Seeking missing pieces in science concept assessments: Reevaluating the Brief Electricity and Magnetism Assessment through Rasch analysis

Lin Ding*

Department of Teaching and Learning, The Ohio State University, Columbus, Ohio 43210, USA

(Received 27 August 2013; published 14 February 2014)

Discipline-based science concept assessments are powerful tools to measure learners' disciplinary core ideas. Among many such assessments, the Brief Electricity and Magnetism Assessment (BEMA) has been broadly used to gauge student conceptions of key electricity and magnetism (E&M) topics in college-level introductory physics courses. Differing from typical concept inventories that focus only on one topic of a subject area, BEMA covers a broad range of topics in the electromagnetism domain. In spite of this fact, prior studies exclusively used a single aggregate score to represent individual students' overall understanding of E&M without explicating the construct of this assessment. Additionally, BEMA has been used to compare traditional physics courses with a reformed course entitled Matter and Interactions (M&I). While prior findings were in favor of M&I, no empirical evidence was sought to rule out possible differential functioning of BEMA that may have inadvertently advantaged M&I students. In this study, we used Rasch analysis to seek two missing pieces regarding the construct and differential functioning of BEMA. Results suggest that although BEMA items generally can function together to measure the same construct of application and analysis of E&M concepts, several items may need further revision. Additionally, items that demonstrate differential functioning for the two courses are detected. Issues such as item contextual features and student familiarity with question settings may underlie these findings. This study highlights often overlooked threats in science concept assessments and provides an exemplar for using evidence-based reasoning to make valid inferences and arguments.

DOI: 10.1103/PhysRevSTPER.10.010105 PACS numbers: 01.40.Fk, 01.40.gf

I. INTRODUCTION

Assessment is an integral component of science education. When properly designed and implemented, assessments can be effectively used to assist learning, monitor student progress, and evaluate educational programs. Given the increasingly heightened attention paid to the outcomes of assessments and their possible ramifications to decision making, educators and researchers are urged to reexamine the quality of educational assessments and particularly the appropriateness of the inferences and actions that are made based on assessment results [1–3]. As emphasized by the National Research Council, the very core of educational assessment is an "evidence-based reasoning" process [1]. This should not only involve carefully crafted assessment instruments that are grounded in learning theories and capable of eliciting students' knowledge and skills, but also require sufficiently accurate analysis models and interpretation mechanisms to allow for valid and reliable arguments about teaching and learning [4–7]. To this end, careful investigations of broadly used educational assessments to inform and reshape future science curricula are warranted.

In the past two decades, a large number of science assessments have been developed to measure students' various cognitive constructs. Among them, discipline-based concept inventories have been an important focus, as they directly target disciplinary core ideas. In physics education, the first of its kind—the Force Concept Inventory (FCI) [8]—has been instrumental in revealing student alternative Newtonian ideas and has served as a catalyst for many physics curricular reforms. Inspired by this, researchers invested a great deal of time and effort to develop similar concept assessments for use in other subject domains such as electricity and magnetism (E&M) as well as in other science disciplines [9,10]. While these instruments are frequently employed to gauge students' learning of disciplinary core ideas and to compare the effectiveness of science curricula, they are often used with *prima facie* credibility without being subject to additional validity and reliability investigations. For example, in most cases a student's understanding of a scientific topic is represented by a single score on an assessment—the sum of the questions the student has correctly answered. This approach to representing student conceptual understanding is based largely on a putative assumption; that is, a single score is a sufficient and meaningful indicator that can lead to valid inferences about the student's understanding

*Corresponding author.
ding.65@osu.edu

of the tested topics [11–13]. However, this assumption may not always hold, particularly in situations where the assessment is designed to test a diverse range of topics [14]. In such cases, a single aggregate score can be problematic or even misleading (as a set of separate scores may be needed to best represent each set of closely related topics). Consequently, analysis and interpretation derived from such aggregate scores are likely to generate inaccurate or even false conclusions.

Similarly, when using a concept assessment to compare the effectiveness of different curricula that cover the same content, a critical yet often overlooked issue is assessment bias [15,16]. Ideally, the selected assessment is appropriate and serves as an unbiased measure for all students in different curricula, if comparable learning opportunities are indeed provided to students in these curricula. However, it is not uncommon that consistent nonzero measurement errors, also known as biases, may occur in some part of the assessment [15]. For instance, a question on an assessment may be situated in a context that is more familiar to students in one class than to those in another. This could result in students in the former class having an inadvertently higher success rate on the question, even though both classes may indeed share a similar mastery of the tested topics. Such issues will likely go undetected if they are not empirically examined and monitored, and as a result, the inferences and conclusions drawn from the assessment outcomes can skew reality. At their worst, these skewed inferences may either inflate or underestimate the effectiveness of a science curriculum, misinforming future decision making on curriculum development and reform.

In this paper we reexamine one of the broadly used concept assessments, the Brief Electricity and Magnetism Assessment (BEMA) [10], to highlight these critically important issues that have been overlooked in prior work. Specifically, drawing on the "reasoning-from-evidence" framework, we seek to uncover two missing pieces in the common practices of using science assessments to compare curriculum effectiveness. One is concerned with the rationale of representing student conceptual understanding by an aggregate score. This issue is foundational because a body of studies relied on this approach to interpret BEMA results and draw conclusions [17,18]. If no evidence is found to support this approach, a significant portion of the prior results will become questionable. The other issue we seek to address is whether the assessment is potentially biased in favor of one group of students over the other. For example, BEMA has been adopted to measure student understanding of key electricity and magnetism concepts in different physics curricula where students were exposed to the same content [18]. Comparisons were often carried out without checking the possibility of bias and hence provided no evidence for the absence of this potential threat.

II. THEORETICAL BACKGROUND

Following the evidence-based reasoning framework, we seek to investigate these missing pieces in the context of BEMA to illustrate their consequential significance in science educational assessment. In this section, we (1) review the content and construct of BEMA, (2) explicate the possible threat in using an aggregate score to represent a student's conceptual understanding, and (3) address the theoretical perspectives on potential measurement bias in using concept assessments for comparative studies. Of these three aspects, the last two directly relate to the two missing pieces we attempt to seek in BEMA.

A. Content and construct of BEMA

BEMA is a 30-item multiple-choice assessment designed to measure student conceptual understanding of key topics in electricity and magnetism [10]. Since it is intended to be a common-denominator assessment suitable for use in various college-level introductory E&M courses, only those that are considered core concepts by instructors of both traditional and reformed courses are included in the assessment [10]. To a large extent, BEMA shares many similarities with concept inventories in terms of design, format, and usage. In other words, as with concept inventories, BEMA was designed to probe student conceptual understanding of disciplinary core ideas, formatted in the multiple-choice mode, and can be used for both pre- and postinstructional measurements to track student learning gains. However, it also differs noticeably from regular concept inventories in terms of the breadth of content covered in the assessment. Typically, a concept inventory is an assessment designed to probe student understandings of a single topic [19–21]. The FCI is such an example that focuses only on the Newtonian concepts of force—one of the many topics that are discussed in mechanics [8]. Alternatively, BEMA covers a broad range of key concepts in the domain of electricity and magnetism. Topics therein range from electric charges and fields that are typically taught at the beginning of an E&M course to electromagnetic inductions, such as Ampere's law and Faraday's law, that are discussed near the end of the course [10] (also see Supplemental Material [22]). To distinguish assessments with a broad content coverage (like BEMA) from those with a narrowed focus (like FCI), researchers refer to the former as concept *surveys* and the latter as concept *inventories* [9].

The broad content coverage in BEMA raises a serious question: Is using a single score by summing up correct responses a meaningful way to represent a student's overall understanding of this broad subject domain? In other words, can we make claims about student conceptual understanding based on this broad concept survey? Or, from a measurement perspective, can the individual questions on BEMA that are aimed at a wide range of topics

morph into a cohesive construct—a trait or a competency of interest? Unlike concept inventories for which it is easier to make an argument about a focused construct due to content homogeneity and hence about the rationale of using a single aggregate score to represent the construct, it is challenging to make a convincing case for concept surveys [14].

