023)
q6	0.685 (0.029)	0.575 (0.069)	0.465 (0.016)	1.341 (0.041)	-0.795 (0.025)
q7	0.386 (0.031)	0.707 (0.064)	0.58 (0.013)	1.814 (0.048)	0.35 (0.019)
q8	0.594 (0.03)	0.604 (0.067)	0.474 (0.015)	1.196 (0.036)	-0.433 (0.023)
q9	0.53 (0.031)	0.573 (0.071)	0.454 (0.016)	1.055 (0.032)	-0.165 (0.023)
q10	0.438 (0.031)	0.578 (0.071)	0.476 (0.015)	1.117 (0.033)	0.253 (0.024)
q11	0.37 (0.03)	0.507 (0.071)	0.424 (0.016)	0.886 (0.029)	0.682 (0.033)
q12	0.717 (0.028)	0.444 (0.072)	0.386 (0.017)	0.988 (0.035)	-1.133 (0.039)
q13	0.274 (0.028)	0.396 (0.072)	0.373 (0.017)	0.766 (0.028)	1.421 (0.054)
q14	0.17 (0.023)	0.13 (0.066)	0.159 (0.019)	0.262 (0.028)	6.135 (0.657)
q15	0.359 (0.03)	0.525 (0.073)	0.453 (0.016)	1.011 (0.031)	0.674 (0.03)
q16	0.325 (0.029)	0.486 (0.072)	0.428 (0.016)	0.925 (0.03)	0.914 (0.036)
q17	0.47 (0.031)	0.554 (0.072)	0.446 (0.016)	0.979 (0.031)	0.122 (0.025)
q18	0.526 (0.031)	0.338 (0.08)	0.278 (0.018)	0.448 (0.024)	-0.25 (0.048)
q19	0.482 (0.031)	0.571 (0.072)	0.459 (0.016)	1.021 (0.031)	0.059 (0.024)
q20	0.236 (0.026)	0.264 (0.072)	0.281 (0.018)	0.527 (0.027)	2.359 (0.117)
q21	0.195 (0.024)	0.264 (0.071)	0.314 (0.018)	0.706 (0.03)	2.207 (0.086)
q22	0.364 (0.03)	0.295 (0.079)	0.259 (0.018)	0.421 (0.024)	1.374 (0.089)
q23	0.591 (0.031)	0.632 (0.065)	0.498 (0.015)	1.365 (0.04)	-0.397 (0.021)
q24	0.282 (0.028)	0.413 (0.075)	0.392 (0.017)	0.898 (0.03)	1.203 (0.042)
q25	0.432 (0.031)	0.567 (0.073)	0.461 (0.016)	1.008 (0.031)	0.303 (0.026)
q26	0.644 (0.03)	0.643 (0.064)	0.513 (0.015)	1.563 (0.046)	-0.576 (0.02)
q27	0.243 (0.027)	0.352 (0.076)	0.373 (0.017)	0.827 (0.03)	1.561 (0.055)
q28	0.563 (0.031)	0.445 (0.076)	0.357 (0.017)	0.704 (0.027)	-0.415 (0.034)
q29	0.268 (0.027)	0.342 (0.079)	0.347 (0.017)	0.728 (0.028)	1.528 (0.06)
q30	0.548 (0.031)	0.461 (0.076)	0.371 (0.017)	0.722 (0.027)	-0.311 (0.032)
q31	0.222 (0.026)	0.427 (0.069)	0.439 (0.016)	1.068 (0.034)	1.423 (0.042)
q32	0.348 (0.029)	0.299 (0.077)	0.26 (0.018)	0.415 (0.024)	1.57 (0.1)

