Multidimensional item response theory and the force and motion conceptual evaluation

Jie Yang, Cabot Zabriskie, and John Stewart*

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

(Received 6 June 2019; published 4 November 2019)

Many studies have examined the structure and properties of the Force Concept Inventory (FCI); however, far less research has investigated the Force and Motion Conceptual Evaluation (FMCE). This study applied Multidimensional Item Response Theory (MIRT) to a sample of $N = 4528$ FMCE post-test responses. Exploratory factor analysis showed that 5, 9, and 10-factor models optimized some fit statistics. The FMCE uses extensive blocking of items into groups with a common stem; these blocks factored together in most models. A confirmatory analysis, which constrained the MIRT models to a theoretical model constructed from expert solutions, produced a model requiring only 8 principles, fundamental reasoning steps. This was substantially fewer than the 19 principles identified in the FCI by a previous study. Correlation analysis also demonstrated that the two instruments were very dissimilar. The reduced number of principles and the repetition of items using a single principle allowed the extraction of eight single-principle subscales, seven with Cronbach's alpha greater than the 0.7 required for acceptable internal consistency. The differences between the FCI and the FMCE suggest that the two instruments could provide complementary, but different, information about student understanding of Newton's laws with the FCI measuring an integrated Newtonian force concept and the FMCE measuring details of that force concept.

DOI: 10.1103/PhysRevPhysEducRes.15.020141

I. INTRODUCTION

The Force and Motion Conceptual Evaluation (FMCE) was introduced in 1998 to measure the understanding of force and motion in one dimension [1]. The FMCE was developed following the success of the Force Concept Inventory (FCI) [2]. The FCI was critical in demonstrating that traditional instruction did little to improve conceptual understanding [3]. The FCI and the FMCE have been exceptionally important to the development of physics education research (PER). For an overview of the role of these instruments in the development of PER, see the synthesis of Doctor and Mestre [4].

A. Prior studies

This work replicates two prior studies which applied constrained Multidimensional Item Response Theory (MIRT) to the FCI and the Conceptual Survey of Electricity and Magnetism (CSEM) [5]. These studies will be referenced as study 1 and study 2 in this work.

1. Study 1

Study 1 examined the item-level structure of the FCI [6]. Both exploratory analysis using unconstrained MIRT to perform an exploratory factor analysis (EFA) and confirmatory analysis constraining the MIRT models to a theoretical model developed from expert solutions were presented. EFA identified a 9-factor solution as optimal on some but not all fit statistics. Much of the factor structure was shown to be related to the practice of "blocking" items. We define an item block as a group of items each referring to a common stem. An item's "stem" defines the physical system but does not pose a question. Each item in an item block can then refer to this physical system. For example, FCI items 21 to 24 all refer to a rocket that turns on its engine while in space. Study 1, then, constructed a model of the FCI from solutions collected from content experts. These solutions were decomposed into individual reasoning pieces, called principles. Beyond the laws and definitions which define Newtonian mechanics, the expert solutions contained many secondary qualitative principles derived from these laws such as "if the acceleration and the velocity are in the same direction, the object speeds up." The expert model was then fit to a large dataset by constraining the MIRT parameter matrix to the model. A small number of theoretically motivated changes to the expert model were also fit, allowing the selection of an optimal model of student reasoning about Newtonian mechanics. The secondary principles were not included

*jcstewart1@mail.wvu.edu

in the optimal model. The optimal MIRT model was also shown to be substantially superior to the model proposed by the authors of the FCI. The optimal model provides a fine-grained picture of the reasoning required to solve the FCI and the interconnection of items in the FCI. This model will be used in the present work to examine differences between the FCI and the FMCE.

2. Study 2

Study 2 applied the methods introduced in study 1 to the CSEM using datasets drawn from 2 different institutions [7]. The optimal models identified were similar but not identical suggesting that while some differences exist, the optimal models may be representative of student thinking at a number of institutions. Unlike the model of the FCI in study 1, both optimal models included some of the secondary principles.