From the measurement theory viewpoint, a clearly delineated construct is essential for valid data interpretation and inferences. However, when the content of an assessment tested by different questions becomes increasingly heterogeneous, these questions run a risk of potentially representing different underlying constructs or distinct dimensions, hence reducing the coherence and interpretability of what the assessment is testing [14,23,24]. In fact, for assessments with a broad content bandwidth, even when classical test theory reports a high reliability, a single construct or unidimensionality still cannot be guaranteed [23]. In this case, using a single score to represent multiple constructs or different dimensions increases measurement uncertainty and obscures the nature of the intended construct. Specifically, two sources of ambiguity are likely to be introduced into test results. One is the lack of clarity in the contribution of each dimension to the composite score. The other is the uncertainty in score comparisons, because "the same composite score is likely to reflect different combinations of constructs for different members of the sample" [14]. In light of this theoretical basis, it is crucial that the construct of BEMA be empirically investigated to offer cogent arguments for the validity of using a single score to represent student learning of a broad range of E&M topics. Unfortunately, this issue has not been addressed in prior studies.

From a different theoretical viewpoint of physical sciences, electricity and magnetism concepts by nature should form a cohesive entirety, because the topics in this subject area, no matter how complex or seemingly diverse they are, can always be traced back to no more than a few fundamental principles regarding charges, fields, and their interactions [25]. Perhaps this is why the content of introductory-level E&M courses has more or less remained constant for the past century. Nevertheless, this scientific grounding lacks empirical verification, especially when it comes to the learning and teaching of these topics. For example, prior studies of the FCI have shown that the construct of Newtonian force conceptions viewed from the scientific perspective were often misaligned with the empirical outcomes from student learning of this topic [26,27]. To this end, it is necessary that we uncover construct-related evidence for BEMA in order to make inferences about the extent to which student understanding of electricity and magnetism can be represented by a single score.

B. Measurement bias and differential item functioning

Measurement bias is another critical issue in the evidence-based practices of science education assessment [15,16]. In many comparative studies, researchers often choose a common concept assessment for use with multiple groups of students to seek meaningful between-group differences. Presumably, each question on the assessment is unbiased; or more specifically, differences in student performance on each item should be solely determined by real differences in the construct being measured. In principle, students at the same level of competency as measured by the assessment should demonstrate the same (or similar) performance on each question regardless of their group membership. If a significant difference exists in student performance on an item between those with the same level of competency in each group, the item is considered to function differentially for different groups. Or simply put, it has differential item functioning (DIF), controlling for student ability levels. It is worth noting that not every between-group difference should be considered as DIF. Only those for matched students—those with the same level of competency that the assessment is intended to test—are considered as DIF [15,16]. Practically, a DIF can be a sign of item bias but does not guarantee it. In other words, DIF is a necessary but not a sufficient condition for item bias [28]. Whether or not a question with DIF is truly biased needs to be examined through analysis of its content and context in relation to the target construct being measured.

Theoretically, DIF represents a potential measurement bias in a question that can be caused by two primary effects: content and context [29]. The content effect lies in the differential learning opportunities that different groups of students may have [15]. For example, if a question tests students' knowledge about musical instruments, those who have been exposed to symphony orchestras may have a better chance to succeed on this question than those who have not. A DIF due to such a content effect does not necessarily mean the question is biased or problematic and therefore may not be the researchers' main concern. On the other hand, a context effect occurs when a change in question settings affects student performance [29]. For instance, if a particular group of students happen to be more familiar with the scenario of a question (not with what the question is meant to test) and hence have a higher chance to answer it correctly, this increased performance is undesirable and needs to be controlled.

Prior research using BEMA to study the relative effectiveness of physics curricula has overlooked the important issue of potential bias in the assessment. In large-scale studies, Kohlmyer *et al.* [18] used BEMA to measure student conceptual understanding of electricity and magnetism in two physics curricula. One is a traditional college-level calculus-based physics course, and the other is a reformed course called Matter and Interactions (M&I) [30]. In both courses, students were required to attend class for the same amount of time, were exposed to similar course content in the same academic term, and were taught

by equally experienced instructors [18]. A main difference between the two courses, however, was that the sequence of the topics in the M&I course was rearranged to highlight the hierarchical structure of physics knowledge centered on a few fundamental principles. Kohlmyer *et al.* compared student total scores between the two curricula and found that students in M&I outperformed their peers in the traditional physics courses. After taking into account many confounding factors, Kohlmyer *et al.* reached the conclusion that the M&I course is more effective in promoting student conceptual understanding of core E&M ideas than a traditional course. However, in view of the evidence-based reasoning framework, an important supporting piece is missing; that is, the threat of potential bias has not been ruled out. It is true that BEMA is designed to be appropriate for both traditional and M&I curricula, and prior studies have established sufficient content-related evidence. Nonetheless, no empirical data have been established to verify that BEMA questions are indeed not inadvertently in favor of the M&I students. Without this supporting evidence, the argument about the increased effectiveness of the M&I curriculum can be dubious.

It is worth noting that in this case there was no evidence suggesting different opportunities for students to learn the tested topics between the two courses. According to Kurz and Elliott [31], learning opportunities are conceptualized as consisting of three key aspects: instruction time, content, and quality. As mentioned earlier, both courses took place in the same academic term and involved the same instruction time. In addition, students in both courses were exposed to similar content, although the M&I students learned the required topics by following a different sequence that underscored the hierarchical structure of the physics enterprise. Moreover, instructors who taught these courses were equally experienced and gave no reason for assuming any significant difference in their teaching quality. Perhaps more importantly, the developers of BEMA stressed that this assessment was intended to be a common-denominator test. Therefore, items testing topics that were not discussed or only treated as of peripheral importance in either of the two courses were not included in BEMA [10,18]. To this end, what BEMA purports to test is presumably those key E&M topics that both traditional and M&I students would have comparable opportunities to access in their respective courses. This indeed needs to be empirically verified, because the comparison between the M&I and traditional courses was predicated on the postulation that BEMA is not in favor of one course over the other [18].

C. Research goals

In this study, we investigate the aforementioned two missing pieces. Specifically, we attempt to answer the following questions. (1) Do the individual questions on BEMA form a cohesive construct to allow a meaningful interpretation by using a single aggregate score? (2) If the answer to the previous question is affirmative, then what exactly is the construct that BEMA is intended to measure? Conversely, if the answer to the previous question is negative, then how should we better represent student performance to allow for valid inferences? (3) What evidence can speak to the issue of potential DIF in BEMA when comparing the two courses, traditional versus M&I?

III. METHODS

A. Student sample and settings

In order to provide empirical answers to the above questions, we administered BEMA to students in science and engineering majors at a large U.S. research university. These students were enrolled in two parallel calculus-based introductory E&M courses in the same academic term. Both courses were the second sequence of their respective two-semester physics curricula and were taught by equally experienced senior faculty members who valued and committed to effective teaching. One was a traditional course, in which students attended three 50-minute lectures and one 2-hour lab each week. The topics covered in this course followed a conventional sequence (see Supplemental Material [22]). The other was the Matter and Interactions E&M course [30,32,33]. Students in this course also attended three 50-minute lectures and a 2-hour lab every week. Although the topics covered in M&I were essentially the same as those discussed in the traditional course, the sequence was reorganized by following a hierarchical, principled structure (see Supplemental Material [22]) to help students increase conceptual coherence [33]. More details on the M&I curriculum can be found in Refs. [30,33]. As with the case in studies performed by Kohlymer *et al.*, students in both courses were provided similar opportunities to learn the tested topics on BEMA, as they were exposed to comparable instruction time, course content, and teacher quality.

We administered BEMA as both a pretest and a posttest to students in the traditional and M&I courses. The pretest was conducted in the first week of the course as part of class activities; a total of 190 students attending the classes on the day that BEMA was administered took the test (102 from the traditional class and 88 from M&I). To secure the test for postinstructional use, no feedback was provided to students, and students were not told they would be retested at the end of the academic term. The posttest was administered in the last week of the course; 165 students attending the classes on the day of the event completed the test (82 from the traditional class and 83 from M&I). Note that there was a significant attendance drop in the traditional class near the end of the semester, the reason for which remains unknown.

B. Rasch analysis of BEMA items and model fit

In order to match the goals of the study, we chose the dichotomous Rasch model to examine the collected data. This decision was made based on the following

considerations. First, Rasch analyses can allow us to examine whether or not the individual items fall under one single dimension to fit the model, and hence can provide evidence for construct-related arguments about BEMA [11–13,34,35]. Second, Rasch analysis can convert ordinal-level raw data to a set of interval-level estimates [36–39]. Strictly speaking, the commonly used total scores are not continuous (although they have orders) and cannot be directly subject to various statistical analyses that only interval data can suit. Rasch analysis can resolve this issue by creating an interval scale of measurement for both items and respondents [11–13]. Another intrinsic advantage of Rasch analysis is that the model-estimated item difficulty and person ability are sample independent, which is also known as invariance of measurement [40]. This means that the item difficulty estimates obtained from Rasch analysis remain more or less constant regardless of the student samples taking the test (given that the model fit is satisfactory for the samples). Similarly, the estimates of person ability are invariant regardless of the difficulty levels of the items that are pooled into the test. Because of the invariant nature, we can use Rasch-generated results to detect DIF in BEMA questions to examine whether or not potential bias exists. In this study, we used the Winsteps software [41] to carry out Rasch and DIF analyses.