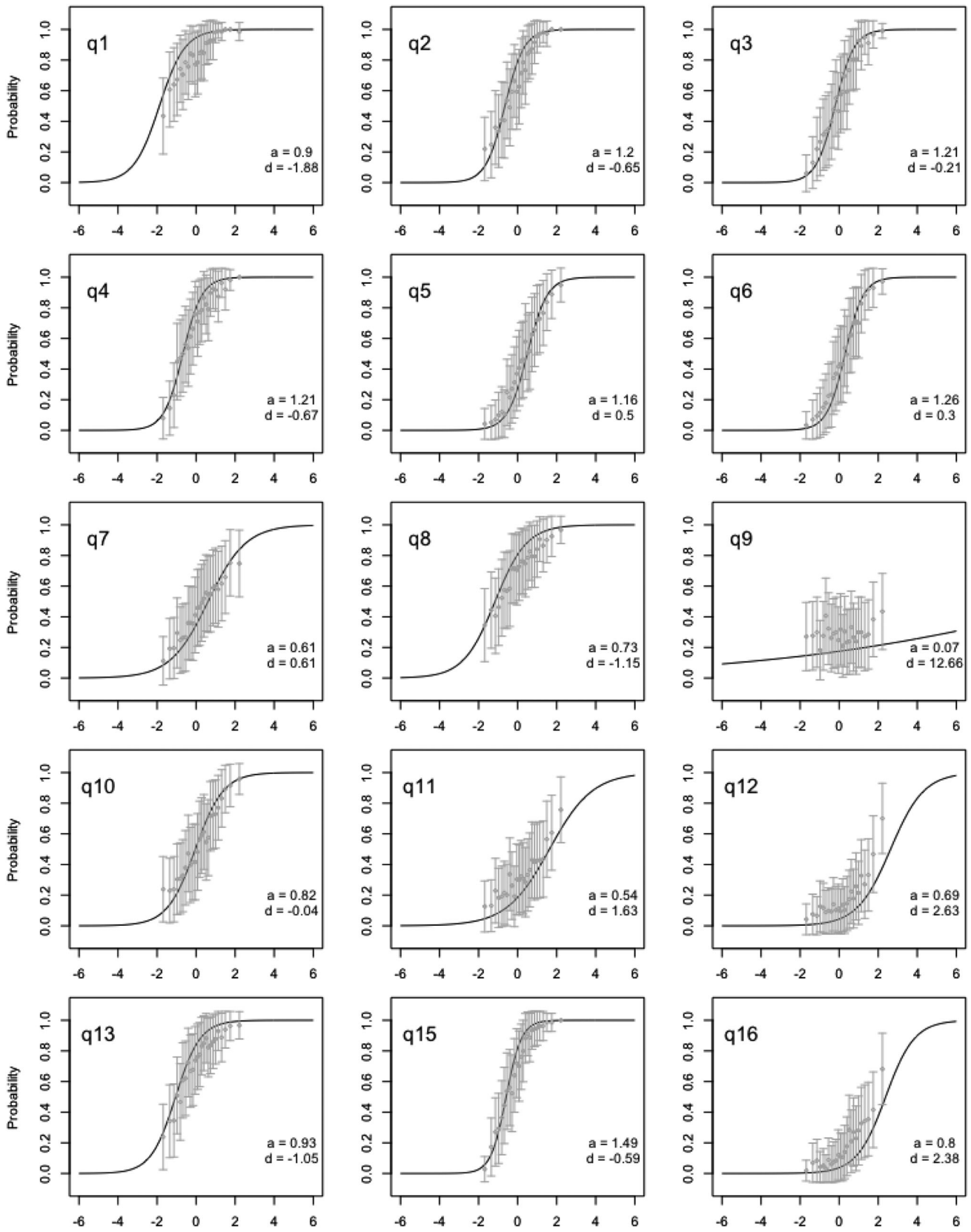


FIG. 11. Item characteristic curves for the item 1–15 on the BEMA, in the CSEM scale. The gray error bars represent 1 standard deviation, from top to bottom, in the student responses. The horizontal axis is student latent ability.

