Multidimensional Item Response Theory and the Conceptual Survey of Electricity and Magnetism

Cabot Zabriskie and John Stewart*

Department of Physics and Astronomy, West Virginia University, Morgantown, West Virginia 26506, USA

(Received 22 June 2018; published 3 July 2019)

[This paper is part of the Focused Collection on Quantitative Methods in PER: A Critical Examination.] While many studies have examined the structure, validity, and reliability of the Force Concept Inventory, far less research has been performed on other conceptual instruments in widespread use in physics education research. This study performs a confirmatory analysis of the Conceptual Survey of Electricity and Magnetism (CSEM) guided by a theoretical model of expert understanding of electricity and magnetism. Multidimensional Item Response Theory (MIRT) with the discrimination matrix constrained to the theoretical model was used to investigate two large datasets ($N_1 = 2014$ and $N_2 = 2657$) from two research universities in the United States. The optimal model identified by MIRT was similar, but not identical, for the two datasets and had very good model fit with comparative fit indices of 0.975 and 0.984, respectively. The most parsimonious optimal model required 23 independent principles of electricity and magnetism and was significantly better fitting than a more general model dividing the CSEM into 6 general topics. The optimal models for the two samples were quite similar, sharing 22 of a possible 26 conceptual principles. Most of the overall item difficulties and discriminations were significantly different between the two samples; however, the rank order of the overall difficulty and discrimination were generally similar. There was much more similarity between the discrimination by item of the individual principles. Five items had a difficulty ranking that was substantially different between the two samples, indicating that while generally similar, relative difficulty does depend on the student population and instructional environment.

DOI: 10.1103/PhysRevPhysEducRes.15.020107

I. INTRODUCTION

The Conceptual Survey of Electricity and Magnetism (CSEM) was introduced nearly 20 years ago and has become one of the most used conceptual instruments for understanding conceptual knowledge of electricity and magnetism [1]. The CSEM was developed following the success of the Force Concept Inventory (FCI) [2] in demonstrating the ineffectiveness of traditional instruction in fostering conceptual learning [3]. Like the FCI, the CSEM was developed to test student misconceptions as well as their physics knowledge. The authors further intended the instrument to serve as a broad summary of student learning in electricity and magnetism rather than a granular measure of student understanding [1]. The other instrument commonly used to measure student conceptual learning in electricity and magnetism is the 30-item Brief Evaluation of Electricity and Magnetism (BEMA) [4]. Pollock found the CSEM and BEMA to be equally effective for evaluating conceptual learning with slight variations in the content covered by each instrument [5]. These and other conceptual instruments have been very important in physics education research (PER) and a general overview of the use of these and other conceptual inventories can be found in the recent synthesis of PER by Docktor and Mestre [6].

The current study explored the structure of the CSEM using Multidimensional Item Response Theory (MIRT) for 2 samples of students in university calculus-based physics. This work replicates a similar study on the FCI, but uses datasets from different institutions to determine whether MIRT results can be considered general [7]. The previous study as well as a large body of social science research [8,9] argued that exploratory methods, where one does not begin with a theoretical model and develops the model from the data, such as factor analysis, identify accidental features of the data and do not provide generalizable results. Conversely, a confirmatory analysis begins with a theoretical model and determines how well the data support the model. As such, this study as well as the previous study of the FCI presents confirmatory analyses beginning with introduction of a theoretical model of the knowledge measured by the instrument. Because the majority of research exploring the

*jcstewart1@mail.wvu.edu

structure of conceptual physics instruments has been exploratory and no such research has been performed on the CSEM, an exploratory analysis is presented in the Supplemental Material [10] to address this absence in the literature. MIRT can be used for either exploratory or confirmatory analyses; for exploratory analysis the MIRT parameters are unconstrained, for confirmatory analysis the MIRT parameters are constrained to the model. Little previous research has compared the results of either exploratory or confirmatory analyses across multiple institutions, so this work should advance the understanding of the generality of MIRT analyses.

The previous study of the FCI showed that MIRT models could be constrained to a theoretical model of Newtonian mechanics and used to explore theoretically motivated modifications of the model. The optimal model identified contained only principles of physics such as Newton's 1st law and the definition of acceleration. The optimal model did not contain qualitative statements about mechanics derived from these statements; for example, "if the acceleration and the velocity are in the same direction, the object speeds up." Testing alternate models allowed for an exploration of how students answered mechanics questions; for example, model fit did not improve when Newton's 2nd law was combined with the vector addition of forces; illustrating that students have different facility with these two mechanics principles.

A. Reliability and validity

The structure, reliability, and validity of PER conceptual instruments is an active area of research; however, most of this research has focussed on the FCI. Few studies have analyzed the item-level validity, reliability, or fairness of the CSEM.

Classical test theory (CTT) provides methods to examine item validity though the calculation of difficulty and discrimination. The difficulty of an item is defined as its average score; a higher difficulty score indicates an easier item and a lower difficulty score a harder item. The discrimination is defined as the difference in the average score of the highest performing students and the lowest performing students. Items with either very high or very low difficulty or low discrimination are "problematic" and present validity threats to the instrument [11,12].

Maloney *et al.* reported CSEM item-level difficulty for both algebra-based and calculus-based introductory electricity and magnetism courses [1]. Their study found four problematic items with item 3 too easy for calculus-based students (difficulty above 0.8) and items 14 (calculus and algebra-based students), 20 (algebra-based students), 31 (algebra-based students) too difficult (difficulty below 0.2). Though item discrimination was evaluated in the study, the item-level results were not reported. Planinic identified six conceptual areas measured by the CSEM in a study comparing Croatian students to American students. These were electric charge and force (items 1-3, 5, 6, and 8), electric field and force (items 9 and 12-15), electric potential and energy (items 11 and 16-20), magnetic field and force (items 21-23, 25, 26, and 28), electromagnetic induction (items 29-32), and Newton's laws (items 4, 7, 10, 24, and 27). The conceptual areas were identified qualitatively by grouping the 11 conceptual areas identified by Maloney *et al.* [1] to produce groups of items sufficiently large for analysis. The difficulty of the items in each conceptual area was calculated finding similar results for both populations [13].

Other studies have focused on only a few items in the CSEM. Meltzer investigated items 18 and 20 to explore changes between pretest and post-test responses regarding the intersection of electric field and potential concepts [14]. Leppävirta investigated CSEM items that probed Newton's 3rd law (items 4, 5, 7, and 24) showing that 20% of students had an alternative model of Newton's 3rd law on the pretest which was reduced to 10% on the post-test [15].

Gender differences in performance on the CSEM have also been investigated. Kohl and Kuo [16] examined the difference in the gender gap on the CSEM before and after switching to studio physics, finding the course transition reduced the gap in normalized gain. Studio physics is an instructional model that integrates short lectures with group work and hands-on activities in a classroom where students are grouped around tables [17]. Kreutzer and Boudreaux [18] also measured a difference in performance by gender in the CSEM. Pedagogical changes such as "affirming domain belongingness in women" greatly reduced the gap. For a more complete synthesis of the study of gender and conceptual inventories see Madsen, McKagan, and Sayre [19].

B. Factor analysis

Extensive work on the factor structure of PER conceptual instruments has been performed; however, the majority of this work has investigated the FCI [7,20–25] and the Force and Motion Conceptual Evaluation (FMCE) [26,27].

Beyond the initial factor analysis conducted by Maloney *et al.* [1] when constructing the CSEM, additional work exploring the factor structure of the instrument has not been reported. The factor structure found by Maloney *et al.* was determined using principle component analysis and found an optimal 11-factor model of the instrument. This model was discarded as containing too many factors with too little variance explained by each. The 11-factor model structure was not reported.

The majority of the factor analyses have been performed on the FCI and have used exploratory methods [exploratory factor analysis (EFA)]. The EFA studies failed to establish a single structure of the instrument and resulted in factor models with 5 factors [23], 6 factors [25], and 9 factors model [7]. An early model produced by Huffman and Heller with a very conservative factor selection criteria

identified only 1 factor [20]. There were some similarities between the 5, 6, and 9 factor models but they were not identical. Further, examination of the individual factors showed that many did not make theoretical sense, mixing items testing very different physical concepts. Stewart *et al.* [7] demonstrated that much of the identified structure was not grounded in physical principles, but instead was due to question blocking or repetition of a few very similar items.

Ramlo explored the factor structure of the FMCE and found 3 factors for the post-test [27]. In this study, items involving similar conceptual topics largely loaded onto the same factor.

C. Item Response Theory

Many studies have employed Item Response Theory (IRT) to probe individual item performance on PER conceptual inventories; however, again, the majority of this research has focused on the FCI. Multiple unidimensional IRT models have been published [28,29] and find the FCI has generally excellent properties unlike some other published instruments in engineering education [12]. Unidimensional IRT models a student's facility with the material with a single ability parameter; MIRT uses multiple ability parameters. Unidimensional models have also been used to investigate gender fairness in the FCI and multiple items have been identified as unfair with the majority unfair to women [30,31]. A similar analysis showed that few CSEM items are substantially unfair to either men or women [32]. IRT has also been used to produce modified versions of the FCI to reduce testing time [33]. MIRT, as will be employed in this study, has been used as an alternate method of performing factor analysis on the FCI [7,24] producing similar but not identical results to traditional EFA.

IRT has also been used to examine physics problems not part of PER conceptual inventories. Changes in student understanding of physics in online learning environments [34] and how different patterns of feedback affect understanding have been explored.

D. The structure of knowledge

Experts tend to categorize conceptual problems in a more deliberate way than novices, focusing on the hierarchical structure of the knowledge starting with the most fundamental principles and branching out from there to the less fundamental principles [35–39]. This more efficient way of organizing understanding allows experts to more expediently solve physics problems from first principles [40–42]. Conversely, novices tend to focus more on the surface features of the problem and their solutions often lack the same deliberate structure of experts [35,36].