C. Analysis of BEMA construct

1. Rasch analysis of unidimensionality and local independence of BEMA items

To seek construct-related evidence for BEMA, we examined the fit of the data to the Rasch model. For each item, Rasch analysis reports a set of fit statistics: infit and outfit mean square residuals and their standardized Z scores (see below for details). These statistics reflect how well the data set conforms to the model [41]. Since the Rasch model assumes all items falling under one single dimension (unidimensionality), the reported fit statistics can help identify which items, if any, do not meet this requirement [13,41].

In addition to checking the fit statistics, the unidimensionality assumption needs further verification [20]. Bejar's total test versus subtest approach is one way to evaluate this assumption [20,42,43]. The key idea is to estimate item difficulty parameters twice, first by using the total test and then by using only a subset of the test. If the assessment items form a cohesive single construct, a scatter plot of these two sets of estimates should show points near parallel to a straight line of slope 1 and intercept of 0. On the other hand, if the plotted points significantly depart from the line, the unidimensionality assumption is violated. In this study, we used this approach to test the unidimensionality of BEMA items.

Related to unidimensionality is another important assumption of the Rasch model: local independence. This means that the correlations between student responses to each item should be explained entirely by two factors: item difficulty and person ability [13,44,45]. Yen's Q3 method provides a practical way to test the local independence assumption [45]. This method looks into the correlations of Rasch residuals after removing the portion of variance that has been explained by the item and person estimates. As recommended by Yen and Fitzpatrick [45], residual correlations with a magnitude less than 0.2 are acceptable. We followed this approach to acquire further evidence regarding whether BEMA items are only related by the construct they are intended to measure.

2. Qualitative analysis of BEMA construct

To articulate what BEMA actually intends to measure, we analyzed the individual items by using a revised two-dimensional Bloom taxonomy [46–49]. These two dimensions in the Bloom taxonomy are content and cognition. The content dimension, represented by a set of nouns, reveals "what" types of knowledge are tested by each item. There are, from the lowest to the highest level, four types: facts, concepts or principles, procedures, and metacognitive knowledge (see Fig. 1). The cognition dimension, which is characterized by verbs, shows "how" mental processes are carried out. With an increasing order of complexity, these cognition levels include remember or recognize,

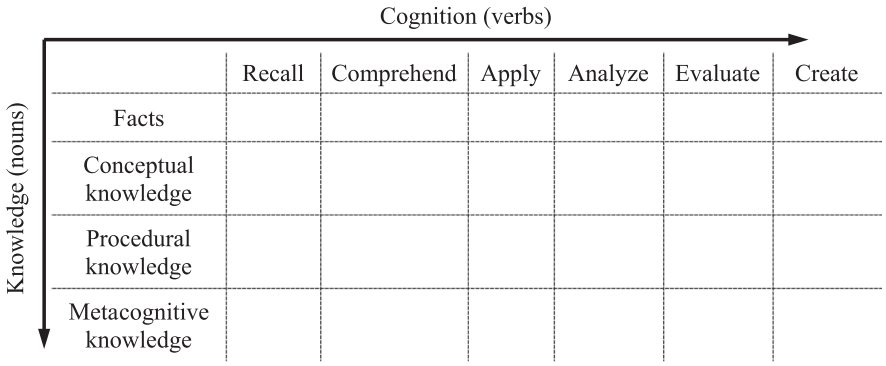

FIG. 1. Two-dimensional Bloom's taxonomy. The vertical dimension contains knowledge types (represented by nouns); the horizontal dimension contains cognition processes (represented by verbs).

comprehend, apply, analyze or synthesize, evaluate, and create. We classified BEMA items onto these two dimensions to explicate at which content and cognition levels this assessment is aimed. This analysis serves two important purposes. One is to provide an articulated description of the BEMA construct, which its developers only vaguely reported as "understanding of basic electricity and magnetism concepts." Since Bloom's taxonomy is a general operational framework for explicating educational objectives, it can help capture the qualitative details of BEMA construct while still preserving an appropriate level of generality. The other purpose is to examine, together with the Rasch measures, whether or not items classified at higher levels of Bloom's taxonomy are indeed more difficult than those at the lower levels. Such results can cast useful light on the details of BEMA construct.

D. Rasch-based analysis of DIF in BEMA

To test potential DIF in BEMA for the two groups of students (traditional versus M&I), we separately analyzed the data from the two courses. As with the previous Bejar's method, we analyzed BEMA item difficulty twice, first by using the data collected from the traditional course and then by using the data from the M&I course. According to the invariant property of Rasch measurement, the two sets of item difficulty estimates, which are both constrained to have a mean of zero by default, should be approximately equal or differ by only a constant [13]. Surely, no measurement is perfect, and in reality error is always involved. Thus, a divergence between the two sets of estimates within a certain error range (for example, 1% or 5%) is acceptable. We used Rasch measures to detect potential DIF for each item on BEMA.

IV. RESULTS

A. Rasch analysis of BEMA: Item quality and model fit

Based on collected data points, the person and item reliability of BEMA are found to be 0.78 and 0.96, respectively, indicating an adequate measure to allow for meaningful subsequent Rasch analysis. Note that person reliability in Rasch analysis is equivalent to the conventional KR-20 index or Cronbach's alpha. It essentially indicates the extent to which person placement can be replicated if a similar test is administered to the same participants. Item reliability, on the other hand, does not have a conventional equivalent. It represents the replicability of item placement along the difficulty hierarchy if the test is administered to a similar cohort of students. In making a judgment of the acceptability of reliability, one can use the traditional criterion as a reference; that is, a value equal to or above 0.7 is typically considered satisfactory [10,50,51].

At the core of Rasch analysis is the collection of construct-related evidence for BEMA. One way to do this is to examine item quality and model fit. As mentioned earlier, Rasch analysis yields a set of interval-level estimates for item difficulty and person ability. Since they are on the same interval scale, we can plot them side by side to check item and person distributions. Such a plot is called a Wright map [12,13,35]. For accurate model estimates, a close match between the item and person distributions is desired [13]. Figure 2 displays a Wright map for BEMA. In this figure, a vertical scale (with increasing values from bottom to top) separates the person ability distribution on the left and the item difficulty distribution on the right. Here, two columns of person distributions are shown; one for the pretest and the other for the posttest. As seen, student pretest performance is noticeably lower than the difficulty levels of most BEMA items. On the other hand, the posttest distribution seems to match the item distribution fairly well. However, two gaps in the item distribution are noticeable. One is at the lower end of the scale between item 1 and item 8, the other is at the higher end of the scale between item 12 and item 28. This suggests that more items with a difficulty level in these two ranges are needed to better estimate student ability.

Rasch analysis also generates a set of fit statistics to allow for inspection of model fit. For each item, two sets of fit statistics are reported: mean square residuals (MNSQ) and standardized Z statistics (ZSTD). Both reflect the difference between the observed data and model-expected values. The MNSQs are an average of squared residuals, whereas the ZSTDs are normalized Z scores of the residuals [12,13]. Depending on how MNSQs and ZSTDs are calculated, each can further generate two statistics: infit and outfit. The infit assigns more weight to those with a close person-item match, whereas outfit puts equal weight on all data points and hence is more sensitive to outliers. Typically, MNSQs within the range of $[0.7, 1.3]$ and ZSTDs within $[2, 2]$ are considered as a reasonable fit [13]. For items with MNSQs greater than 1.3 and ZSTDs greater than $+2$, there is more variance in the observed data than predicted by the model—also known as underfit. Conversely, MNSQs less than 0.7 and ZSTDs less than 2 signify that there is less variance in the data than predicted—which is also known as overfit. Overfit indicates that the data are too predictable and lack randomness, so it does not degrade model fit [13,41].

Table I shows BEMA item fit statistics. The majority of the items seem to have a reasonable fit to the model within the acceptable range. Four items (item 5, item 6, item 15, and item 16) fall below the lower end of the range and represent an overfit. Given that these items yield data that are too predictable but do not degrade the measurement, they are less of a concern. On the other hand, two items (item 9 and item 17) exceed the upper limit of the range and therefore represent an underfit. These two items warrant further inspection and need to be revised in future studies.