This study informs the present work by showing that the results of MIRT have some generality across institutions, as well as showing that for some instruments the optimal models do contain secondary principles.

B. Research questions

The current work replicated both the exploratory and confirmatory MIRT analyses performed in studies 1 and 2 for the FMCE. This produced a detailed model of the knowledge measured by the instrument and allowed a more thorough comparison of the FCI and the FMCE than was previously possible.

In this paper, we seek to answer the following research questions:

RQ1 What is the optimal model of the FMCE identified using exploratory factor analysis? To what extent does the blocking of items explain the factor structure?

RQ2 What is the optimal model of the FMCE using constrained MIRT?

RQ3 How do the number of principles and connectivity of the principles in the optimal MIRT models of the FMCE and the FCI compare?

C. Item level analysis of the FMCE

Relatively little item-level analysis has been performed on the FMCE unlike the more thoroughly studied FCI. Multiple qualitative subdivisions of the instrument have been proposed. Thornton and Sokoloff proposed four subgroups of items: "Force Sled" questions (items 1–7), "Cart on a Ramp" questions (items 8–10), "Coin Toss" questions (items 11–13), and "Force Graph" questions (items 14–21) [1]. The subgroup of items 27, 28, and 29 is represented by the notation 27–29. They identified some items as problematic: items 5, 6, and 15. This led to the modification of the subgroups; the force sled questions became items 1–4 and 7 while the force graph questions became items 14 and 16–21.

After the instrument's initial publication, Thornton *et al.* provided additional analysis which ultimately lead to the suggestion of an alternate scoring scheme [8]. This scoring scheme combined groups of items into clusters which received two points if all the items in the cluster were correct and zero points if any were incorrect. The clusters identified were items 8_10, 11_13, and 27_29. A cluster of items 27, 28, and 29 that is scored together is represented by the notation 27_29. Six items were also eliminated from the instrument in this analysis (items 5, 15, 33, 35, 37, and 39) because students without a Newtonian understanding of physics often answered them correctly. Item 6 was also eliminated because content experts often answered it incorrectly.

The subgroups introduced by Thornton and Sokoloff have been further investigated and refined [9,10]. Wittmann suggested a subdivision of the instrument into five subgroups "Velocity" (items 40–43), "Acceleration" (items 22–29), "Force (Newton I and II)" (items 1–4, 7–14, 16–21), "Newton III" (items 30–32, 34, 36, 38), and "Energy" (items 44–47) [9]. By applying a resource framework, Smith and Wittmann proposed a set of subgroups refining Wittmann's subgroups: Force sled (items 1–4, 7), reversing direction (items 8–13, 27–29), force graphs (items 14, 16–21), acceleration graphs (items 22–26), Newton III (items 30–32, 34, 36, 38), velocity graphs (items 40–43), and energy (items 44–47) [9]. Note, all subgroups eliminate the items removed in the revised scoring. Smith, Wittmann, and Carter later used the revised subgroup structure to allow a deeper analysis of the effect of instruction [10].

Ramlo explored the reliability and factor structure of the FMCE [11]. For a sample of 146 students, the pretest had Cronbach's alpha of 0.742 and the post-test 0.907 showing the instrument had excellent reliability when used as a post-test. Ramlo also examined the instrument with exploratory factor analysis. The pretest factor structure consisted of three factors. Each factor contained a mix of items measuring different concepts and the same concept was distributed among more than one factor. Thus, Ramlo concluded that the FMCE pretest factor structure was undefined. The post-test factor structure contained three factors. These factors generally contained items testing the same concept with 21 out of 29 questions measuring Newton's 1st or 2nd law in factor 1, 8 out of 10 questions associated with Newton's 3rd law in factor 2, and 8 out of 8 questions related to velocity in factor 3.

Talbot [12] investigated the item-level difficulty and discrimination of FMCE items finding some items outside of the preferred range established in classical test theory [13]. Items 11, 12, 36, and 38 were problematic on the pretest; many items were problematic on the post-test because of a ceiling effect. The sample was very high performing with 51 of 336 students receiving perfect scores. Recent studies of the FMCE have explored the ranking of incorrect responses to show how student ideas

TABLE I. MIRT fit statistics for an exploratory factor analysis of the FMCE.