Multiple theoretical frameworks have been advanced to understand the differences in expert-novice problem solving. One model of student knowledge proposes that students categorize knowledge into "ontological categories." This provides an explanation of the prevalence of commonly held misconceptions where students miscategorized their knowledge, storing it in overly broad categories [43–45]. Another model proposes "knowledge in pieces" where knowledge lies not in broad principles, but in granular facts that are activated as needed to solve problems [46–48].

Research has shown that novice problem solutions are strongly context dependent and rely on how the current problem relates to previous problems that the student has solved [49–51]. This context sensitivity of problem solving suggests that it may be appropriate to treat novice problem solving as composed of granular knowledge pieces instead of the broad knowledge structures probed by factor analysis or cluster analysis.

The current work produced a fine-grained model of the information needed to solve CSEM problems. A similar model was created for the FCI by Stewart *et al.* [7]; these models are similar to those produced by research into complex problem solving by Simon and Newell [52]. Their research paradigm dominated problem-solving research for three decades and is summarized by Ohlsson [53]. Simon and Newell constructed models that replicated the human problem-solving sequence. This sequence was identified by examination of think-aloud transcripts. The models were sufficiently detailed that they could be converted to computer code and executed to reproduce the sequence of steps taken by the human solver. This method was productive in the understanding of problem solving in kinematics and dynamics and many other subjects [54,55]. In a related effort, Reif and Heller created a detailed model of problem solving in mechanics [40]. While not computationally executable, their model was intended to be a prescription of expert problem solving in mechanics.

The model we will construct for the CSEM has a similar structure to the computational models of Larkin *et al.* [55] and the model of Reif and Heller [40].

E. Research questions

This study seeks to answer the following research questions:

RQ1: What is the optimal model of student knowledge measured by the CSEM? Are the principles forming the optimal model consistent across samples?

RQ2: Are the parameters of the optimal models consistent between samples?

II. METHODS

A. Conceptual Survey of Electricity and Magnetism

The CSEM is a 32-question conceptual instrument designed to measure student understanding of electricity and magnetism. This instrument covers concepts often found in introductory electricity and magnetism courses such as the Coulomb force law, electric and magnetic fields, induction, and electric potential [1]. The CSEM was originally developed by Maloney *et al.* by combining

the concepts from two prior surveys from Hiegeelke and O'Kuma, which probe understanding of electricity and magnetism separately [56]. These two surveys were combined after many iterations and the resulting version of the CSEM can be found at PhysPort [57]. A complete list of the concepts the CSEM was designed to measure can be found in Maloney *et al.* [1].

B. Sample

This study will examine two samples drawn from different institutions in the United States.

Sample 1: Sample 1 was drawn from 14 semesters of calculus-based introductory electricity and magnetism courses at a large southern land-grant university serving approximately 25 000 students. The undergraduate population had ACT scores ranging from 23 to 29 (25th to 75th percentile) [58]. The institution held a Carnegie Classification of "Highest Research Activity" for all semesters studied [59]. The overall undergraduate demographics were 77% White, 8% Hispanic, 5% African American, 2% Asian with other groups each 3% or less. The sample was primarily male (77%) [58].

For the entire study, the course was comprised of two 50-min lectures per week with an additional two 2-h weekly laboratories. The CSEM was given as a laboratory quiz pre- and postinstruction with the student's scores counting toward the course grade. The course was taught and overseen by the same instructor for each of the 14 semesters included in this study. The aggregate dataset contains 2014 students who completed the course for a grade and received credit for the CSEM pretest and post-test. The dataset were also analyzed by Henderson *et al.* [32] to explore gender fairness.

Sample 2: Sample 2 was drawn from 13 semesters of calculus-based introductory electricity and magnetism courses at a large eastern land-grant university serving approximately 30 000 students. The undergraduate population had ACT scores ranging from 21 to 26 (25th to 75th percentile) [58]. The institution was first rated as highest research activity a year prior to the completion of data collection [59]. The overall undergraduate demographics were 79% White, 8% Hispanic, 6% International, 5% African American, 4% Hispanic, 2% Asian with other groups each 4% or less. The sample was primarily male (81%) [58].

Unlike sample 1, the instructional environment for sample 2 was variable. For the first 4 years of the study, the course was taught by 6 separate instructors with standings ranging from full professor to late career graduate student. For this period, the course was comprised of four 50-min lectures and a single 2-h laboratory each week. A learning assistant (LA) program was implemented to improve conceptual learning [60]. Undergraduate students, LAs, who had previously completed the course were hired to work as helper instructors. The first hour of lab was dedicated to students working on the University of Washington *Tutorials in Introductory Physics* [61] with the LA serving as the lead lab instructor with the assistance of a graduate teaching assistant (TA). For second half of the laboratory, students performed traditional lab experiments under the instruction of the TA. LAs were required to attend a course in science teaching from an expert from the College of Education and were overseen and further trained by an experienced physics instructor. Once the LA program was discontinued at the end of its funding in 2015, the course was modified to a different structure with three 50-min lectures per week with one 3-h weekly lab. After 2015, all courses were team taught by a pair of experienced instructors. The CSEM grading policy was set by the individual instructors. The aggregated dataset contains 2657 students who completed the course for a grade and completed CSEM pretest and CSEM post-test. This dataset was also analyzed by Henderson *et al.* [32].

As with any analysis, it is preferable to have a stable research environment. Theoretically, IRT should be "sample independent" and return the same results regardless of population; however, this assumes all populations receive relatively consistent coverage of the material so that the ordering of items by the IRT difficulty parameter is consistent for all students. The variability in sample 2 means this assumption is unlikely to hold for this sample and we find that the difficulty parameters are indeed different between samples. It is also likely that this variability influenced the standard deviations of the parameters in sample 2.

C. Unidimensional Item Response Theory

Unidimensional IRT uses a logistic function to model the effect of a single latent trait called "ability" on the probability of a student successfully answering an item [62]. The simplest form of IRT is called the Rasch model where the probability π_{ij} of a student i correctly answering an item j is given as a function of the latent trait θ_i and the item difficulty b_j. The Rasch model is often extended by the addition of a discrimination a_j for each item to form the two parameter logistic (2PL) model:

$$\pi_{ij} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}. \tag{1}$$

The Rasch model is the 2PL model with the discrimination constrained to one, $a_j = 1$. This model can be further extended to the 3PL model, which includes a parameter for random guessing. The 3PL model has also been used to understand the properties of the FCI [28].

D. Multidimensional Item Response Theory

Unidimensional IRT uses a single ability trait; however, conceptual inventories like the CSEM are designed to probe multiple topics such as electric fields, magnetic force, and induction. MIRT extends the IRT model to include multiple latent ability traits. If k latent traits are to be

modeled, then student i's ability becomes the k component vector $\boldsymbol{\theta}_i$. Each item has k discrimination parameters given by the vector $\boldsymbol{a}_j$. MIRT models can be constructed in two forms: compensatory and noncompensatory. The compensatory form of MIRT assumes that the solution does not depend on the latent traits independently and that a deficiency in one trait can be compensated for by a strength in one of the other traits. The compensatory MIRT model is

$$\pi_{ij} = \frac{\exp[\boldsymbol{a}_j \cdot \boldsymbol{\theta}_i + d_j]}{1 + \exp[\boldsymbol{a}_j \cdot \boldsymbol{\theta}_i + d_j]}, \tag{2}$$

where d_j is related to the difficulty of the item and is assumed to be the same for each of the latent traits. In the 2PL model, $d_j = -a_j b_j$. Conversely, the noncompensatory model limits the degree to which one latent ability can compensate for the lack of another. This model does not assume the same difficulty for each item and provides an independent difficulty for each latent trait. Noncompensatory models require a doubling of the parameters estimated and, in our analysis, these models failed to converge.

E. Model fit statistics

IRT uses maximum likelihood (ML) estimation techniques to determine model parameters. The model is used to calculate the likelihood function L representing the probability that a specific observation occurred given the model. ML techniques iteratively search the parameter space for the values of the parameters which maximize L. To determine if ML models fit the data well, several statistics have been developed and should be used in conjunction to evaluate models [63]. This paper will report the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI).

Both AIC and BIC measure the relative information lost when using the model in comparison to the "true" model and correct for overfitting as additional parameters are added to a model [64,65]. Smaller AIC or BIC represent better fitting models. The definition of AIC [Eq. (3)] and BIC [Eq. (4)] follows:

$$\mathrm{AIC} = 2k - 2\ln(L), \tag{3}$$

$$\mathrm{BIC} = k\ln(n) - 2\ln(L), \tag{4}$$

where n is the sample size and k is the number of parameters estimated. When interpreting AIC differences, Burnham and Anderson [64] recommend a difference of greater than 2 as significant and the model with significantly lower AIC should be selected. BIC follows a similar rule with Raftery defining differences of $\Delta\mathrm{BIC} \leq 2$ as "weak," $2 < \Delta\mathrm{BIC} \leq 6$ as "positive," $6 < \Delta\mathrm{BIC} \leq 10$ as "strong," and $\Delta\mathrm{BIC} > 10$ as "very strong" [66]. Both methods penalize the additions of parameters with BIC doing so more strongly and representing a more conservative estimate. Because of the similarity of the two measures, we will adopt Raftery's convention for both AIC and BIC. The likelihood L is the probability that the measured data were observed given the MIRT probability model. For most multiple parameter models with a large sample, this probability is very small. AIC and BIC primarily depending $-2\ln(L)$, which tends to be large because L is very small. As such, changes in AIC and BIC represent exponential changes in the probability of the observed data being represented by the model. If the sample size and number of parameters is constant, then a reduction of AIC or BIC by 10, $\Delta\mathrm{AIC} = -10$ or $\Delta\mathrm{BIC} = -10$, means that the lower AIC or BIC model is $e^5 = 148$ times more likely.