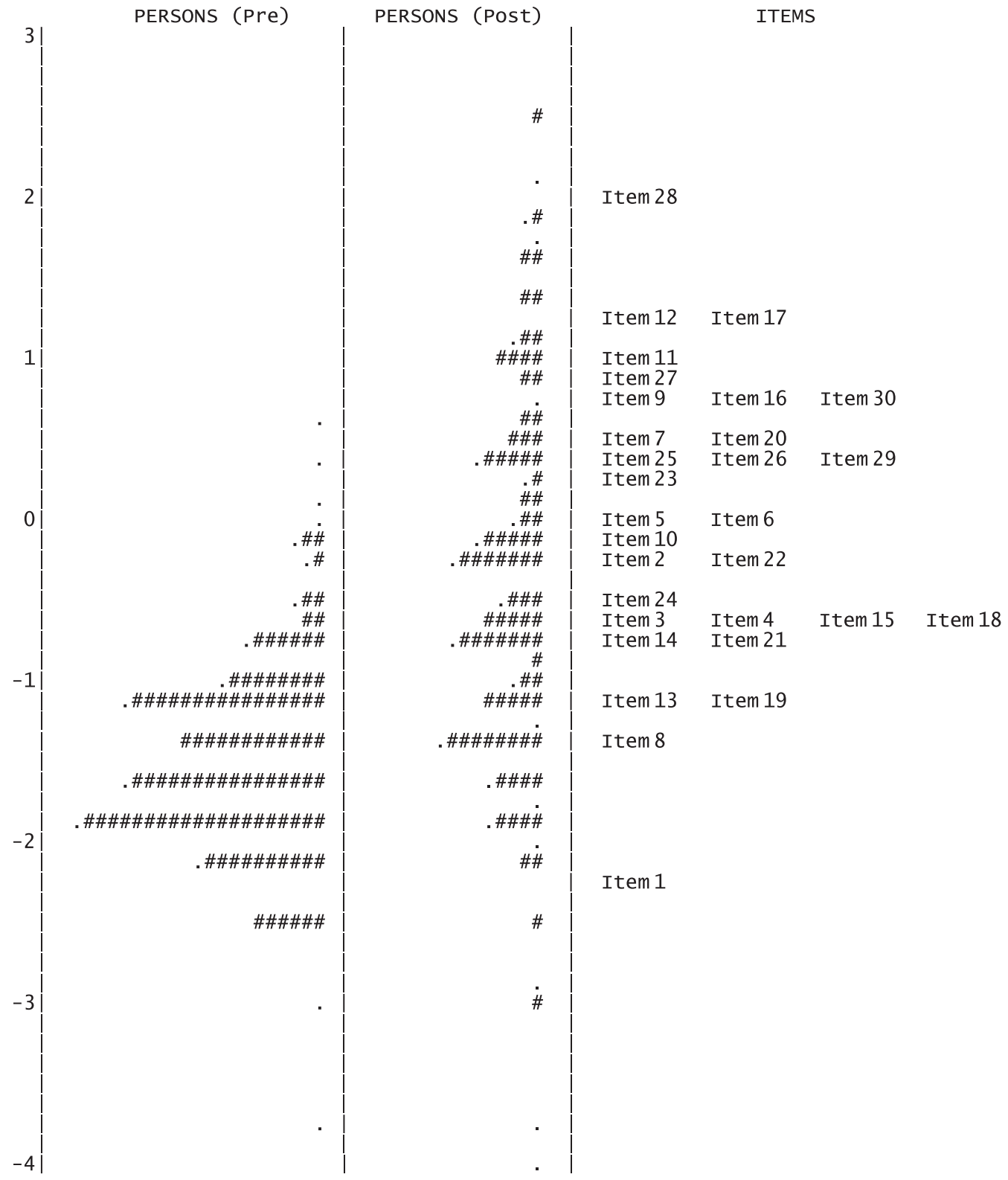

FIG. 2. A BEMA Wright map based on all data. On the left-hand side of the logit scale are two columns of student ability distribution; on the right-hand side of the logit scale is the item difficulty distribution.

TABLE I. BEMA posttest item difficulty estimates, model standard errors (SE), and infit and outfit statistics [mean square residuals (MNSQ) and Z scores].

			Infit		Outfit	
Item	Difficulty	SE	MNSQ	Z	MNSQ	Z
1	−2.31	0.21	1.02	0.2	0.92	−0.8
2	−0.23	0.16	1.03	0.5	1.07	0.7
3	−0.6	0.17	0.96	−0.6	0.91	−0.9
4	−0.62	0.16	1.02	0.3	0.98	−0.1
5	0.01	0.17	0.83	−2.8	0.83	−1.7
6	−0.04	0.17	0.86	−2.4	0.79	−2.2
7	0.55	0.18	1.04	0.5	1.09	0.9
8	−1.36	0.17	1	0	0.99	−0.1
9	0.72	0.18	1.37	4	1.71	5.8
10	−0.17	0.16	1.03	0.5	1.03	0.3
11	0.99	0.19	1.17	1.7	1.17	1.5
12	1.24	0.2	1.08	0.8	1.14	1.3
13	−1.14	0.17	0.88	−1.8	0.88	−1.2
14	−0.7	0.16	1.05	0.8	1.1	1
15	−0.65	0.16	0.74	−4.9	0.67	−3.6
16	0.81	0.18	0.79	−2.5	0.65	−3.9
17	1.28	0.2	1.16	1.4	1.65	5.3
18	−0.61	0.17	1.12	2	1.15	1.4
19	−1.12	0.17	0.97	−0.4	0.99	0
20	0.43	0.17	0.89	−1.6	0.89	−1.1
21	−0.8	0.17	0.88	−2	0.87	−1.3
22	−0.25	0.17	0.9	−1.8	0.84	−1.6
23	0.31	0.17	0.9	−1.4	0.88	−1.2
24	−0.52	0.16	1	0.1	1	0
25	0.4	0.17	1	0.1	1	0.1
26	0.33	0.17	1.13	1.7	1.17	1.6
27	0.84	0.19	1.05	0.6	1.11	1.1
28	2.03	0.25	0.99	0	1.14	1.3
29	0.42	0.17	0.87	−1.8	0.91	−0.9
30	0.78	0.18	1.03	0.4	1.16	1.5

Specifically, item 9 asks students to determine the current in salt water by using the drift velocity and the number of ionic charges therein. Although this question targets a content-relevant topic of polarization in an ionic solution, it requires students to formulate the answer in mathematical symbols. Perhaps this math component makes the question deviate from what it is originally intended for. The other question that shows an underfit is item 17. A quick glance at this question does not flag any problematic issues: it tests a key concept in the electromagnetism domain—electric potential in an open circuit—and it does not require nonphysics-related knowledge. A closer look at Table I shows that the infit statistics are in the acceptable range, but the outfit statistics fail to meet the requirement. This suggests that students may have made careless mistakes or lucky guesses in answering the question [41]. Indeed, nearly 50% of the students in the top quintile (according to Rasch-generated ability estimates) mistakenly chose zero as an answer. In other words, these students overlooked the battery in the circuit and solely focused on the open part of the circuit. Conversely, students in the bottom two quintiles had a correct rate of 12%, close to the overall 19% average success rate. These students may have guessed correctly on this question or had previously encountered a similar question and thus memorized the answer.

B. Rasch analysis of BEMA Construct: Unidimensionality and local independence

While the fit statistics suggest BEMA items, in general, can hold together as a meaningful measurement of one construct, the unidimensionality assumption of Rasch analysis needs to be further verified. We used Bejar's approach to compare two sets of item difficulty parameters based on the entire BEMA and a subset of the BEMA items, respectively. For a rigorous evaluation, we followed Bejar's recommendation [20,43] to split the items into two areas of most dissimilar content: electricity and magnetism. Of course, one can choose to split the items in numerous other ways. However, the more dissimilar the items are between two subsets, the more useful information can be revealed. On BEMA, the first 19 items (item 1–item 19) target electricity concepts and the remaining items (item 20–item 30) target magnetism or electromagnetic induction concepts. We estimated item difficulty parameters separately for these two subsets of questions and then compared them with those based on the entire BEMA. Figure 3(a) shows a scatter plot of total-test-based versus subtest-based estimates for the electricity items, and Fig. 3(b) shows a similar plot for the magnetism items.

The plotted dots for the 19 electricity questions in Fig. 3(a) lie near the identity line (solid line with a slope angle of 45°). A linear regression of these dots, namely, a regression axis, yields a line with a slope of 1.01 (a slope angle of 45.3°) and an intercept of 0.22, nearly parallel to the identity line. Similarly, the dots for the 11 magnetism questions in Fig. 3(b) also locate near the identity line, forming a regression axis with a slope of 1.12 (a slope angle of 48.2°) and an intercept of −0.41. According to Bejar, unidimensionality should result in a close parallelism between the principle axis and the identity line. Based on Fig. 3, there seems to be no evidence to support the hypothesis that the unidimensionality assumption is violated.