Factors	AIC	BIC	RMSEA	TLI	CFI
1	170 746	171 298	0.109	0.90	0.90
2	161 277	162 099	0.086	0.94	0.94
3	156 928	158 013	0.074	0.95	0.96
4	152 050	153 392	0.055	0.97	0.98
5	149 663	151 255	0.053	0.98	0.98
6	148 329	150 164	0.126	0.86	0.90
7	147 744	149 817	0.128	0.86	0.90
8	147 345	149 649	0.130	0.85	0.91
9	147 088	149 617	0.131	0.85	0.91
10	147 037	149 784	0.158	0.78	0.88
11	147 146	150 104	0.144	0.82	0.90
12	147 250	150 414	0.161	0.78	0.89

develop over the span of an introductory mechanics course [14] and developed a hierarchy of responses [15]. Item response theory was used to explore the responses to item 18 and develop a hierarchy of responses with the correct answer (B) the best response, the incorrect response (A) second best and responses (D), (G) or (H) weaker than either (A) or (B) with responses (C) and (F) the weakest [15].

Thornton *et al.* [8] performed a detailed comparison of pretest and post-test results of the FCI and the FMCE for a large sample of students at many institutions and found a strong correlation between the results of the two instruments ($r = 0.78$). They, however, note that the coverage of the two instruments is different with 22 of the 30 FCI items outside of the coverage of the FMCE.

Henderson *et al.* [16] performed an analysis of the item-level fairness of the FMCE for men and women. The majority of the FMCE items were significantly more difficult for women; however, few items were substantially unfair.

D. Factor analysis of the FCI

Many studies have examined the factor structure of the FCI; however, no consensus on the structure of the instrument has emerged from these efforts. Huffman and Heller identified only one factor for a sample of college students [17]. This was substantially different than the structure suggested by the authors of the FCI [2], leading to a lively debate about what the instrument actually measured [17–19]. In more recent work, Scott, Schumayer, and Gray reported a model with 5 factors [20], Semak *et al.* found 6 factors [21], and study 1 (Table I [6]) found 9 factors. The 5, 6, and 9-factor models have some similarities but are not identical. Study 1 demonstrated that much of the identified factor structure could be attributed to the blocking of items and to the existence of a few repeated groups of conceptually similar items.

E. Item Response Theory

While many studies have applied Item Response Theory (IRT) to the FCI, little work has investigated the FMCE. Unidimensional IRT uses a single ability parameter to model a student's facility with the material while MIRT extends the unidimensional model with multiple ability parameters. Several studies have reported unidimensional IRT models of the FCI [22–25]. Item characteristic curves (ICC) plot the IRT response models. The ICCs and model parameters reported in these works showed that FCI items were generally well functioning with positive discriminations in the desired range. Unidimensional IRT has also been used to investigate the gender fairness of the FCI [25,26] with multiple studies reporting many unfair items. The majority of the unfair items were unfair to women. More recently MIRT has been used to perform EFA on the FCI finding the instrument contained from 5 to 9 factors [6,27].

F. The structure of knowledge

The current work produced a fine-grained model of the reasoning needed to solve FMCE items. This model is similar to those pioneered Simon and Newell [28] to understand complex problem solving. Their methodology, which built computationally functional models of reasoning, was central to problem-solving research for decades [29]. Larkin *et al.* applied this methodology to understand expert and novice differences in kinematics and dynamics [30,31]. This methodology eventually lost favor because it could not be used to understand general problem-solving strategies. Reif and Heller created a related model of expert problem solving in mechanics [32]; however, their model was not computationally functional. The expert model of the FMCE constructed in the current study shares many features with the models of Larkin *et al.* [31] and the model of Reif and Heller [32].

The model of Newtonian mechanics produced in the current work, as well as that of Reif and Heller or Larkin *et al.*, all represent only the correct Newtonian model. Substantial work has shown that students also have robust misconceptions and novicelike habits that are important in their solution (often incorrectly) of physics problems.