The root mean square error of approximation is a badness-of-fit statistic with values ranging from 0 to 1. Interpretation of the RMSEA relies on an analysis of the 90% confidence intervals (CI) of the statistic. When using the RMSEA, three hypotheses are tested. The first is the exact fit hypothesis H_0: RMSEA $= 0$, which is rejected if the lower bound of the CI includes zero. The second is the "not-close-fit" hypothesis H_0: RMSEA ≥ 0.05, which is rejected if the upper bound of the CI is ≤ 0.05, thus indicating a close fitting model. Finally the "poor-fit" hypothesis H_0: RMSEA ≥ 0.10 is rejected if the upper bound on the CI is less than 0.10 [67]. The statistical software used in this analysis reports the more common 95% confidence interval; we will use this more conservative test in our analysis.

The final two fit statistics reported are closely related: the CFI and TLI. These quantities are incremental-fit-index goodness-of-fit statistics which measure the departure of the tested model from the null model assuming independence, that all parameters are uncorrelated. There exists some debate as to the appropriate cutoff values for good fit using the CFI and TLI with Kline [67] recommending 0.90 as the minimum for acceptable model fit while others [63,68] recommend that a CFI or TLI of 0.97 or greater represents a good fit relative to the independence model and 0.95 or greater is an acceptable fit. We will use the more conservative and more common 0.95 cutoff for good model fit.

F. Additional analyses

Bootstrapped means and standard deviations were calculated for the MIRT parameters $\boldsymbol{a}_j$ and d_j. Bootstrapping generates many subsamples of the data with replacement and runs the desired statistical test on each subsample to generate a normal distribution of fit parameters from which the mean value is calculated. For this work, 100 subsamples were generated requiring one week on a modern personal computer. All analyses were carried out using the "R" programming language [69]. MIRT analysis and fit statistics were generated using the "mirt" package [70]. The bootstraps

were run using the "boot" package [71,72]. Visualizations of the correlation matrix presented in the Supplemental Material [10] were produced with the "qgraph" package [73].

G. Supplemental material

See the Supplemental Material [10] for exploratory MIRT factor analysis of the CSEM and the ability correlation matrices for the optimal models presented in Table IV. An extension of this table with the standard errors for all parameters is also presented.

III. RESULTS

A theoretical model of the CSEM expert solution structure was constructed and tested using MIRT as a confirmatory analysis. This model was then transformed with theoretically justifiable modifications until a best fitting model was found.

Sample 1 contained $N_1 = 2014$ subjects with an average CSEM post-test percentile score of 63.7%. Sample 2 contained $N_1 = 2657$ subjects with an average CSEM post-test percentile score of 44.7%.

A. Exploratory and confirmatory MIRT

MIRT can be used as both an exploratory and confirmatory method. For exploratory analysis, the discrimination matrix $\boldsymbol{a}$ is not constrained and each element may take on any value. For each item j, $\boldsymbol{a}_j$ is a vector of length k forming a matrix with elements a_{jk}. Each column in this matrix $\boldsymbol{a}_k$ represents a "factor." The number of factors in the model is incrementally increased. Successive models are compared using fit statistics to identify the optimal number of factors. An exploratory analysis of the CSEM is presented in the Supplemental Materials [10]. A confirmatory analysis begins with the selection of a theoretical model; the model for this work is described in the next section. The model identifies a small number of concepts covered by each CSEM item. Each concept is associated with a column k in $\boldsymbol{a}$. If item j is not associated with concept k in the theoretical model, the discrimination is constrained to be zero, $a_{jk} = 0$, for the item. The constrained model is then fit to the data and model fit statistics are examined. If fit is acceptable, a small number of related models are then explored to determine the optimal model. The set of models to be investigated is outlined in a model transformation plan before fitting the initial model.

B. Theoretical framework

A theoretical model of the knowledge structure measured by the CSEM was developed using multiple expert solutions of the instrument applying the same methods as Stewart *et al.* [7]. Instructors in the classes studied and members of the research team were asked to provide detailed solutions to the CSEM. These solutions were decomposed to the sentence level. Sentences expressing the same physical reasoning were grouped together and a general statement of that reasoning, called a *principle*, was constructed. Table I presents the list of principles identified. Each principle was classified as a law (L) representing important physical laws such as Newton's 2nd law, as a definition (DF) introducing an important new quantity, or as a fact (F) representing knowledge about the universe that did not rise to the level of a law. From these fundamental pieces of information, corollaries (C) were derived as important secondary results. The expert solutions contained a number of qualitative statements that interpreted the laws, definitions, and corollaries; these were called lemmas (LM). Table I shows the classification of the principles into laws, definitions, facts, corollaries, and lemmas as well as the higher order principle from which a lower order principle was derived. As was found for the FCI, expert solutions did not contain all the higher order principles from which the lemmas and corollaries were derived; these higher order principles were inferred and added to the model. Table I also presents the CSEM items requiring each principle for their solution.

The theoretical model in Table I differs from the models of Planinic [13] and Maloney *et al.* [1] because it is grounded in the reasoning found in expert solutions. A course-grained model related to that of Planinic [13] is presented, but has substantially worse model fit than the model described above.

This process was substantially less straightforward for the CSEM than for the FCI. There were two sets of principles that could not be distinguished by the items in the instrument because they all loaded on the same items. A principle will be said to "load" on an item if it is required for the solution of the item following the terminology of factor analysis. The principles "charge is conserved" and "charge does not rapidly escape to the environment" both were used only in items 1 and 2 and are labeled L3. Faraday's law and the definition of magnetic flux both loaded on items 29, 30, and 32 and are labeled L8. Corollaries C3 and C4, involving the behavior of conductors, were coded as single principles, but could be derived from a number of other principles not independently tested within the instrument. As such, they were left without a derivation because their structure could not be further resolved by the instrument. There were insufficient items in the instrument to separate the addition of electric and magnetic fields (L9), so these were combined. The instrument is ambiguous about the items involving the magnetic fields of currents (23, 24, 26, 28, and 30). While some items are represented as three-dimensional wires (24, 30), some are simply shown as current into or out of the page. The expert solutions all addressed these problems using the field of an infinite straight wire and the form of the right-hand rule for this system (grab wire with right hand, fingers curl in the direction of the field). Both the field of the wire and the right-hand rule are derived from the more general Biot-Savart law (L6) and the right-hand

TABLE I. Theoretical model of electricity and magnetism as tested by the CSEM. The optimal model column indicates the samples for which the principle was included in the optimal model; M1-6 for sample 1 and M2-6 for sample 2. For items with two possible solution paths, the item number is followed by parentheses which enclose the solution path number.

Label	Optimal model	Derived from	CSEM No.	Principle
				Mechanics
DF1	1,2		6, 8, 9 (1)	The net force is the vector sum of the forces (forces add as vectors).
L1	1,2		10, 31	Newton's 2nd law.
L2	1,2		7(1), 24	Newton's 3rd law.
C1				Objects moving in a curved trajectory will experience centripetal acceleration.
C2	1,2	C1, L1	22	If a particle is turning in some direction, there is a force in that direction.
				Electrostatics
L3	1,2		1, 2	Charge is conserved. Charge does not rapidly escape to the environment.
F1	1,2		2	Charge cannot move through an insulator.
C3	1,2		13, 14	A conductor shields its interior from the electric field and force.
C4	1,2		1	Charge spreads out over the outer surface of a conductor.
L4	1,2		3, 7 (2)	Coulomb's law for the electric force ($\vec{F} = \frac{kq_1q_2}{r^2}\hat{r}$).
L5	1,2			Coulomb's law for the electric field ($\vec{E} = \frac{kq}{r^2}\hat{r}$).
LM1		L4	6, 7, 8, 9 (1), 14	Opposite charges attract and likes repel.
DF2	1,2		9(1), 10, 12, 15, 19, 20	Definition of electric field ($\vec{F} = q\vec{E}$).
DF3	1,2		15	The electric field is tangent to electric field lines.
LM2	1,2	L5	9(2)	Electric fields point away from positive charge.
				Electric potential
DF4	1			Definition of electric potential ($\Delta V = \frac{W_{\text{ext}}}{q} = -\int E dx$).
LM3	2	DF4	16	The electric potential contains an arbitrary constant.
LM4	2	DF4	17	Relation of work and electric potential ($W_{\text{ext}} = q\Delta V$).
C6	1,2	DF4	18, 20	Relation of electric potential and field ($\vec{E} = -\frac{dV}{dx}\hat{x}$).
LM5	2	DF4	11, 19, 20	Electric field points to lower potential.
				Magnetostatics
L6	1,2		23, 24 (2), 26, 28, 30	Biot-Savart law ($d\vec{B} = \frac{\mu_0}{4\pi}\frac{Id\vec{\ell}\times\hat{r}}{r^2}$).
L7	1,2		22, 24 (2), 25, 31	Lorentz force ($\vec{F} = q\vec{v}\times\vec{B}$ or $d\vec{F} = Id\vec{\ell}\times\vec{B}$).
LM6		L7	21, 27	The magnetic force on a stationary charge is zero.
LM7	1,2	L6, L7, DF5	24(1)	Like currents attract and opposites repel.
F2	1,2		29	The magnetic field of a permanent magnet weakens with distance.
DF5	1,2		22, 23, 24 (2), 26, 28, 30, 31	Right-hand rule for cross products.
DF6	1,2		25	Magnitude of the cross product ($\lvert\vec{A}\times\vec{B}\rvert = \lvert\vec{A}\rvert\lvert\vec{B}\rvert\sin\theta$).
				Induction
L8	1,2		29, 30, 32	Faraday's law ($emf = -\frac{d\Phi}{dt}$). Definition of magnetic flux ($\Phi = \int_S \vec{B}\cdot\hat{n}dA$).
				Superposition
L9	1,2		9(2), 23, 28	Electric and magnetic fields add as vectors.

rule for the cross product (DF5). There were no items that made the distinction between the infinite wire field and the field of an infinitesimal current element, so the field of the wire was combined with the Biot-Savart law and the right-hand rule for a wire with the right-hand rule for the cross product.