We also evaluated the local independence assumption by using Yen's methods. The correlations of Rasch residuals between each item were calculated. Yen and Fitzpatrick [45] considered a residual correlation $|r| < 0.2$ as an indication of local independence. In the present study, nine correlations (out of a total of 435 item-pair correlations) were found to fall outside of this range: three of them are greater than 0.2 and six are less than −0.2 (see Table II). While overall the local independence assumption appears to hold for the BEMA items, these nine residual correlations warrant further investigation (see Sec. V A).

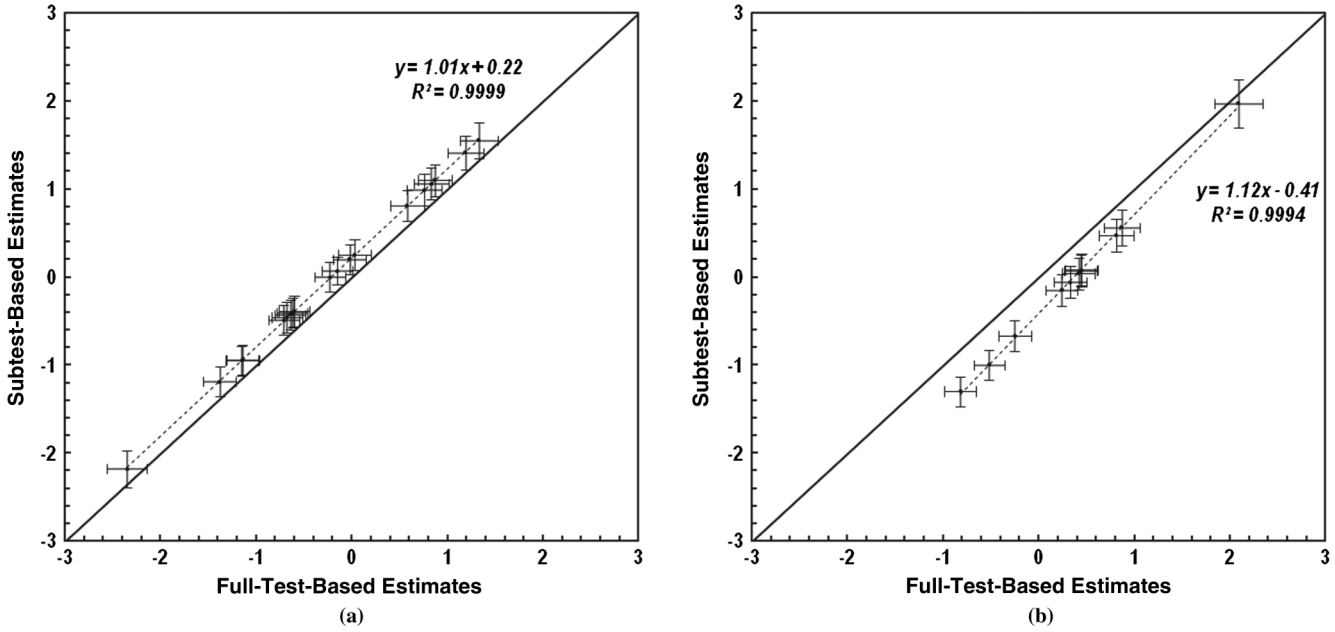

FIG. 3. Scatter plots for full-test-based versus subtest-based BEMA item estimates. (a) Scatter plot for electricity items (item 1–item 19). (b) Scatter plot for magnetism items (item 20–item 30).

C. Qualitative analysis of BEMA construct: Bloom's content and cognition levels

Based on the above Rasch analyses, there seems to be no strong evidence suggesting that BEMA items cannot function together to measure the same construct. But a lingering question is, what is this construct? As mentioned by the BEMA designers, this assessment is meant to test student understanding of key electricity and magnetism topics [10]. Though it may be true, this vague description of BEMA offers little information as to what exactly it is intended to measure. A more detailed account of the construct of BEMA is needed for better interpretation of assessment outcomes. We used the revised two-dimensional Bloom's taxonomy to classify BEMA items [46,47].

We categorized each item twice along the content and cognition dimensions, respectively (see Fig. 1). A panel of two physics education researchers and one physicist independently classified all the items using the taxonomy [47]. During the initial classification, it was found that both the conceptual and procedural knowledge categories can capture the content of BEMA items. According to Krathwohl [47], conceptual knowledge is defined as "the interrelationships among the basic elements within a larger structure that enable them to function together," for example, "knowledge of principles and generalizations" or "knowledge of theories, models, and structures." Krathwohl also defined procedural knowledge as "methods of inquiry and criteria for using skills, algorithms, techniques and methods," for example, "knowledge of subject-specific skills and algorithms" or "knowledge of criteria for determining when to use appropriate procedures." The panel recognized that BEMA items require students not only to know the meaning of physics laws and principles but also to know when and how to use them in a logical way to answer the questions. Therefore, the conceptual and procedural categories should go hand in hand in the analysis of BEMA items. We therefore combined the two content categories for use in this study.

TABLE II. BEMA item pairs with residual correlations $|r| > 0.2$.

Residual correlation $r > 0.2$		Residual correlation $r < -0.2$	
Correlation	Item pair	Correlation	Item pair
0.34	Q21–Q22	−0.26	Q18–Q21
0.34	Q15–Q16	−0.24	Q21–Q26
0.21	Q2–Q3	−0.24	Q9–Q15
		−0.23	Q9–Q21
		−0.21	Q16–Q26
		−0.21	Q12–Q22

The three panel members independently categorized all items, and then classification notes were compared to check interrater reliability. The initial agreement between all panel members before discussion was 90% for the content categorization and 87% for the cognition categorization. For the remaining cases, there was always an agreement between two of the three panel members and the third disagreed by only one level. The divergences were then discussed among the panel and were eventually resolved. As a result, 23 items (item 1–item 23) were categorized as requiring students to apply an E&M principle by carrying out the application procedures in a specific context (apply concepts or procedures). The other seven items (item

24–item 30) were classified as requiring students to synthesize both electricity and magnetism concepts for analysis in a complex system (analyze concepts or procedures). (Also see Supplemental Material [22].) Using the Rasch-generated item measures, we further compared the difficulty of these two categories of items. Overall, the mean difficulty estimates for items that require "application of concepts or procedures" and "analysis of concepts or procedures" are -0.19 [standard error(SE) $= 0.19$] and 0.61 (SE $= 0.29$), respectively, with the former being statistically lower than the latter at the 4% error rate ($p < 0.04$, effect size $d = 0.92$).

D. Evaluation of DIF in BEMA

While the emerging evidence allows us to make inferences about the construct of BEMA, it is still unclear whether or not we can use this assessment to compare different E&M courses and draw valid conclusions. We used data from the traditional and M&I courses to seek evidence for DIF (potential bias) in BEMA. To establish baseline information, we compared Rasch-generated person ability estimates between the two courses (see Table III). For the pretest, no significant between-group difference is detected, but for the posttest there is a significant difference. This means students in both courses started at a similar performance level, but the M&I students finished with a higher level than those in the traditional course—a result consistent with what was reported in the literature [18]. In this case, the validity of the difference in the posttest becomes our major concern, because it could have been due to the potential bias in BEMA that was in favor of the M&I curriculum.

We reestimated BEMA item parameters by using the data from the traditional and M&I courses separately. Since the difference in the posttest is our main concern and the pretest data lacks a sufficient person-item match (Fig. 2), we used the post data for DIF analysis. The two sets of item difficulty parameters (estimated based on the traditional and M&I courses, respectively) are presented in a scatter plot as shown in Fig. 4. With measurement error in mind, we plotted 95% confidence bands (dotted curves) and 99% confidence bands (solid curves) in the plot (also see Ref. [13]). The dots within these bands represent items

TABLE III. Pre- and posttest person ability estimates for students in the traditional and M&I physics courses.