Multiple theoretical frameworks have been constructed to explain differences in expert-novice problem solving. One framework explains expert-novice differences by the existence of "misconceptions," common sense beliefs about how the world works developed through life experiences [33–35]. Another framework proposes "knowledge in pieces" where student understanding consists of a number of small segments of reasoning that are activated to solve problems [36–38]. The "ontological categories" framework explains novice reasoning as students categorizing knowledge into incorrect broad categories [39–41]. Hammer proposed a unification of the misconception and knowledge-in-pieces frameworks by introducing the concept of a resource [38,42,43].

See study 1 [6] for a more complete summary of the application of factor analysis and IRT to the FCI and a more thorough exploration of the structure of student knowledge.

II. METHODS

A. The FMCE and the FCI

The FMCE was constructed to test students' understanding of Newtonian mechanics in one dimension. The original instrument contained 43 items; a revised instrument added 4 items to measure the understanding of energy. These items are often not included in the score of the FMCE and are not included in the analysis in the current study. Each item has a minimum of six possible responses with some items having nine responses. All items include a "none of the above" response which may cause psychometric problems [44]. Since all items include a none of the above response, each item should suffer from having a distractor which is preferentially not selected by the student; thus no item will be affected more than other items. This should have less effect than in other commonly used conceptual instruments because of the large number of distractors used in most items. The none of the above response is not the correct answer for any item which may serve to limit its negative effects. The instrument groups items into 9 blocks where all items in each block refer to a common stem. Only one item is not included in a problem block. The revised version of the FMCE is available at PhysPort [45].

The FCI is a 30-item multiple-choice instrument designed to measure a student's Newtonian force concept. Each item has five possible responses. The items in the FCI were developed to both probe Newtonian knowledge and common misconceptions. The instrument uses limited blocking of items with 13 of the 30 items in item blocks.

Both the FMCE and the FCI cover one-dimensional kinematics and Newton's laws, while the FCI also includes two-dimensional motion under constant acceleration (parabolic motion), impulsive forces, and circular motion. Unlike the FCI, the FMCE includes many items requiring the interpretation of graphs.

B. Sample

The sample for this study was collected during 13 semesters from spring 2011 to spring 2017 at a large eastern land-grant university serving approximately 30 000 students. The undergraduate demographics of the institution were 79% White, 7% international, 4% African American, 4% Hispanic, 4% students reporting two or more races, with other groups 1% or less [46]. The ACT scores of the institution ranged from 21 to 26 (25th to 75th percentile). The sample included 3719 FMCE post-test responses (80% men) collected in the introductory, calculus-based mechanics class taken by scientists and engineers. This sample was analyzed in an earlier work [16] where it was referenced as sample 3A. The class was presented by a variety of instructors; most used some form of Peer Instruction in the lecture. The course also required the students attend a laboratory each week; the laboratory session presented a variety of interactive engagement activities. The instructional environment for the sample was discussed in detail in the previous study.

C. Item Response Theory

Item Response Theory produces statistical models of a student's responses to a test [47]. The unidimensional IRT model employs the logistic function to model the probability of answering an item correctly as a function of a single latent trait called "ability." Many unidimensional IRT models have been used in previous PER studies. The most closely related to classical test theory [13] is the two-parameter logistic (2PL) model. The 2PL model uses two item-level parameters: the item difficulty b_j and the item discrimination a_j, where j is the item number. It assumes that the probability π_{ij} of a student i correctly answering an item j is given by the logistic function

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

where the latent trait θ_i measures the general ability of student i to answer any item correctly.

D. Multidimensional Item Response Theory

Multidimensional Item Response Theory is an extension of unidimensional IRT which uses multiple ability parameters for each student and multiple discrimination parameters for each item. The ability of student i becomes the k component vector $\boldsymbol{\theta}_i$ where each element of the vector measures a different dimension of the student's ability with mechanics; k latent ability traits are estimated for each student. The item discrimination $\boldsymbol{a}_j$ is also modeled as a vector with k components where j represents the item number. The MIRT model of the probability π_{ij} of a student i correctly answering an item j 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 the overall difficulty of the item.