Two equally likely solution paths were identified for three of the items: 7, 9, and 24. Both solution paths were added to Table I and will be explored with MIRT. For these items, the solution path number (1 or 2) was placed in parenthesis next to the item number. For example, 7(2) in row L4 means principle L4 is used in the second solution path for item 7. If a principle was used in both solution paths, the parentheses were dropped.

While many physics questions have multiple solution paths and one goal of physics instruction is for students to

see physics as a set of linked concepts, items with multiple solution paths in an assessment instrument make it difficult to determine what the instrument actually measures. To resolve what an item with multiple common solution paths actually measures, multiple related items as required probing the same concepts in different ways. MIRT allowed the exploration of the multiple solution paths and the determination of the path measured by the instrument if the principles in each path are sufficiently probed elsewhere in the instrument.

Previous work on the FCI [7] showed that placing problems in a group sharing a common stem could generate correlations between the items which were not grounded in the student's understanding of the items. Their work removed all but first item of each problem group. The CSEM contains three problem groups: items 3, 4, and 5; items 10 and 11; and items 17, 18, and 19. Each problem group was examined to determine if spurious correlations were likely. Items 4 and 5 both depend on the answer to item 3 and cannot be treated independently. Items 4 and 5 were removed from the analysis. Item 11 depends on the answer to item 10 and was also removed. Items 17, 18, and 19 can be answered relatively independently and were retained.

Several additional items were removed from the first stage of the analysis and only analyzed after an initial optimal model was constructed: items 9, 14, 31, and 32. Item 9 was the only item directly testing Coulomb's law for the electric field (L5). The item could also be solved using Coulomb's law for the electric force (L4) and the relation of force and field ($\vec{F} = q\vec{E}$) (DF2). Many items probe these two principles. As such, first models were constructed to determine the correct structure of the electric force principles. Once this model was determined, the two solution paths for item 9 were then investigated. Individual experts produced multiple solution paths for items 14, 31, and 32. Some of these solutions required multiple principles not measured by other items in the CSEM. As such, items 14, 31, and 32 were not included in the initial analysis. We will call these items "reserved" items. They were analyzed after the optimal model was constructed by adding a separate "unknown" principle which captured any additional reasoning needed to solve the item.

The principles in Table I will be mapped using MIRT onto a set of latent traits θ_{ki} representing the ability of each student i to apply principle k.

C. Model transformation plan

Confirmatory analyses first fit the most complete theoretical model for a system of data and then carry out a small number of theoretically motivated transformations of the model to potentially improve model fit. Following this methodology, the most complete theoretical model (Table I) of the CSEM was fit first. The expert solutions to the CSEM identified two solution paths to items 7 and 24; these alternate solutions were then explored and compared to the most complete model. The first solution path, indicated by the number in parentheses in Table I, was fit as part of the initial model. The second solution path for items 7 and 24 was then fit and the best model selected for each. To test an alternate solution path, the MIRT parameter matrix is changed, constraining the parameters of the first solution path to be zero and allowing the parameters related to the alternate solution path to be nonzero.

One of the fundamental questions about the structure of student knowledge is how granular or fine grained the knowledge is. This can be tested by determining if the lemmas (LM) in Table I improve the model or if the model improves if the lemmas are eliminated. When a principle, such as a lemma, is removed from the model, the latent trait θ_k representing that principle is no longer used in the model. Removing a principle does not change the number of items in the CSEM being modeled. For the next sequence of transformations, lemmas were removed from the model by replacing them with the higher level principle from which they were derived. This was performed in three stages. First, L5 (Coulomb's force law) was combined with LM1 (opposites attract and likes repel). All items loading on either LM1 or L5 were set to load on L5. Second, lemmas involving electric potential (LM3, LM4, and LM5) were collapsed to the principle from which they were derived, DF4 (the definition of electric potential). Third, LM6 (the magnetic force on a stationary charge is zero) was combined with L7 (the Lorentz force law). Each of these was tested in turn; the order was arbitrary and could be rearranged with no effect on the results.

Finally, a model using only the general categories (mechanics, electrostatics, electric potential, magnetostatics, induction, and superposition) from Table I was tested. This represented a collection of principles within general topics and was the model most closely related to previous work on evaluating the structure of the CSEM [1,13].

D. Constrained MIRT

The complete model presented in Table I eliminating blocked items 4, 5, and 11 and "reserved" items 9, 14, 31, and 32, which will be explored later, was fit to each sample. For items 7 and 24, where multiple likely solution paths were identified, the first solution path was used in this initial model. The model was fit by constraining the MIRT discrimination matrix $\boldsymbol{a}_j$ so that discrimination parameters that did not conform to the model were zero. For example, the discrimination parameter associated with conservation of charge (L3) was constrained to be zero, $a_{L3} = 0$, except for items 1 and 2 (see Table I). Following Stewart *et al.*, one discrimination parameter a_0 was allowed to load on all items. This parameter is associated with a general ability θ_0 to solve conceptual electricity and magnetism questions. The $a_{j\neq 0}$ parameters then capture the additional discrimination of the item for an individual principle j. The initial

TABLE II. Sample 1 model transformation. Differences in AIC and BIC determine whether the models are statistically different; CFI, TLI, and RMSEA indicate the quality of fit for each model.

Transformed model	Transformation	Original model	AIC	BIC	CFI	TLI	RMSEA	Superior
M1-1			54,941	55,485	0.964	0.953	0.025(0.022,0.028)	
M1-2	Solution path 2 to item 7.	M1-1	54,941	55,485	0.964	0.953	0.025(0.022,0.028)	M1-1
M1-3	Solution path 2 to item 24.	M1-1	54,928	55,484	0.967	0.956	0.024(0.021,0.027)	M1-3
M1-4	Combine LM1 with L4.	M1-1	54,914	55,458	0.969	0.959	0.023(0.020,0.026)	M1-4
M1-5	Combine LM3, LM4, LM5 with DF4.	M1-4	54,893	55,437	0.970	0.960	0.023(0.020,0.026)	M1-5
M1-6	Combine LM6 with L7.	M1-5	54,860	55,404	0.975	0.967	0.021(0.018,0.024)	M1-6
M1-7	Collapse to general categories.	M1-6	54,969	55,434	0.948	0.936	0.029(0.027,0.032)	M1-6

theoretical model was fit to both samples producing models M1-1 and M2-1 where the first number is the sample number and the second number is the model number. The results of fitting this model for sample 1 are shown in Table II and sample 2 in Table III. The models are referenced to the transformed model column in the tables. For both samples, the models had good fit indices: CFI > 0.96, TLI > 0.95, and RMSEA < 0.3.

A sequence of more parsimonious models was then fit where transformations proceeded according to the model transformation plan in Sec. III C. The first transformed models, M1-2 and M2-2, investigated an alternate solution to item 7 as indicated by the 7(2) notation in Table I, where the 2 represents the solution path number. The original model was fit with 7(1) constraints. Item 7 asks for the magnitude and direction of the forces on unequal point charges. Solution path 1 used opposites attract and likes repel (LM1) and Newton's 3rd law (L2). The second solution path also used opposites attract and likes repel but applied Coulomb's force law (L4) to obtain the magnitude. In sample 1, there was no difference in the fit of the two solution paths. In sample 2, the model fit was significantly worse for the second solution with an increase in both AIC and BIC of 63, a very strong change using Raftery's classification [66]. As such, the model with the first solution path was retained in both cases. Students solve item 7 using the opposites attract and likes repel (LM1) and Newton's 3rd law (L2) rather than applying Coulomb's law to obtain the magnitude. Tables II and III show the fit parameters for the transformed model, the model from which it was transformed and is being compared (original model), and which of the models was retained (superior model).

Models M1-3 and M2-3 investigated an alternate solution path to item 24. The first solution path, used in the initial model, solved the item by applying like currents attract and opposites repel (LM7) and Newton's 3rd law (L2). The second solution path began with first principles from the Biot-Savart law (L6) and applied the Lorentz force law (L7) using the right-hand rule for the cross product (DF5) to find the direction. Newton's 3rd law (L2) was again applied to find the second force. Solution path 2 showed a significant improvement in AIC of 13 for sample 1, a very strong change, but no significant change in BIC. For sample 2, the second solution path was significantly worse with AIC increasing by 49 and BIC by 61, both very strong changes. With only the change in AIC in sample 1 supporting solution path 2 and much stronger changes in sample 2 supporting path 1, path 1 was retained for all future models. As such, students solve item 24 by applying like currents attract and opposites repel (LM7) rather than the more fundamental Biot-Savart law.

Models M1-4 and M2-4 through M1-6 and M2-6 test whether condensing some of the lemmas into broader principles, laws, and definitions improves model fit. Models M1-4 and M2-4 replace opposites attract and likes repel (LM1) with Coulomb's force law (L4) from which it is derived. This significantly improved model fit over M1-1

TABLE III. Sample 2 model transformation. Differences in AIC and BIC determine whether the models are statistically different; CFI, TLI, and RMSEA indicate the quality of fit for each model.