	Person ability mean value (± standard deviation)		
	Traditional course	M&I course	Two-sample t test p value
Pretest	$-1.46(\pm 0.59)$	$-1.59(\pm 0.59)$	$p = 0.11$
Posttest	$-0.84(\pm 1.03)$	$-0.28(\pm 1.21)$	$p < 0.003$

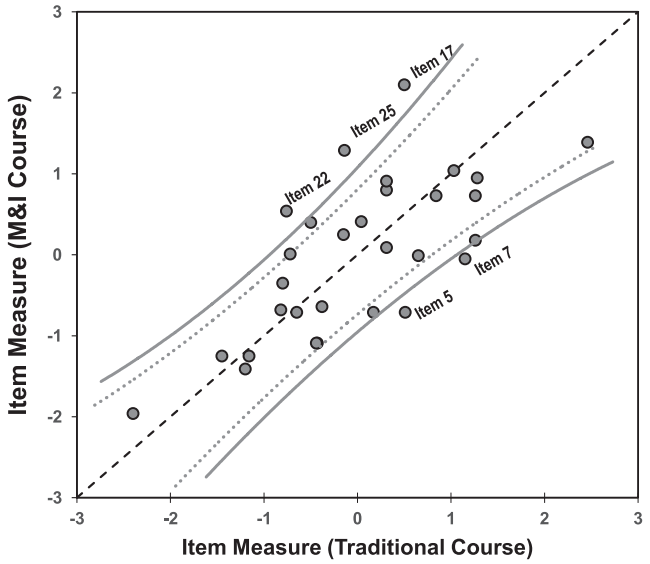

FIG. 4. A scatter plot of BEMA item difficulty estimates based, respectively, on data from the traditional and M&I courses. Dotted curves are 95% confidence bands, solid curves are 99% confidence bands.

of similar functioning for the students in different courses. Those that depart from the bands are items with a DIF and need to be examined. In this plot, the majority of the dots are within or in the immediate vicinity of the 95% confidence bands (specifically within the 99% bands), suggesting no significant DIF in these items. Five dots fall out of the 99% confidence bands, signaling a significant DIF in these items. Among them, two (item 5 and item 7) are located below the lower-limit band and therefore are in favor of the M&I students, the other three (item 17, item 22, and item 25) are above the upper-limit bands, hence, in favor of the students in the traditional course. Additionally, we examined the effect sizes of DIF, namely, DIF contrasts, by taking the difference in item estimates between the two groups [41]. It was found that the DIF contrasts for the items falling out of the 99% bands were at least 1.12 in size ($|\mathrm{DIF}|_{\mathrm{item}\,5} = 1.17$, $|\mathrm{DIF}|_{\mathrm{item}\,7} = 1.12$, $|\mathrm{DIF}|_{\mathrm{item}\,17} = 1.57$, $|\mathrm{DIF}|_{\mathrm{item}\,22} = 1.24$, and $|\mathrm{DIF}|_{\mathrm{item}\,25} = 1.36$). For the remaining items, the DIF contrasts were all immediately near or below 1, with seven of them displaying a moderate size with $|\mathrm{DIF}| \geq 0.64$ (see Ref. [41]).

To further explicate the DIF in these items, we divide the students from each course into five quintiles according to their Rasch ability estimates. For each item, the proportions of correct responses in each quintile are plotted as a function of person ability (see Fig. 5). For the two items in favor of the M&I students (item 5 and item 7), the curve of the M&I course lies higher than that of the traditional course. Conversely, for the three items in favor of the traditional course (item 17, item 22, and item 25), the pattern is reversed. These plots reveal the ability levels at which each item functions differentially for the two groups

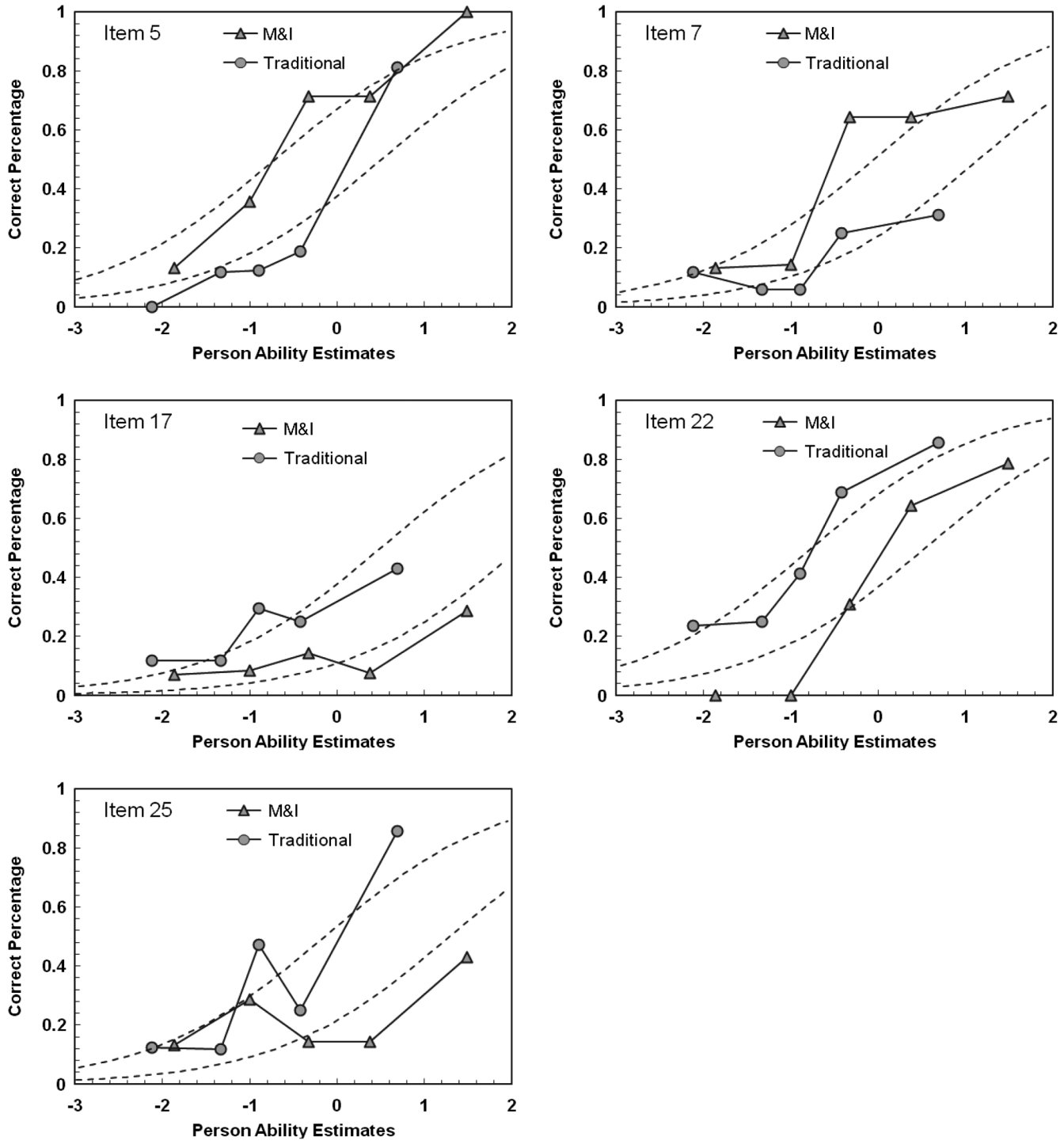

FIG. 5. Item characteristic curves (showing proportion of correct responses as a function of person ability) for BEMA items with DIF. Solid lines are empirical results based on the data from the traditional and M&I courses respectively, dashed lines are modeled results.

of students. For example, item 22 displays a consistent DIF between the two groups of students regardless of the ability levels. Alternatively, the DIF in item7 comes mostly from the difference in the high ability region.

In addition to seeking empirical evidence for differential functioning at the item level, we also evaluated possible differential functioning at the assessment level. One approach is to plot students' total scores as a function of their ability levels estimated separately for the two courses and then examine the deviation between the two plots [52]. If the two plots overlap, there is no differential functioning at the assessment level. Otherwise, the assessment as an entirety functions differentially for different groups of students. Figure 6 shows the two plots for the traditional

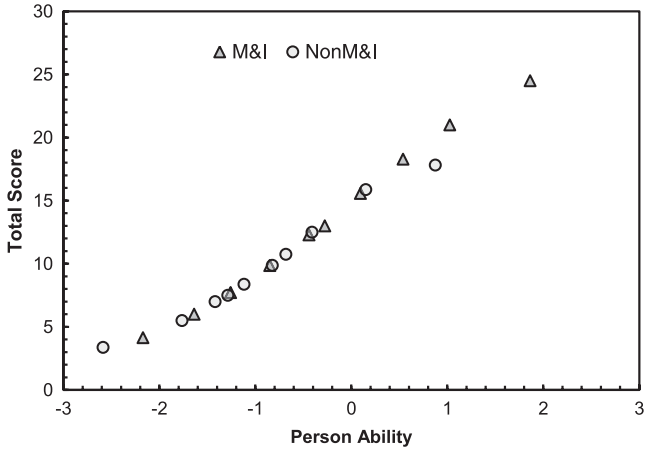

FIG. 6. BEMA test characteristic curves (showing total scores as a function of person ability estimates) based on the data from the traditional and M&I courses, respectively.

and M&I courses, respectively. Here, the two curves virtually fall onto the same *S*-shaped curve.