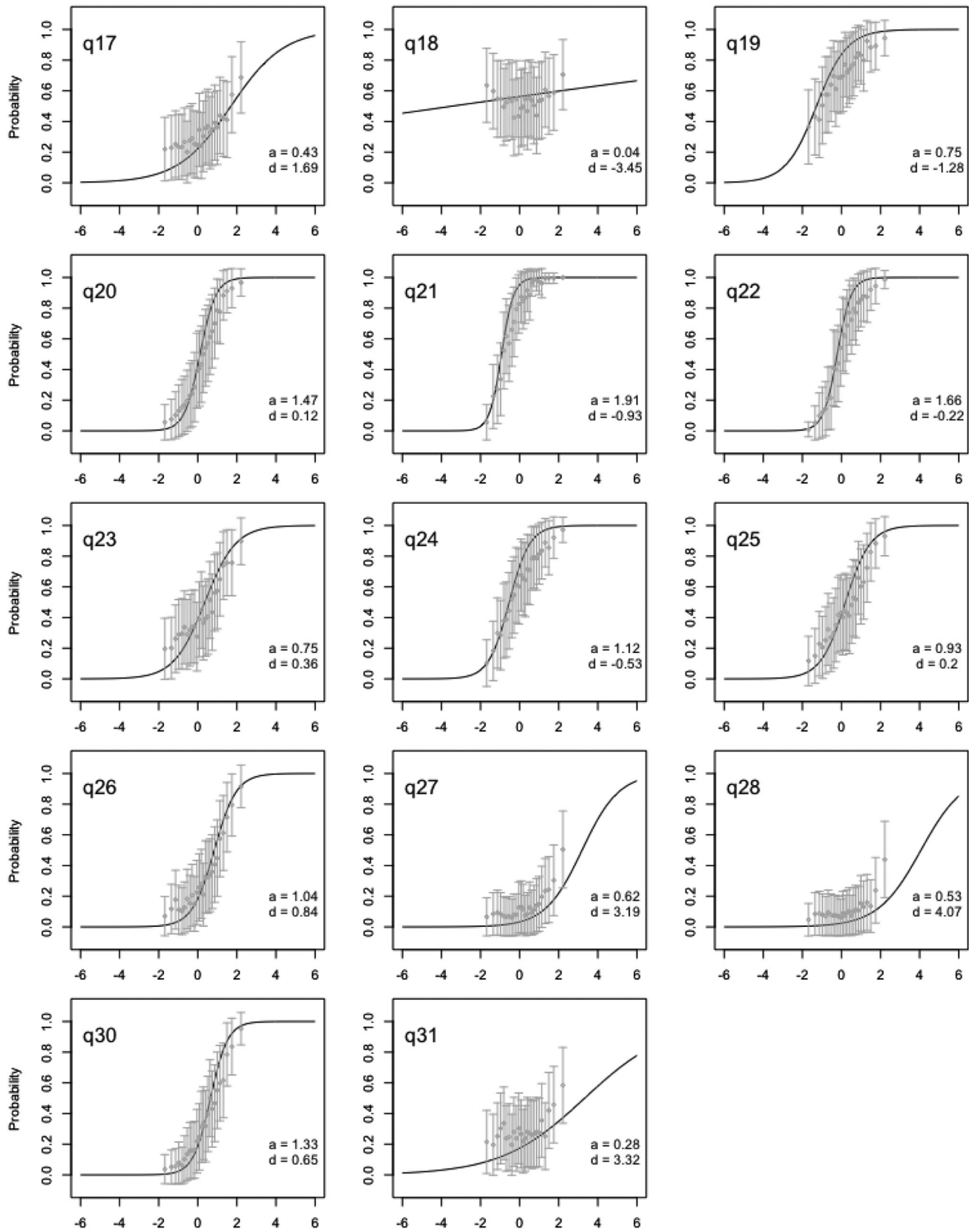


FIG. 12. Item characteristic curves for the item 16–31 on the BEMA, in the CSEM scale. The gray error bars represent 1 standard deviation, from top to bottom, in the student responses. The horizontal axis is student latent ability.