Transformed model	Transformation	Original model	AIC	BIC	CFI	TLI	RMSEA	Superior
M2-1			77,330	77,901	0.983	0.978	0.021(0.019,0.024)	
M2-2	Solution path 2 to item 7.	M2-1	77,393	77,964	0.980	0.970	0.023(0.021,0.026)	M2-1
M2-3	Solution path 2 to item 24.	M2-1	77,379	77,962	0.978	0.970	0.025(0.022,0.027)	M2-1
M2-4	Combine LM1 with L4.	M2-1	77,282	77,853	0.983	0.978	0.021(0.019,0.024)	M2-4
M2-5	Combine LM3, LM4, LM5 with DF4.	M2-4	77,308	77,879	0.984	0.980	0.021(0.018,0.023)	M2-4
M2-6	Combine LM6 with L7.	M2-4	77,265	77,835	0.984	0.980	0.020(0.018,0.023)	M2-6
M2-7	Collapse to general categories.	M2-6	77,315	77,803	0.975	0.969	0.025(0.023,0.028)	M2-6

in sample 1 with AIC and BIC decreasing by 27, both very strong changes. In sample 2, model fit was also improved when compared to the model M2-1 with AIC decreasing by 48 and BIC by 48, both very strong changes. As such, transformed models M1-4 and M2-4 were retained as the superior models. This change also served to collapse the two solution paths for item 7 into one path. As such, students understanding of the electric force was less granular than initially represented in the theoretical model.

The next models, M1-5 and M2-5, combined several principles of electric potential (LM3, LM4, LM5) into the definition of electric potential (DF4) from which they were derived while retaining the changes made in models M1-4 and M2-4. This model was a significant improvement over model M1-4 in sample 1 with AIC and BIC decreasing by 21, both very strong changes. However, in sample 2, model M2-5 was significantly inferior to model M2-4 with AIC and BIC increasing by 26, both very strong changes. As such, model M1-5 was retaining for sample 1 as the superior model, but not for sample 2. This marked the first substantial deviation between the two datasets. For sample 1, students had a more integrated understanding of potential allowing the combination of LM3 (potential contains an arbitrary constant), LM4 (the relation of work and potential), and LM5 (electric field points to lower potential) into a single definition of potential (DF4). Students in sample 2 had differing reasoning abilities on these lemmas.

Models M1-6 and M2-6 combined the principle that the magnetic force on a stationary charge is zero (LM6) with the principle from which it is derived, the Lorentz force law for magnetic fields (L7). This change significantly improved model fit for model M1-5 in sample 1 with AIC decreasing by 33 and BIC by 33, both very strong changes. In sample 2, model M2-6, which made the same modifications to model M2-4, significantly improved model fit with AIC decreasing by 17 and BIC by 18, both very strong changes. As such, models M1-6 and M2-6 were retained as the superior models. The reasoning of students in both samples about stationary magnetic force was not differentiated from reasoning about nonzero magnetic force.

Models M1-6 and M2-6 represent the most parsimonious models which the authors felt could be theoretically justified. Many studies have sought to produce even more general models of the FCI and the FMCE through exploratory methods such as factor analysis as described in the introduction. These methods model the internal structure of an instrument through a small number of factors thought to represent subsets of the instrument that are conceptually similar. To test whether this was the correct way to model the CSEM, models M1-7 and M2-7 condensed models M1-6 and M2-6 to the bolded general categories in Table I (mechanics, electrostatics, electric potential, magnetostatics, induction, and superposition). Model M1-7 had significantly poorer fit than model M1-6 with an increase in AIC of 109 and BIC of 30, both very strong changes. Therefore, model 1-6 represents the optimal model of student knowledge for sample 1. Model M2-7 made a similar transformation to model M2-6; the model fit indices to this transformation were mixed. AIC increased by 50, but BIC decreased by 32, both very strong changes; however, CFI, TLI, and RMSEA all support the choice of model M2-6 as the optimal model for sample 2. For both samples, the theoretical model grounded in specific principles of physics was superior to a model using broad general topics.

The sequence of models used progressively fewer parameters; model fit usually increases with the addition of parameters. AIC and BIC both penalize the addition of parameters to correct for overfitting. BIC penalizes additional parameters more strongly than AIC.

E. Analysis of optimal models

The full optimal models for sample 1, model M1-6, and sample 2, M2-6, are shown in Table IV. Each item is reported with the individual principles required for its solution. The number in parenthesis is the discrimination, a_{jk}, of item j on principle k. The means of the discriminations were calculated by bootstrapping with 100 subsamples. The standard error of the mean was also calculated by bootstrapping and is presented in the Supplemental Material [10]. The principle discrimination a_{jk} represents how well the item j discriminates between high and low ability students above the discrimination a_{j0} of the item on a general facility with conceptual electricity and magnetism. Table IV also reports the results of a t test for each discrimination as a superscript to determine if the discrimination parameter is significantly different from zero. A Bonferroni correction has been applied to adjust for the number of statistical tests performed. The table also reports d_j, the overall difficulty of the item.

The optimal models for sample 1 and sample 2 differ slightly because of the way electric potential was modeled. For sample 1, only DF4 was included (model M1-5), but in sample 2 DF4 was expanded into lemmas LM3, LM4, and LM5 (model M2-4). These differences were retained as optimal models M1-6 and M2-6 were constructed. To determine how similar the models are, a single model must be selected. Because model M1-6 is the more parsimonious, it was selected for comparisons between the two samples. This model was fit to sample 2 and bootstrapping was repeated. A comparison of the fits of this model for the two samples is shown in Table V where the mean fit values for sample 2 have been subtracted from those obtained from the fit of sample 1 to form Δa_{jk} and Δd_j. The significance of the differences between the parameters was calculated with t tests with a Bonferroni correction. Significance values are reported as superscripts. The difference in overall discrimination, Δa_{j0}, and difficulty, Δd_j, is statistically significant ($ps < 0.001$) for the majority of the items. Many of the principle discriminations a_{jk} were not

TABLE IV. Optimal MIRT model for samples 1 and 2. The first column shows the CSEM item number (No.). Not all CSEM items were modeled. The number in parenthesis is the discrimination a_{jk} for principle k of item j. a_{j0} is the discrimination for a factor loaded on all items and d_j is related to the overall difficulty of the item. Both parameters are also rank ordered from smallest to largest. The significance of a t test with Bonferroni correction to determine if the difficulty and discrimination are different from zero is reported as a superscript. A superscript of "a" represents the corrected equivalent of $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$.

	Sample 1					Sample 2				
		a_{j0}	d_j	a_{j0}	d_j		a_{j0}	d_j	a_{j0}	d_j
No.	Principles			Rank	Rank	Principles			Rank	Rank
1	C4(0.16)^c L3(0.48)^c	0.68^c	2.67^c	8	12	C4(0.22)^c L3(0.47)^c	0.82^c	1.38^c	11	23
2	F1(0.23)^c L3(0.46)^c	0.58^c	0.98^c	5	10	F1(0.27)^c L3(0.49)^c	0.59^c	−0.66^c	6	9
3	L4(0.67)^c	0.81^c	2.04^c	12	19	L4(0.32)^c	0.66^c	1.58^c	7	24
6	L4(0.41)^c DF1(0.08)^c	1.23^c	2.05^c	22	20	L4(0.33)^c DF1(0.18)^c	1.39^c	0.99^c	21	21
7	L4(−0.15)^c L2(0.15)^c	1.25^c	1.23^c	23	15	L4(−0.01) L2(0.59)^c	2.03^c	−0.05^c	25	15
8	L4(0.26)^c DF1(0.13)^c	0.86^c	1.15^c	15	12	L4(0.22)^c DF1(0.20)^c	0.97^c	0.62^c	15	19
10	L1(0.31)^c DF2(0.23)^c	1.09^c	0.23^c	19	5	L1(0.25)^c DF2(0.10)^c	1.30^c	−0.50^c	19	11
12	DF2(0.28)^c	0.78^c	2.66^c	11	21	DF2(0.15)^c	0.94^c	1.68^c	13	25
13	C3(0.21)^c	0.94^c	1.57^c	18	17	C3(0.25)^c	0.81^c	−0.97^c	10	7
15	DF2(0.07)^c DF3(0.26)^c	0.93^c	0.72^c	16	9	DF2(0.06)^c DF3(0.22)^c	1.26^c	−0.94^c	18	8
16	DF4(0.31)^c	0.82^c	0.54^c	13	7	LM3(0.27)^c	0.95^c	−1.16^c	14	5
17	DF4(0.19)^c	0.69^c	1.36^c	9	16	LM4(0.27)^c	1.15^c	−0.31^c	16	13
18	C6(1.06)^c	0.40^c	1.18^c	3	14	C6(0.82)^c	0.30^c	0.32^c	2	17
19	DF2(0.13)^c DF4(0.91)^c	1.13^c	2.80^c	21	24	DF2(0.12)^c LM5(0.50)^c	1.40^c	−0.05^c	22	14
20	DF4(0.45)^c DF2(0.02) C6(1.00)^c	0.49^c	−0.62^c	4	1	DF2(0.17)^c C6(0.75)^c LM5(0.39)^c	0.55^c	−2.10^c	4	1
21	L7(1.03)^c	0.75^c	1.68^c	10	18	L7(0.61)^c	0.40^c	−1.40^c	3	4
22	C2(0.20)^c L7(0.19)^c DF5(0.23)^c	0.38^c	−0.49^c	1	2	C2(0.23)^c L7(0.04)^c DF5(0.08)^c	0.02^c	−0.53^c	1	10
23	L9(0.01) DF5(0.26)^c L6(0.44)^c	1.12^c	2.79^c	20	23	L9(0.21)^c DF5(0.25)^c L6(0.45)^c	1.42^c	0.74^c	23	20
24	L2(0.13)^c LM7(0.27)^c	0.85^c	−0.15^c	14	3	L2(0.58)^c LM7(0.31)^c	1.35^c	−0.97^c	20	6
25	L7(0.30)^c DF6(0.33)^c	0.93^c	0.49^c	17	6	L7(0.37)^c DF6(0.32)^c	1.16^c	−0.48^c	17	12
26	DF5(0.39)^c L6(0.52)^c	1.64^c	3.71^c	25	25	DF5(0.25)^c L6(0.45)^c	1.69^c	1.22^c	24	22
27	L7(0.68)^c	0.67^c	1.18^c	7	13	L7(0.54)^c	0.72^c	−1.55^c	9	3
28	L9(0.24)^c DF5(0.18)^c L6(0.09)^c	0.40^c	0.64^c	2	8	L9(0.21)^c DF5(0.07)^c L6(0.08)^c	0.56^c	0.57^c	5	18
29	L8(0.33)^c F2(0.23)^c	1.27^c	−0.10^c	24	4	L8(0.23)^c F2(0.17)^c	0.84^c	−1.83^c	12	2
30	DF5(0.08)^c L6(0.08)^c L8(0.24)^c	0.66^c	1.02^c	6	11	DF5(0.06)^c L6(0.03)^c L8(0.31)^c	0.68^c	0.16^c	8	16

significantly different between the samples, suggesting that many of CSEM items perform similarly at different institutions once overall differences in ability are removed.