V. DISCUSSION

A. Construct of BEMA

The above results derived from Rasch analysis suggest that BEMA items, albeit testing a broad range of topics, can, in general, hold together to collectively measure the same construct. That said, two items fail to fit into this construct. One item (item 9) requires students to formulate answers in mathematical notations—a possible confounding factor in a physics concept survey. Therefore, revisions aimed at removing this factor are recommended. The other item (item 17) represents a situation in which students may have made careless mistakes due to some unknown reasons. A close monitoring of this item in future studies will be useful. Besides checking the fit statistics, the unidimensionality of BEMA is further evaluated by using Bejar's approach. No evidence suggests that this assumption is violated. Moreover, an evaluation of the Rasch residual correlations indicates that the local independence assumption by and large is supported. Thus, the model we used can satisfactorily explain the relationships among items; or simply put, the assumption that BEMA items are related by a shared construct holds [45].

Nevertheless, nine pairs of items show a stronger association in their residuals than expected (Table II). Among them, three have a strong positive association ($r > 0.2$) and six have a strong negative association ($r < -0.2$). The three positive correlations suggest that there may exist between each pair of items some common factor extraneous to the shared construct of the assessment. A closer look at these items reveals that these are all consecutive questions that share the same question stems and diagrams. It is possible that this commonality may have led to the positive residual correlations. Future studies are recommended to separate these items either using different question stems and diagrams or placing them at different locations of the assessment. As such, the local independence can be more accurately tested to examine the presence (or absence) of unintended factors among them. Conversely, the six negative correlations indicate that there may be some inherent differences between the items in each pair. In fact, for all of the six pairs, each contains one magnetism question and one electricity question (or a synthesis of electricity and magnetism as seen in item 26). Therefore, the content dissimilarity between the paired items may account for the negative residual correlations. This result seems to suggest that magnetism and electricity items may not fit well together. In light of this possibility, the totality of the acquired evidence is examined to draw credible inferences. Recall that, in testing the unidimensionality assumption, we split BEMA into two subsets of electricity and magnetism items, and the results support one construct. In terms of the local independence assumption, while six out of total 435 pairs show a strong negative association, the majority are in the desired range. In addition, the Rasch fit statistics yield no evidence for a large number of underfits. Combining all of these considerations, it is unlikely that separating electricity items from magnetism items is warranted. Given the above emerging evidence, it is therefore reasonable to infer that a single aggregate score on BEMA can be used to represent student understanding of E&M topics.

In the present study, the construct of BEMA, which was vaguely described by its designers as "understanding of basic electric and magnetism concepts," is qualitatively explicated through Bloom's taxonomy. Specifically, BEMA is aimed at testing students' proficiency in applying or analyzing E&M principles through logical procedures to predict various electromagnetic phenomena. As manifested by their average item difficulty measures, questions that require students to perform analysis or synthesis, in general, are more challenging than those that require only application, and the difference in their difficulty measures is of a large size ($d = 0.92$). This result is consistent with the hierarchical nature of Bloom's taxonomy (see Fig. 1).

B. Potential bias in BEMA items

Based on the Rasch measures, no significant DIF is detected for the majority of BEMA items. This means that most items function similarly for students who have the same ability levels regardless of which course they attended. However, our analysis also reveals possible evidence of DIF for five items. Among them, two are in favor of the M&I course and the other three are in favor of the traditional course.

Typically, DIF can be attributed to two primary causes. One is due to the different opportunities that students have of being exposed to the tested content, and the other lies in the context in which questions are situated. In the present

study, there is strong evidence suggesting that the two courses provided comparable opportunities for students to learn the tested topics in terms of instruction time, course content, and teacher quality—the three-pronged conceptualization of learning opportunities [31]. More importantly, BEMA was intentionally designed to be a common-denominator assessment to test key E&M topics that students in both traditional and M&I classes would have comparable opportunities to access in their respective courses [10,18]. Therefore, possible causes of the detected DIF in the present study likely rest with the contexts in which these items are situated. To some extent, DIF can also be considered as an indication of possible extraneous factors in questions that may have caused differential outcomes for students with comparable ability levels; or simply put, there may be some subtlety in the item design that interferes with the functioning of the item [16].

Take the two items in favor of M&I, for example, item 5 and item 7. Item 5 requires students to apply the supposition principle to determine an electric field and is presented in the case of an electric dipole. Item Q7 requires students to determine the polarization in an insulator produced by an external electric field. Both items are posed in a context highly familiar to the M&I students. Specifically, electric dipoles are often used as a scenario in practice problems in the M&I course, whereas the same setting in the traditional course is primarily invoked as a special example of the superposition of electric fields. Since the purpose of item 5 is to test student application of the superposition principle, perhaps a context other than electric dipoles can be tried to mitigate the possible DIF. Similarly, the context of item 7 (particularly the diagram presented therein) highly resembles the context of Scotch tape polarization examples that are frequently used in the M&I course [30]. Conceivably, when answering these questions, the M&I students likely have an advantage over those in the traditional course.

However, the potential DIF of the three items in favor of the traditional course is difficult to understand. A further examination of the curricula materials for both courses provides no interpretable account (neither in terms of course content nor student familiarity with item contexts). One postulation, however, is that the students in the traditional course might have encountered similar questions not long before taking the BEMA, and therefore were likely to succeed on these items. Item 22, perhaps, is such an example, since students at all levels in the traditional course outperformed their M&I counterparts of the same person ability (Fig. 5). Another alternative explanation is that the M&I students made careless mistakes or a lucky guess in responding to these items. For example, when answering item 17, nearly 60% of the M&I students (as opposed to 40% of those in the traditional course) chose "0 volt." As discussed earlier (see Sec. IV A, last paragraph), these students likely overlooked the battery and focused solely on the "open" part of the circuit. In light of the data we have, it is nevertheless unclear why the M&I students would have made such a mistake.

Additionally, the differential functioning at the assessment level is evaluated. The plots of total score versus person ability are compared between the M&I and traditional courses. The fact that the two curves fall on the same *S*-shaped curve suggests that, despite the detected DIF in a few BEMA items, the entire assessment functions similarly for the two groups of students. Based on this evidence and the aforementioned inferences about BEMA construct, it is reasonable to conclude that the higher posttest performance of the M&I students on BEMA is unlikely caused by bias in the assessment. Since both groups of students started with a similar preinstructional ability measure, we can now more confidently attribute the statistically better posttest measure of the M&I students to the instruction they received in the course. Clearly, the M&I course improved students' overall ability of applying and analyzing various electromagnetic phenomena (the construct of BEMA), not just increased their performance on specific topics due to potential biases (as otherwise would be manifested by DIF in a large number of items).

As noted earlier and demonstrated in the above analysis, detection of DIF often requires both quantitative and qualitative inspection of the items of interest. A statistically significant DIF can be a sign of item bias but certainly does not guarantee it. In order to properly infer item bias, a careful analysis of item content and context is needed. In our study, the uniqueness of content similarity between the two courses serves as an anchor for our subsequent analysis regarding the contextual issues pertinent to item bias. Had we chosen for investigation some other physics courses that differed in pedagogy or learning goals, the DIF results perhaps would have been different. This is not to say that we must hastily change the assessment every time we identify DIF. Instead, as our study has illustrated, inferences and decisions derived from evidence-based reasoning and supported by both quantitative and qualitative analysis often can be more beneficial in the long run.

C. Significance and implications

While this study focuses primarily on the technicality of BEMA and provides long-missing but much-needed justifications for its use as a measurement tool, the significance goes beyond the targeted assessment. As discussed earlier, science concept assessments can be designed into different types. Depending on the breadth of its content coverage, an assessment can test either a narrow topic (in the case of concept inventories) or a broad range of topics (in the case of concept surveys) [9]. When it comes to the latter, extra caution must be taken to empirically verify the existence of a single construct [14]. Otherwise, the action of using an aggregate score to represent student performance on the assessment is not warranted. Even when mounting evidence is accumulated to support an overall single

construct at the assessment level, there may still be item-level or even finer-level issues that are in discordance with the overall construct, such as item locations, content variations, and contextual features. These issues are likely to be revealed through a set of evaluations like those we conducted with BEMA. To this end, it is important to weigh both the totality and the individuality of evidence to draw a balanced inference. It is useful to remember that no assessment is perfect. The more closely one inspects an assessment, the more issues one will discover. However, these issues, at the very least, can provide key guidance to the effective revisions of assessment items.