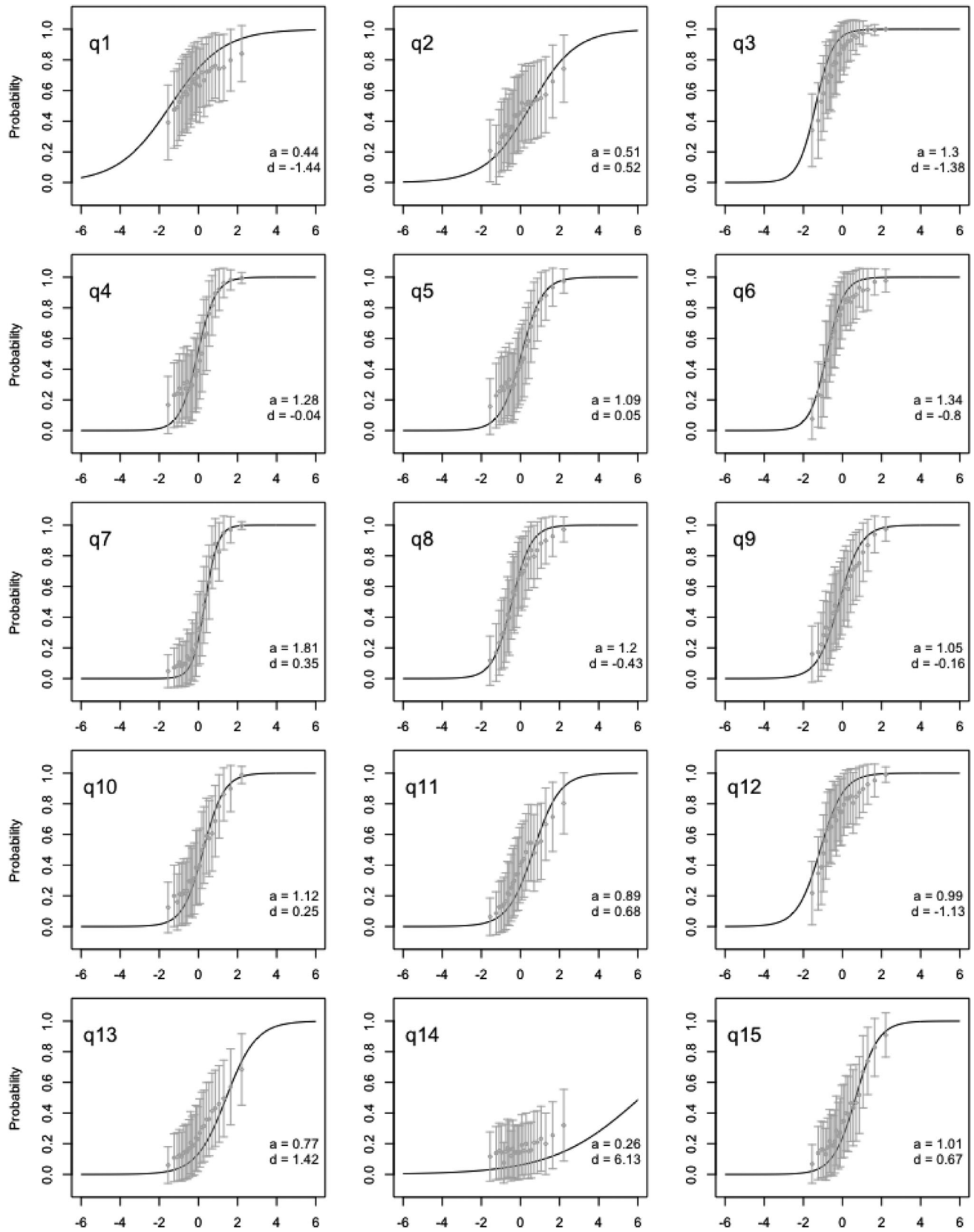


FIG. 13. Item characteristic curves for the item 1–15 on the CSEM. The gray error bars represent one standard deviation, from top to bottom, in the student responses. The horizontal axis is student latent ability.