For sample 1, items 3, 12, 13, 16, 17, 18, 21, and 27 load on a single principle and probe six individual concepts: L4, DF2, C3, DF4, C6, and L7. These questions could be used to investigate student knowledge about these concepts independent of other principles. The remaining 19 items load on multiple principles; however, many have a single principle that has a discrimination at least twice as large in absolute value as the next largest (items 1, 2, 6, 8, 15, 19, 20, 24, and 30). These items could be used to characterize student knowledge on the high discrimination principle.

For sample 2, items 3, 12, 13, 16, 17, 18, 21, and 27 also load on a single factor. The remaining 18 items load on multiple principles; however, many have a single factor that has a discrimination at least twice as large in absolute value as the next largest (items 1, 7, 10, 15, 19, 22, 28, and 30). These items could be used to characterize student knowledge on the high discrimination principles.

F. Analysis of reserved items

Item 9 was initially withheld from the analysis because the expert solutions provided two equally plausible solution paths, one relying primarily on reasoning using electric force, the other relying on reasoning using the electric field. Only item 9 directly probed the electric field of a point charge (L5). As such, an optimal model for other principles was identified before exploring item 9. Two models were tested using the two possible solutions to item 9 as shown in Table I. The first solution path used a positive test charge, opposites attacts and likes repel (LM1, now L4), the vector addition of forces (DF1), and the relation of force and field (DF2). The second solution used that electric field points away from positive charge (LM2) and the vector addition of fields (L9). Solution path 1 involving electric force produced the optimal model for sample 1 with very strong changes in AIC and BIC. Solution path 2 involving electric field produced the superior model for sample 2 with a positive change in AIC and a strong change in BIC. Model fit statistics and the solution path selected for analysis in

TABLE V. The difference in parameters between samples 1 and 2 using the optimal model for sample 1 (model M1-6). The number in parenthesis is the difference in discrimination Δa_{jk} for item j. Δa_{j0} is the difference in discrimination for a principle loaded on all items and Δd_j is the difference in the difficulty of the item. The standard error of each parameter is also reported. Each difference was t tested with a Bonferroni correction. A superscript of "a" represents the corrected equivalent of $p < 0.05$, "b" $p < 0.01$, and "c" $p < 0.001$.

CSEM No.	Principles	Δa_{j0}	Δd_j
1	C4(-0.06 ± 0.01)[a] L3(0.02 ± 0.02)	-0.13 ± 0.01[c]	1.28 ± 0.02[c]
2	F1(-0.10 ± 0.02)[b] L3(-0.02 ± 0.02)	-0.02 ± 0.01	1.65 ± 0.01[c]
3	L4(0.38 ± 0.02)[c]	0.18 ± 0.01[c]	0.50 ± 0.02[c]
6	L4(0.07 ± 0.01)[a] DF1(-0.11 ± 0.01)[c]	-0.14 ± 0.01[c]	1.07 ± 0.02[c]
7	L4(-0.16 ± 0.01)[c] L2(-0.45 ± 0.01)[c]	-0.77 ± 0.02[c]	1.29 ± 0.01[c]
8	L4(0.05 ± 0.01)[a] DF1(-0.06 ± 0.01)[a]	-0.10 ± 0.01[c]	0.56 ± 0.01[c]
10	L1(0.04 ± 0.02) DF2(0.13 ± 0.01)[c]	-0.21 ± 0.01[c]	0.73 ± 0.01[c]
12	DF2(0.13 ± 0.01)[c]	-0.15 ± 0.01[c]	0.99 ± 0.02[c]
13	C3(0.01 ± 0.01)	0.15 ± 0.01[c]	2.53 ± 0.01[c]
15	DF2(0.02 ± 0.01) DF3(0.05 ± 0.02)	-0.34 ± 0.01[c]	1.67 ± 0.01[c]
16	DF4(0.18 ± 0.01)[c]	-0.01 ± 0.01	1.59 ± 0.01[c]
17	DF4(0.05 ± 0.01)[c]	-0.32 ± 0.01[c]	1.65 ± 0.01[c]
18	C6(0.22 ± 0.02)[c]	0.10 ± 0.01[c]	0.87 ± 0.01[c]
19	DF4(0.60 ± 0.02)[c] DF2(-0.05 ± 0.01)	-0.12 ± 0.02[c]	2.87 ± 0.03[c]
20	DF4(0.12 ± 0.01)[c] DF2(-0.22 ± 0.01)[c] C6(0.25 ± 0.02)[c]	-0.02 ± 0.01	1.50 ± 0.02[c]
21	L7(0.39 ± 0.02)[c]	0.37 ± 0.01[c]	3.10 ± 0.02[c]
22	C2(-0.01 ± 0.01) L7(0.16 ± 0.01)[c] DF5(0.15 ± 0.01)[c]	0.36 ± 0.01[c]	0.03 ± 0.01[b]
23	L9(-0.20 ± 0.01)[c] DF5(0.00 ± 0.02) L6(0.03 ± 0.02)	-0.28 ± 0.01[c]	2.07 ± 0.02[c]
24	L2(-0.46 ± 0.01)[c] LM7(-0.03 ± 0.02)	-0.48 ± 0.01[c]	0.83 ± 0.01[c]
25	L7(-0.07 ± 0.01)[c] DF6(0.06 ± 0.02)	-0.23 ± 0.01[c]	0.97 ± 0.01[c]
26	DF5(0.11 ± 0.02)[b] L6(0.11 ± 0.02)	-0.05 ± 0.02	2.52 ± 0.04[c]
27	L7(0.13 ± 0.01)[c]	-0.07 ± 0.01[c]	2.73 ± 0.01[c]
28	L9(0.06 ± 0.01)[a] DF5(0.11 ± 0.01)[c] L6(0.02 ± 0.01)	-0.15 ± 0.01[c]	0.09 ± 0.01[c]
29	L8(0.08 ± 0.01)[c] F2(0.04 ± 0.01)	0.41 ± 0.01[c]	1.78 ± 0.01[c]
30	DF5(0.00 ± 0.01) L6(0.07 ± 0.01)[c] L8(-0.07 ± 0.01)[b]	-0.03 ± 0.01	0.87 ± 0.01[c]

TABLE VI. Sample 1 and 2 reserved item comparisons. Differences in AIC and BIC determine whether the models are statistically different; CFI, TLI, and RMSEA indicate the quality of fit for each model.

CSEM No.	Solution path	AIC	BIC	CFI	TLI	RMSEA	Superior path
				Sample 1			
9	Path 1	56,675	57,247	0.973	0.965	0.022(0.020,0.025)	Path 1
	Path 2	56,691	57,258	0.973	0.964	0.022(0.019,0.025)	
14	Path 1	57,184	57,750	0.970	0.961	0.022(0.019,0.025)	Path 1
	Path 2	57,188	57,760	0.971	0.962	0.022(0.019,0.025)	
31	Path 1	56,925	57,492	0.967	0.957	0.024(0.021,0.027)	Path 2
	Path 2	56,910	57,482	0.967	0.957	0.024(0.021,0.027)	
32	Path 1	57,661	58,221	0.970	0.961	0.022(0.019,0.025)	Path 1
	Path 2	57,680	59,246	0.970	0.961	0.022(0.020,0.025)	
				Sample 2			
9	Path 1	80,398	80,998	0.986	0.982	0.019(0.017,0.022)	Path 2
	Path 2	80,393	80,988	0.986	0.982	0.019(0.017,0.022)	
14	Path 1	79,457	80,051	0.987	0.983	0.018(0.015,0.020)	Path 1
	Path 2	79,497	80,097	0.986	0.981	0.019(0.016,0.021)	
31	Path 1	79,073	79,667	0.987	0.983	0.018(0.016,0.021)	Path 2
	Path 2	79,044	79,644	0.985	0.981	0.019(0.017,0.022)	
32	Path 1	80,703	81,291	0.985	0.981	0.019(0.017,0.022)	Path 1
	Path 2	80,725	81,319	0.984	0.979	0.020(0.017,0.022)	

this section are shown in Table VI. Note, AIC and BIC for this section cannot be compared directly to values in Tables II and III because the number of items fit have changed.

Items 14, 31, and 32 were investigated by adding a separate unknown principle to their model in Table I. Expert solutions were quite varied for these items and often contained additional principles not tested elsewhere in the instrument. Models with this additional principle were tested independently and compared for each item. The unknown principle was used to capture any reasoning not already captured by the principles included in Table I. First, the loadings for items 14, 31, and 32 presented in Table I were added to the already identified optimal models M1-6 and M2-6 and model fit recalculated. These models are identified as solution path 1. The fit of this model was then compared to a model that added the unknown principle to one of the reserved items, solution path 2. In both sample 1 and sample 2, the addition of the unknown principle only improved the model fit for item 31, implying the students were using additional reasoning beyond L7 (Lorentz force) and DF5 (right-hand rule for cross products) to solve the item.