Also illustrated in this study is the significance of using available evidence from assessments to make valid inferences about learning and teaching. For example, had we not sought potential biases in BEMA, we would have overlooked the differential functioning in some items. Although there is no indication to nullify the overall better performance of the M&I students, seeking and documenting this empirical evidence is crucial for making credible arguments. In this study, the reported evidence allows our heightened confidence in the effectiveness of the M&I curriculum in promoting students' core content knowledge. Of course, there are other confounding factors that may need additional investigation or technical issues that can be further improved. For instance, the sample sizes in the current study were relatively small, which may have limited the analysis power of detecting DIF with moderate size. Nevertheless, the evaluation practices like those presented in this paper can undoubtedly help us accumulate evidence for proper interpretation and use of educational assessments. With increasing evidence at hand, our confidence in the extent to which our inferences and conclusions are valid will also increase. After all, educational assessment is an "evidence-based reasoning" process, and it indeed, as aptly stated by Messick [53], is "an integrated evaluative judgment of the extent to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment."

ACKNOWLEDGMENTS

The author wishes to thank William Boone, Irene Neumann, and Knut Neumann for useful discussions. We also thank the three reviewers for their insightful comments that have led to significant improvement of the manuscript. This study is partially supported by the National Science Foundation (NSF Grant No. DRL 1252399).

[1] National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment* (National Academy Press, Washington, DC, 2001).
[2] N. B. Songer and M. A. Ruiz-Primo, Assessment and science education: Our essential new priority?, J. Res. Sci. Teach. **49**, 683 (2012).
[3] J. W. Pellegrino, Assessment of science learning: Living in interesting times, J. Res. Sci. Teach. **49**, 831 (2012).
[4] M. T. Kane, in *Educational Measurement*, edited by R. L. Brennan, (Praeger, Westport, CT, 2006), pp. 17–64.
[5] E. H. Haertel and W. A. Lorie, Validating standards-based test score interpretations, Meas. Interdiscip. Res. Perspect. **2**, 61 (2004).
[6] American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), *Standards for Educational and Psychological Testing* (American Psychological Association, Washington, DC, 1999).
[7] R. J. Mislevy, L. S. Steinberg, and R. G. Almond, On the Structure of Educational Assessments, Meas. Interdiscip. Res. Perspect. **1**, 3 (2003).
[8] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[9] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).
[10] 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).
[11] W. J. Boone and K. Scantlebury, The role of Rasch analysis when conducting science education research utilizing multiple-choice tests, Sci. Educ. **90**, 253 (2006).
[12] X. Liu, *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach* (Information Age Publishing, Charlotte, NC, 2010).
[13] T. G. Bond and C. M. Fox, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Routledge, New York, 2007), 2nd ed.
[14] M. Strauss and G. Smith, Construct validity: Advances in theory and methodology, Annu. Rev. Clin. Psychol. **5**, 1 (2009).
[15] G. Camilli, in *Educational Assessment*, edited by R. L. Brennan (Praeger, Westport, CT, 2006), pp. 221–256.
[16] S. J. Osterlind and H. T. Everson, *Differential Item Functioning* (Sage Publications, Thousand Oaks, CA, 2009), 2nd ed.
[17] S. J. Pollock, Longitudinal study of student conceptual understanding in electricity and magnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020110 (2009).
[18] M. A. Kohlmyer, M. D. Caballero, R. Catrambone, R. W. Chabay, L. Ding, M. P. Haugan, M. J. Marr,

B. A. Sherwood, and M. F. Schatz, Tale of two curricula: The performance of 2000 students in introductory electromagnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020105 (2009).

[19] P. M. Sadler, H. Coyle, J. L. Miller, N. Cook-Smith, M. Dussault, and R. R. Gould, The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 national science standards, Astron. Educ. Rev. **8**, 010111 (2009).

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

[21] J. M. Bailey, Concept inventories for ASTRO 101, Phys. Teach. **47**, 439 (2009).

[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevSTPER.10.010105 for Bloom's levels of BEMA items, and the topical sequences of electricity and magnetism in traditional and M&I courses.

[23] O. John and V. Benet-Martinez, in *Handbook of Research Methods in Social and Personality Psychology*, edited by H. Reis and C. Judd (Cambridge University Press, Cambridge, England, 2000), pp. 339–369.

[24] S. Downing, Validity: On the meaningful interpretation of assessment data, Med. Educ. **37**, 830 (2003).

[25] F. Seroglou, P. Koumaras, and V. Tselfes, History of science and instructional design: The case of electromagnetism, Sci. Educ. **7**, 261 (1998).

[26] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).

[27] T. F. Scott and D. Schumayer, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).

[28] B. Clauser and K. Mazor, Using statistical procedures to identify differential functioning test items, Educ. Meas. Issues Prac. **17**, 31 (1998).

[29] M. Michaelides, A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in item equating, Front. Psychol. **1**, 1 (2010).

[30] R. Chabay and B. Sherwood, *Matter & Interactions, Volume II: Electric and Magnetic Interactions* (John Wiley & Sons, Hoboken, NJ, 2011).

[31] A. Kurz and S. Elliott, in *Assessing Student in the Margin: Challenges, Strategies, and Techniques*, edited by M. Russell and M. Kavanaugh (Information Age Publishing, 2011), pp. 31–58.

[32] R. Beichner, R. Chabay, and B. Sherwood, Labs for the Matter & Interactions curriculum, Am. J. Phys. **78**, 456 (2010).

[33] R. Chabay and B. Sherwood, Restructuring the introductory electricity and magnetism course, Am. J. Phys. **74**, 329 (2006).

[34] I. Neumann, K. Neumann, and R. Nehm, Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test, Int. J. Sci. Educ. **33**, 1373 (2011).

[35] M. Wilson, *Constructing Measures* (Taylor & Francis, New York, 2005).

[36] W. J. Boone, J. S. Townsend, and J. Staver, Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data, Sci. Educ. **95**, 258 (2011).

[37] B. D. Wright and J. Linacre, M., Observations are always ordinal; measurements, however, must be interval, Arch. Phys. Med. Rehabil. **70**, 857 (1989).

[38] B. D. Wright, A history of social science measurement, Educ. Meas. Issues Prac. **16**, 33 (1997).

[39] D. Borsboom and A. Z. Scholten, The Rasch model and conjoint measurement theory from the perspective of psychometrics, Theory Psychol. **18**, 111 (2008).

[40] A. A. Rupp and B. Zumbo, Understanding parameter invariance in unidimensional IRT models, Educ. Psychol. Meas. **66**, 63 (2006).

[41] J. M. Linacre, *A User's Guide to Winsteps, Rasch Measurement Program* (MESA Press, Chicago, 2009).

[42] R. Childs and S. Oppler, Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high-stakes testing program, Educ. Psychol. Meas. **60**, 939 (2000).

[43] I. Bejar, A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates, J. Educ. Measure. **17**, 283 (1980).

[44] W. M. Yen, Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, Appl. Psychol. Meas. **8**, 125 (1984).

[45] W. M. Yen and A. Fitzpatrick, in *Educational Measurement*, edited by R. L. Brennan (Praeger, Westport, CT, 2006), pp. 111–153.

[46] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Longman, New York, 2001).

[47] D. R. Krathwohl, A revision of Bloom's taxonomy: An overview, Theory Into Practice **41**, 212 (2002).

[48] L. Ding, R. Chabay, and B. Sherwood, How do students in an innovative principle-based mechanics course understand energy concepts?, J. Res. Sci. Teach. **50**, 722 (2013).

[49] R. E. Teodorescu, C. Bennhold, G. Feldman, and L. Medsker, New approach to analyzing physics problems: A taxonomy of introductory physics problems, Phys. Rev. ST Phys. Educ. Res. **9**, 010103 (2013).

[50] L. Ding and X. Liu, in *Getting Started in PER—Reviews in PER*, edited by C. Henderson and K. Harper (American Association of Physics Teachers, College Park, MD, 2012), Vol. 2, pp. 1–42.

[51] J. S. Aslanides and C. M. Savage, Relativity concept inventory: Development, analysis, and results, Phys. Rev. ST Phys. Educ. Res. **9**, 010108 (2013).

[52] N. Raju and B. Ellis, in *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis*, edited by F. Drasgow and N. Schmitt (Jossey-Bass, San Francisco, 2002), pp. 156–188.

[53] S. Messick, in *Educational Measurement*, edited by R. L. Linn (MacMillan, New York, 1989), 3rd ed., pp. 13–103.