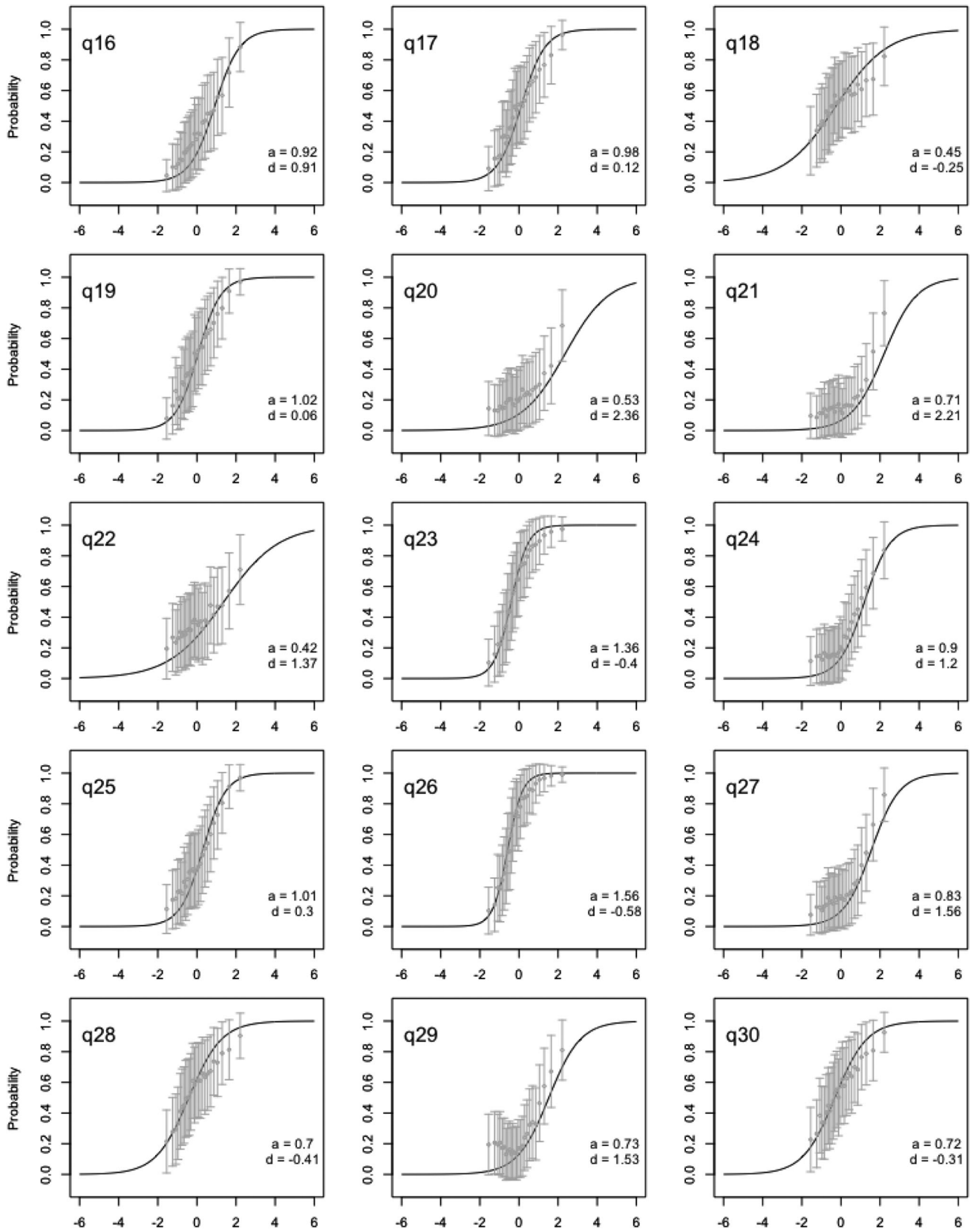


FIG. 14. Item characteristic curves for the item 16–30 on the CSEM. The gray error bars represent one standard deviation, from top to bottom, in the student responses. The horizontal axis is student latent ability.