MIRT allows the estimation of the ability for each principle for each student. Stewart *et al.* [7] used the correlations of these abilities as a method to explore the degree to which student knowledge is integrated. The correlation matrices for the student abilities $\boldsymbol{\theta}_i$ are presented in the Supplemental Material [10].

IV. DISCUSSION

This study investigated two research questions; they will be discussed in the order proposed.

RQ1: What is the optimal model of student knowledge measured by the CSEM? Are the principles forming the optimal model consistent across samples? The optimal model for sample 1 required 23 principles, while the optimal model for sample 2 required 25. The optimal models had 22 principles in common. As such, while there were some differences between the two optimal models, in general they were very similar despite large differences in instructional environment and the student's overall performance on the CSEM.

The optimal model for sample 1 was comprised of a model with most of the lemmas collapsed into the higher level principles from which they were derived. Two lemmas, LM2 and LM7, were retained in optimal model M1-6. The collapse of LM2 into L5 could not be investigated because the CSEM did not contain other items which employed L5. For this sample, student knowledge of electricity and magnetism is better represented by the general laws, definitions, facts, and corollaries defining the topic without the additional set of qualitative principles. This observation is consistent with a similar result found for the FCI [7].

The optimal model for sample 2 included three additional lemmas (LM3, LM4, and LM5) rather than the general definition of electric potential (DF4). In all, 5 of the original 7 lemmas were retained in optimal model M2-6. Students in sample 2 have a less integrated understanding of electric potential than students in sample 1, perhaps explained by their overall weaker performance on the CSEM. For these students, a model with detailed coverage of the implications of the general laws better fit the student understanding of electricity and magnetism. The understanding of these students is less well integrated than that of students in sample 1.

Collapsing the optimal models further to very general categories such as electrostatics or magnetostatics (models M1-7 and M2-7) reduced model fit and, as such, student knowledge of electricity and magnetism is more granular than these broad topics.

The models of the two samples also differed for reserved item 9; this may have resulted from the instruction provided to students in the two samples. The lead instructor for sample 1 reported presenting the material from the standpoint of inserting a positive test charge; the solution path using electric force produced the optimal model for item 9 in sample 1. Conversely, many instructors taught the classes in sample 2 and presented the addition of electric field in many different ways. In this case, the solution using the principle that fields point away from positive charges produced the superior model. This suggests that MIRT could be used to probe differences in the effect of specific instructional choices on student understanding.

Exploration of the rest of the reserved items (14, 31, and 32) through the addition of an unknown principle showed that these items could reasonably be explained using the theoretical model already developed for this instrument. With the exception of item 31, none of the models including the unknown factor performed better than the ones without it. Again, the optimal models for the two samples were similar but not identical.

The differences between the optimal models for the two samples shows the optimal model for the CSEM does vary somewhat between institutions. The difference, however, was restricted to the decision to retain lemmas LM3, LM4, and LM5. Further, unlike the FCI, both optimal models of the CSEM did include one lemma, LM7, and potentially a second lemma, LM2; combining LM2 with L5 could not be tested because of the structure of the instrument. While model M2-5 and models M1-3 and M2-3 were significantly less well fitting, they still possessed excellent fit characteristics with CFI > 0.96, TLI > 0.95, and RMSEA < 0.25. As such, variations between institutions were present, but these variations produced models with similarly excellent fit. As such, it may be reasonable to use the model of the CSEM eliminating all lemmas when comparing results between institutions.

RQ2: Are the parameters of the optimal models consistent between samples? The uniformly larger difficulty

values, d_j, in sample 1 indicate that the CSEM was a much easier test for students in this sample. All differences in overall difficulty Δd_j were significant in Table V. This difference was expected as the students in sample 1 were generally higher performing with higher overall CSEM post-test scores than students in sample 2. The instructional environment in sample 1 was also more controlled and enriched and should have lead to stronger learning outcomes. While most of the overall discriminations a_{j0} of the items were significantly different between samples, only items 7, 10, 15, 17, 21, 22, 23, 24, 25, and 29 had differences in overall discrimination greater than 0.2, approximately one-third of the items; therefore, most of the discrimination differences were fairly small.

In general, most items had overall discriminations a_{j0} and principle discriminations a_{jk} that suggested the items were both well functioning with positive discrimination values. Only item 7 in sample 1 had a principle discrimination less than zero suggesting that it may not be functioning correctly. While some items have principle discriminations substantially different from zero, many items had principle discriminations near zero. These items do not contribute additional information about student understanding beyond a general understanding of electricity and magnetism.

The results for the principle discriminations a_{jk} were similar. Of the 47 discrimination parameters measured, 31 were significantly different between the samples (66%), 21 were significantly different at the $p < 0.001$ level (45%); however, only 7 were different by more than 0.2 (15%). Given the differences in student population and instructional environment, the measured discrimination parameters were somewhat similar, suggesting the optimal models produced may be of general applicability.

The difference in overall difficulty for the two samples makes it challenging to interpret Table IV. To partially eliminate the effect of overall difficulty, the items have been rank ordered from lowest difficulty and lowest overall discrimination to highest in Table IV. For most items the order of difficulty was generally similar; however, items 13, 19, 21, and 27 had difficulty ranks at least 10 positions higher in sample 1 than in sample 2 (they were much easier for students in sample 1). Item 28 had a difficulty rank 10 positions higher in sample 2. In general, the average absolute difference in difficulty rank was 4.7; if items 13, 19, 21, 27, and 28 are removed the average absolute difference falls to 3.2 indicating most items were fairly close to each other in rank; difficult questions in sample 1 were also generally difficult in sample 2. Only item 29 had a difference in overall discrimination rank of over 10. The average absolute difference in overall discrimination $a_0 j$ was 2.8, which fell to 2.4 if item 29 was eliminated.

It is likely that some of the differences in the discrimination parameters were a result of the overall difference in student performance for the two samples. While superficially independent in the MIRT model, Eq. (2), the effective window on the difficulty produces correlations between difficulty and discrimination. Most difficulty parameters are between -3 and $+3$; very easy or very hard items have a limited range in which to discriminate between students. This effect can be quantified by calculating the correlation between the rank order of the difficulty and overall discrimination. For sample 1, the correlation is $r = 0.33$, a medium effect size, and, for sample 2, $r = 0.25$, a small to medium effect size.

This work replicated the method applied by Stewart *et al.* [7] to the FCI. As noted above, the optimal models for the CSEM contained more secondary principles (lemmas) than the FCI. Further, the principle discriminations a_{jk} were in general smaller than those found for the FCI. Only 7 principle discriminations were of magnitude 0.5 or greater in each of sample 1 and sample 2. For the FCI, 17 principle discriminations were greater than or equal to 0.5, showing the FCI provides better resolution of the individual principles in its optimal model.

The selection of solution path 1 in models M1-2 and M2-2 supports Leppävirta's identification of item 7 as a Newton's 3rd law item [15].

V. IMPLICATIONS

The optimal theoretical model presented in Table I is a very limited representation of the conceptual material covered by an introductory electricity and magnetism course. Many topics are missing or are weakly represented, such as the electric field of a point charge or the dipole nature of the magnetic field. Other topics are overrepresented such as the magnetic force on a stationary charge. Some items are difficult to interpret with experts producing substantially different solutions. Through the lens of the theoretical model in Table I, the CSEM seems a weak instrument for a general evaluation of electricity and magnetism. The combination of an exhaustive theoretical model extending the model in Table I to include a more complete coverage of introductory electricity and magnetism and constrained MIRT may provide the appropriate framework for creating more robust and reliably interpreted instruments. The weak coverage could be identified by the expert solution analysis alone, which can be performed during instrument development. MIRT provided confirmatory evidence that the expert model was a good model for student knowledge and allowed an optimal version of the theoretical model to be developed.

Ideally MIRT models of an instrument would have the same behavior across multiple samples; this was only partially supported by this study. The overall optimal models were very similar, differing only in a few principles. The difficulty was very different due to the substantial difference in performance of the two populations. Overall discriminations were also different, but principle discriminations were more similar. This suggests the parameters of

the MIRT models are sensitive to student population and instructional environment and cannot treated as universal. The structure of the optimal models was more general suggesting additional research will be able to identify a model that has acceptable fit for most institutions.

Ideally, discriminations should be consistent across populations, assuming the instrument was developed with a sufficiently large and academically diverse population. Difficulty will vary with the overall performance of the different student samples. There is a relation between difficulty and discrimination that arises when items are either very high scoring or very low scoring because of windowing effects. If the items average score is well away from these extremes, then discriminations should be consistent across populations because MIRT uses the relative difficulty of the items to set the scale for the ability traits θ_k. Topical coverage, however, could modify this relative difficulty ranking and produce differences in discrimination which may be the origin of the differences measured in this study.

Instructors can use the results in Table IV to further understand CSEM results. Items that have a high principle discrimination are good measures of that principle. Items that have a single principle or only one principle with a high discrimination are particularly strong measures of that principle. For example, using sample 1 and Table IV, items 3, 12, 13, 16, 17, 18, 21, and 27 all depend on a single principle and have principle discriminations that are significantly different from zero. The item difficulty d_j for these items allows the comparison of student understanding for these principles; students understand L4, Coulomb's law for the electric force, (item 3, $d_3 = 2.04$) and DF2, the definition of electric field, (item 12, $d_{12} = 2.66$) substantially better than C6, the relation of potential and field, (item 18, $d_{18} = 1.18$) and DF4, the definition of electric potential, (items 16 and 17, $d_{16} = 0.54$, $d_{17} = 1.36$). Items such as item 1, which have multiple principles but discriminate more strongly on one principle can be used to measure understanding of that principle. Items which discriminate relatively equally on multiple principles may be used to characterize understanding of that combination of principles.