E. Model fit statistics

The likelihood function L of a MIRT model represents the probability that a specific observation occurred assuming the model in Eq. (2). To fit the MIRT model, maximum likelihood (ML) estimation techniques are used to find the values of the parameters which maximize L. To determine if the model fits the data well, several statistics have been developed. Hu and Bentler recommend using multiple statistics to evaluate models [48]. In this work, we 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).

AIC and BIC are minimized by the optimal model; BIC penalizes the addition of parameters more strongly than AIC. A difference of greater than 2 is considered significant for the AIC [49]. For the BIC, Raftery suggested a difference of 2 as "weak," from 2 to 6 as "positive," from 6 to 10 as "strong," and above 10 as "very strong" [50]. The AIC and BIC are mathematically very similar; we will also use Raftery's criteria for the AIC.

The root mean square error of approximation is a statistic with values ranging from 0 to 1 that measures the badness of fit. Well fitting models should have RMSEA less than 0.05 while poor fitting models have RMSEA greater than 0.1. CFI and TLI are incremental-fit statistics which measure the difference of the tested model and a null model. For good model fit, modern criteria suggest CFI greater than 0.95 or TLI greater than 0.95. See study 2 or Eaton and Willoughby for additional information on fit statistics [24].

III. RESULTS

This study reports both exploratory and confirmatory analyses of the FMCE. Exploratory methods do not proceed from a theoretical model and allow the model to emerge from the data. Confirmatory methods proceed from a theoretical model and explore whether the model is supported by the data. A substantial body of evidence suggests that confirmatory methods are less likely to yield spurious results [51,52]. While both study 1 and study 2 cautioned about the dangers of using purely exploratory methods, virtually no exploratory results for the FMCE have been reported, unlike the wealth of exploratory research into the FCI. Exploratory methods were also productive in study 1 to help understand the effects of blocking on the properties of the instrument. As such, an exploratory factor analysis of the instrument is reported. A correlation analysis was then performed to help understand the factor structure. The confirmatory constrained MIRT methodology introduced in study 1 was then applied to the FMCE. Expert solutions of the FMCE were used to construct a theoretical model of the reasoning needed to solve the instrument. An optimal confirmatory MIRT model was constructed from this initial theoretical model by making small, theoretically motivated changes to the expert model. The optimal model suggested that the FMCE might be productively divided into subscales; the properties of these subscales were calculated.

A. Exploratory factor analysis with MIRT

An exploratory factor analysis of the FMCE was performed using MIRT. Models with 1 to 12 factors were fit; not all model fit statistics were optimized for the same model. Table I shows the model fit statistics for each number of factors. While the 10-factor model minimized AIC, a 9-factor model minimized BIC. These models had relatively poor RMSEA, CFI, and TLI. A 5-factor model had superior RMSEA, CFI, and TLI statistics. As such, as in study 1 (Table II [6]), the fit statistics did not clearly select a single optimal model. The inconsistent identification of the best model by different fit statistics is a result of the different goals that went into the creation of the statistics. AIC and BIC are related to the absolute fit of the model with some penalty for the addition of parameters (BIC penalizes additional parameters more strongly). Both CFI and TFI are comparative indices which compare model fit with a null model (a model assuming the items are uncorrelated). RMSEA is a badness-of-fit statistic designed to detect poorly constructed models. Table II presents the 5-factor model (varimax rotation) and Table III the 10-factor model, the most fully resolved model which maximized any fit statistic. The first column of each table shows the item number. Bolded item numbers represent the first item of an item block. Factors are reported as columns and labeled "FC." For both the 5-factor and 10-factor models, items in the same block generally have their highest loading on the same factor. The last column of the table reports the difficulty d; easier items have larger d. The 10-factor model had two factors (FC8 and FC9) which did not load strongly on any item. It also had one factor, FC10, which had similar loadings to FC4 for the group of items 27–29. This factor seemed to be splitting the subgroup 8–13 and 27–79 into two subgroups. Factor FC6 also seemed to split the subgroup of items 30–39 in FC2 extracting the block of items 35–38. Neither of these splittings were successful in that the loadings in the original factor were generally commensurate to those of the new factor. The constrained MIRT analysis which follows will suggest these divisions are inappropriate; these additional factors may have resulted from the blocking of the instrument. The 5-factor model failed to separate the subgroup of items 40–43 from the subgroup of items 22–26 in FC3; constrained MIRT will also suggest this is inappropriate.