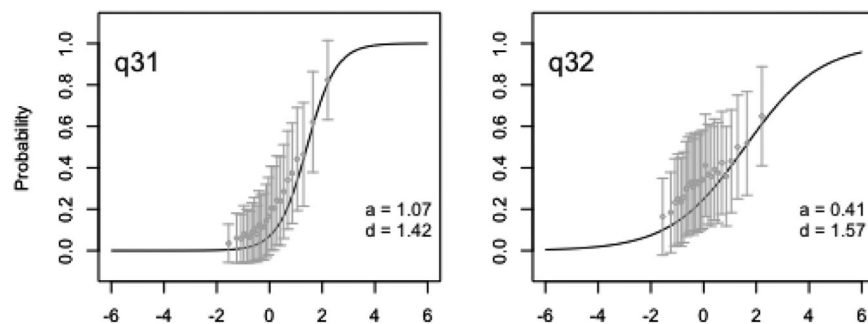


FIG. 15. Item characteristic curves for the item 31–32 on the CSEM. The gray error bars represent one standard deviation, from top to bottom, in the student responses. The horizontal axis is student latent ability.

- [1] 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).
- [2] D. P. Maloney, T. L. O’Kuma, C. J. Hieggelke, and A. V. Heuvelen, Surveying Students? Conceptual knowledge of electricity and magnetism, *Am. J. Phys.* **69**, S12 (2001).
- [3] M. W. McColgan, R. A. Finn, D. L. Broder, and G. E. Hassel, Assessing students’ conceptual knowledge of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **13**, 020121 (2017).
- [4] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the force and motion conceptual evaluation and the conceptual survey of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **14**, 020103 (2018).
- [5] S. J. Pollock, Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA, *AIP Conf. Proc.* **1064**, 171 (2008).
- [6] D. C. Giancoli, *Physics for Scientists and Engineers* (Pearson Education International, Upper Saddle River, NJ, 2008).
- [7] PhysPort, Security FAQ for the Assessment Data Explorer, 2017.
- [8] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, Hillsdale, NJ, 1988).
- [9] D. George and P. Mallery, *IBM SPSS Statistics 23 Step by Step: A Simple Guide and Reference*, 14th ed. (Routledge, New York, 2016).
- [10] G. F. Kuder and M. W. Richardson, The theory of the estimation of test reliability, *Psychometrika* **2**, 151 (1937).
- [11] Rodney L. Doran, Implications for measurement and evaluation from the trends of science education, *Sci. Educ.* **60**, 199 (1976).
- [12] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020103 (2009).
- [13] W. Wiersma and S. G. Jurs, *Educational Measurement and Testing* (Allyn & Bacon Boston, MA, 1985).
- [14] T. Kline, *Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses* (SAGE Publications, Inc., Thousand Oaks, CA, 2005).
- [15] M. D. Reckase, Unifactor latent trait models applied to multifactor tests: Results and implications, *J. Educ. Stat.* **4**, 207 (1979).
- [16] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, *Am. J. Phys.* **78**, 1064 (2010).
- [17] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010124 (2018).
- [18] T. A. Brown, *Confirmatory Factor Analysis for Applied Research* (Guilford Publications, NY, New York, 2014).
- [19] R. Philip Chalmers *et al.*, MIRT: A multidimensional item response theory package for the R environment, *J. Stat. Softw.* **48**, 1 (2012).
- [20] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford Publications, NY, New York, 2013).
- [21] B. LeBeau and A. McVay, *Validity of the Three Parameter Item Response Theory Model for Field Test Data ITP Research Series* (ITP Research Series, University of Iowa, 2017).
- [22] M. J. Kolen and R. L. Brennan, *Test Equating, Scaling, and Linking* (Springer, New York, 2004).
- [23] W.-C. Lee and J.-C. Ban, A comparison of IRT linking procedures, *Appl. Meas. Educ.* **23**, 23 (2009).
- [24] W. Lin, Q. Jiahe, and L. Yi-Hsuan, Exploring alternative test form linking designs with modified equating sample size and anchor test length, *ETS Res. Rep. Ser.* **2013**, i (2013).
- [25] J. Yasuda, N. Mae, M. M. Hull, and M. Taniguchi, Analyzing false positives of four questions in the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010112 (2018).
- [26] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the force concept inventory, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010105 (2009).