The large number of principles identified (26) for a 32-item instrument meant that many principles were only included on small number of items and often mixed with other principles on the same item. This makes identifying what each individual item measures more challenging.

VI. LIMITATIONS

This work compared two large samples from two institutions. Additional samples should be tested to determine if the results are general, particularly from institutions with different student demographics than the institutions studied.

The theoretical model presented in Table I was constructed from the solutions of a small set of expert practitioners. Other models are possible and should be explored. Most experts would agree on some segments of the model but there are other segments were multiple different models are possible. This should not be viewed as the end of the modeling process for the CSEM, but as the beginning. We feel constrained MIRT is the proper tool to explore alternate models. Any researcher interested in testing a model on the datasets in this paper may request the data from the corresponding author.

VII. FUTURE WORK

This analysis technique will be extended to the Force and Motion Conceptual Evaluation (FMCE) [26]. The optimal model for this instrument can then be compared to the optimal model found by Stewart *et al.* [7] for the FCI. Further work should investigate whether the results are consistent for groups of students traditionally underrepresented in physics classes.

VIII. CONCLUSIONS

This work examined models of the CSEM for two large datasets drawn from different institutions. The optimal models identified were similar but not identical, sharing 22 of the 26 principles included in either model. The optimal models had excellent model fit characteristics for both samples. Beyond the laws, definitions, facts, and corollaries needed to define the physics content of the instrument, both optimal models also contained additional qualitative principles derived from the more general principles. The overall difficulty and discrimination of the individual items were significantly different in most cases; however, the principle discriminations were more similar. The rank ordered overall difficulties were also similar, but five items stood out as being more relatively difficult for the students in one of the samples. Therefore, while the models had many similarities, they were not identical; the optimal MIRT model for the CSEM does vary between institutions.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) as part of the evaluation of improved learning for the Physics Teacher Education Coalition, PHY-0108787. Data collection was supported by NSF Grants No. EPS-1003907 and No. ECR-1561517.

[1] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Huevelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

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

[3] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[4] 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).

[5] S. J. Pollock, Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA, AIP Conf. Proc. **1064**, 171 (2008).

[6] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).

[7] John Stewart, Cabot Zabriskie, Seth DeVore, and Gay Stewart, Multidimensional item response theory and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010137 (2018).

[8] L. J. Cronbach and P. E. Meehl, Construct validity in psychological tests, Psychol. Bull. **52**, 281 (1955).

[9] L. A. Clark and D. Watson, Constructing validity: Basic issues in objective scale development, Psychol. Assess. **7**, 309 (1995).

[10] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.020107 for exploratory MIRT factor analysis of the CSEM and the ability correlation matrices for the optimal models presented in Table IV. An extension of this table with the standard errors for all parameters is also presented.

[11] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, Mason, OH, 1986).

[12] N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. DiBello, and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, J. Eng. Educ. **104**, 454 (2015).

[13] M. Planinic, Assessment of difficulties of some conceptual areas from electricity and magnetism using the Conceptual Survey of Electricity and Magnetism, Am. J. Phys. **74**, 1143 (2006).

[14] D. E. Meltzer, Analysis of shifts in students' reasoning regarding electric field and potential concepts, AIP Conf. Proc. **883**, 177 (2007).

[15] J. Leppävirta, The effect of naïve ideas on students' reasoning about electricity and magnetism, Res. Sci. Educ. **42**, 753 (2012).

[16] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre- and post-studio transition, AIP Conf. Proc. **1179**, 173 (2009).

[17] J. M. Wilson, The CUPLE physics studio, Phys. Teach. **32**, 518 (1994).

[18] K. Kreutzer and A. Boudreaux, Preliminary investigation of instructor effects on gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **8**, 010120 (2012).

[19] A. Madsen, S. B. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

[20] D. Huffman and P. Heller, What does the force concept inventory actually measure?, Phys. Teach. **33**, 138 (1995).

[21] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[22] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).

[23] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a force concept inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).

[24] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[25] M. R. Semak, R. D. Dietz, R. H. Pearson, and C. W. Willis, Examining evolving performance on the Force Concept Inventory using factor analysis, Phys. Rev. Phys. Educ. Res. **13**, 010103 (2017).

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

[27] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, Am. J. Phys. **76**, 882 (2008).

[28] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[29] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010).

[30] S. Osborn Popp, D. Meltzer, and M. C. Megowan-Romanowicz, Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, in *Proceedings of the 2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).

[31] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

[32] 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).

[33] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the force concept inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[34] Y. Lee, D. J. Palazzo, R. Warnakulasooriya, and D. E. Pritchard, Measuring student learning with item response theory, Phys. Rev. ST Phys. Educ. Res. **4**, 010102 (2008).

[35] G. S. Gliner, College students' organization of mathematics word problems in relation to success in problem solving, School Sci. Math. **89**, 392 (1989).

[36] G. S. Gliner, College students' organization of mathematics word problems in terms of mathematical structure vs. surface structure, School Sci. Math. **91**, 105 (1991).

[37] B. S. Eylon and F. Reif, Effects of knowledge organization on task performance, Cognit. Instr. **1**, 5 (1984).

[38] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[39] A. H. Schoenfeld and D. J. Herrmann, Problem perception and knowledge structure in expert and novice mathematical problem solvers, J. Exp. Psychol. Learn. **8**, 484 (1982).

[40] F. Reif and J. I. Heller, Knowledge structure and problem solving in physics, Educ. Psychol. **17**, 102 (1982).

[41] I. D. Beatty and W. J. Gerace, Probing physics students' conceptual knowledge structures through term association, Am. J. Phys. **70**, 750 (2002).

[42] G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychol. Rev. **63**, 81 (1956).

[43] M. T. H. Chi and J. D. Slotta, The ontological coherence of intuitive physics, Cognit. Instr. **10**, 249 (1993).

[44] M. T. H. Chi, J. D. Slotta, and N. De Leeuw, From things to processes: A theory of conceptual change for learning science concepts, Learn. Instr. **4**, 27 (1994).

[45] J. D. Slotta, M. T. H. Chi, and E. Joram, Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change, Cognit. Instr. **13**, 373 (1995).

[46] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[47] A. A. diSessa and B. L. Sherin, What changes in conceptual change?, Int. J. Sci. Educ. **20**, 1155 (1998).

[48] D. Hammer, Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions, J. Learn. Sci. **5**, 97 (1996).

[49] A. A. diSessa, N. M. Gillespie, and J. B. Esterly, Coherence versus fragmentation in the development of the concept of force, Cogn. Sci. **28**, 843 (2004).

[50] R. J. Dufresne, W. J. Leonard, and W. J. Gerace, Making sense of students' answers to multiple-choice questions, Phys. Teach. **40**, 174 (2002).

[51] R. N. Steinberg and M. S. Sabella, Performance on multiple-choice diagnostics and complementary exam problems, Phys. Teach. **35**, 150 (1997).

[52] A. Newell and H. A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, NJ, 1972).

[53] S. Ohlsson, The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm, J. Prob. Solving **5**, 7 (2012).

[54] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, Science **208**, 1335 (1980).

[55] J. H. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Models of competence in solving physics problems, Cogn. Sci. **4**, 317 (1980).

[56] C. Hieggelke and T. O'Kuma, The impact of physics education research on the teaching of scientists and engineers at two-year colleges, AIP Conf. Proc. **399**, 267 (1997).

[57] Physport, https://www.physport.org. Accessed 8/8/2017.

[58] U.S. News & World Report: Education, US News and World Report, Washington, DC, https://premium.usnews.com/best-colleges. Accessed 4/30/2017.

[59] The Carnegie Classification of Institutions of Higher Education, Center for Postsecondary Research, Indiana University School of Education, Bloomington, IN, http://carnegieclassifications.iu.edu/. Accessed 9/21/2017.

[60] V. Otero, S. Pollock, and N. Finkelstein, A physics department's role in preparing physics teachers: The Colorado learning assistant model, Am. J. Phys. **78**, 1218 (2010).

[61] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Prentice Hall, Upper Saddle River, NJ, 1998).

[62] W. J. van der Linden, Unidimensional Logistic Response Models, in *Handbook of Item Response Theory*, Vol. 1 (CRC Press, Taylor & Francis Group, New York, NY, 2016), pp. 13–30.

[63] L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Struct. Equ. Modeling **6**, 1 (1999).

[64] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach* (Springer-Verlag, New York, NY, 2003).

[65] McElreath, *Statistical Rethinking: A Baysian Course with Examples in R and Stan* (CRC Press, Taylor & Francis Group, Boca Raton, FL, 2016).

[66] A. E. Raftery, Bayesian model selection in social research, Socio. Meth. **25**, 111 (1995).

[67] R. B. Kline, *Principles and practices of structural equation modeling*, 4th ed. (Guilford Publications, New York, NY, 2016).

[68] K. Schermelleh-Engel, H. Moosbrugger, and H. Müller, Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures, Meth. Psychol. Res. Online **8**, 23 (2003).

[69] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2013).

[70] R. P. Chalmers, mirt: A multidimensional item response theory package for the R environment, J. Stat. Softw. **48**, 1 (2012).

[71] A. Canty and B. D. Ripley, Boot: Bootstrap R (S-Plus) Functions (2017), R package version 1.3–20.

[72] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications* (Cambridge University Press, Cambridge, England, 1997).

[73] S. Epskamp, A. O. J. Cramer, J. L. Waldorp, V. D. Schmittmann, and D. Borsboom, qgraph: Network visualizations of relationships in psychometric data, J. Stat. Softw. **48**, 1 (2012).