While not as detailed, many of the fit statistics suggest the 5-factor model presented Table II as the superior model. The problems noted above suggest the 10-factor model should not be selected as the optimal model. The 5-factor model combines some of the factors in the 10-factor model; the optimal constrained MIRT model allows further exploration of the relation of the 5-factor and 10-factor models and is presented in the Sec. IV.

Within the lens of the optimal MIRT model, the 5-factor model seems to generally make conceptual sense. Factors FC1 and FC5 both load most strongly on items testing Newton's 1st and 2nd law, with factor FC1 requiring more graphical reasoning than FC5. Factor FC2 loads most strongly on Newton's 3rd law items. Factor FC3 loads most strongly on items using graphical reasoning to apply the definition of acceleration. Factor FC4 loads most

TABLE II. Factor loadings for the 5-factor model using exploratory factor analysis with Multidimensional IRT (varimax rotation). The bolded item numbers represent the start of item blocks which are also separated by horizontal lines. Loadings of magnitude greater than 0.3 are shown.

FMCE No.	FC1	FC2	FC3	FC4	FC5
1	0.72				0.56
2	0.67				0.54
3	0.42			0.31	0.65
4	0.68				0.56
5	0.44				0.47
6	0.38			0.39	0.50
7	0.35			0.31	0.65
8	0.34			0.81	
9	0.31			0.74	
10	0.31			0.71	
11	0.32			0.83	
12				0.80	
13	0.33			0.80	
14	0.86		0.34		
15			0.38		
16	0.88				
17	0.85		0.35		
18	0.78		0.33	0.31	
19	0.80			0.31	
20	0.53			0.31	
21	0.39		0.39	0.45	
22	0.47		0.76		
23	0.46		0.71		
24	0.42		0.82		
25	0.46		0.66		
26	0.42		0.81		
27			0.44	0.66	
28			0.40	0.65	
29			0.39	0.66	
30		0.89			
31		0.79			
32		0.92			
33		0.53	0.42		
34		0.91			
35		0.51			
36		0.85			
37		0.36	0.32		
38		0.85			
39		0.69			
40			0.56		0.37
41			0.41		
42		0.33	0.39		0.31
43			0.32		

strongly on items involving motion under gravity. The only substantial difference between 5-factor model and the constrained MIRT model is the failure to resolve items 40 to 43 as a separate factor. This may be a result of the generally weak properties of these items; when these items were used as a subscale, they had substantially weaker internal consistency than the other subscales.

B. Correlation analysis

To further understand the structure identified by EFA, the correlation and partial correlation matrices were calculated. Figure 1(a) presents a visualization of the FMCE correlation matrix created with the "R" qgraph package [53]. Solid lines (green) represent positive correlations greater than 0.3 (Cohen's criteria for medium effect size); the thickness of the lines represent the magnitude of the correlation coefficient. There were no correlations less than −0.3. The nodes are placed for visual effect only. The correlation matrix has a clear clustered structure that largely follows the structure of the item blocks with items in the same block strongly correlated. The correlation matrix also provides evidence of structure beyond item blocking with clusters within the matrix formed of multiple item blocks. The correlation matrix is strikingly different from that of the FCI published in study 1 (Fig. 1 [6]). The FCI correlation matrix was sparsely connected, while the FMCE matrix in Fig. 1(a) contains two tightly connected subgroups; one of which contains the majority of the items in the instrument.

Figure 1(b) shows the FMCE partial correlation matrix controlling for total FMCE score; correlations with $r > 0.1$ (Cohen's criteria for a small effect) are shown. The partial correlation matrix is calculated by first regressing the total FMCE score on the item score, then calculating the correlation matrix of the residuals of these regressions. Items within an instrument may be correlated because high performing students tend to answer most items correctly; a partial correlation matrix corrects for this effect. The blocked structure of the instrument is evident in Fig. 1(b), which contains groups of highly positively correlated items within some item blocks; these items are negatively correlated to items in other blocks. The combination of item blocks into larger groups (for example, items 8–13 and 27–29) is also supported. In general, the negative correlations were smaller than the positive correlations, and therefore, no thick red dashed lines are shown. The negative correlations presented in Fig. 1(b), but not in 1(a), represent pairs of items that are anticorrelated after correcting for total test score. While items within item blocks still vary together after correcting for total test score, many blocks of items are anticorrelated with items in other blocks after correcting for the total test score; this may result from the item blocks testing different physical concepts.

C. Theoretical model

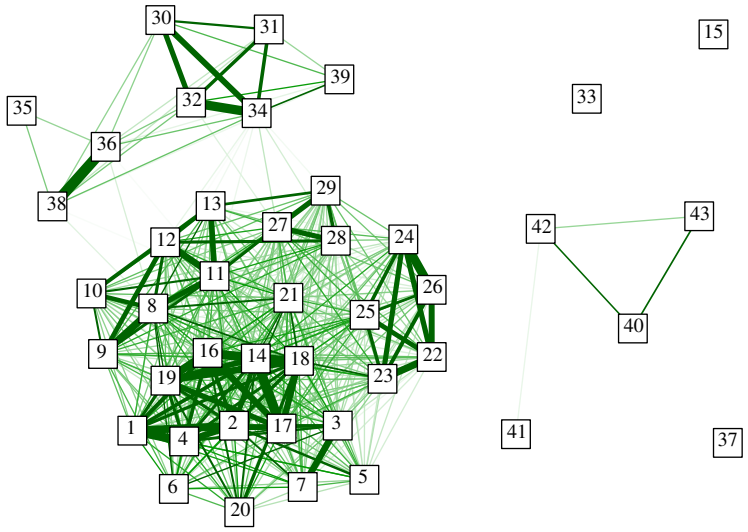
Study 1 introduced a methodology for producing a theoretical model of an instrument from expert solutions of the instrument. First, solutions are collected from a set of content experts. These solutions are textually decomposed into small fragments representing independent

TABLE III. Factor loadings the for 10-factor exploratory factor analysis performed with multidimensional IRT (varimax rotation). The bolded item numbers represent the start of item blocks which are also separated by horizontal lines. Loadings greater than 0.3 are shown. Larger values of d represent easier items. The horizontal header is repeated after item 21 to improve readability.

FMCE No.	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10	d
1	0.46						0.78				0.62
2	0.45						0.71				0.11
3			0.42				0.68				0.38
4	0.43						0.76				0.38
5							0.53				0.07
6				0.36			0.54				0.25
7			0.44				0.64				0.22
8				0.83							1.67
9				0.77					0.33		0.75
10				0.74							0.73
11				0.82							0.54
12				0.79							−0.02
13				0.80							0.55
14	0.81				0.35						−2.31
15			0.33					0.38			−0.17
16	0.83						0.32				−1.23
17	0.82				0.36						−1.73
18	0.72				0.35		0.31				−0.24
19	0.75						0.31				−0.68
20	0.47						0.34				−0.04
21				0.43	0.4		0.33				0.14
FMCE No.	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10	d
22	0.31				0.83						0.86
23	0.31				0.78						0.56
24	0.32				0.83						0.52
25	0.35				0.71						0.18
26	0.32				0.82						0.50
27				0.52	0.40					0.53	−1.29
28				0.53	0.31					0.57	−1.70
29				0.54	0.35					0.49	−1.17
30		0.85									0.08
31		0.76									−0.06
32		0.9									0.43
33		0.58	0.37								−0.04
34		0.89									0.57
35		0.34				0.44					0.00
36		0.46				0.84					0.00
37		0.34									0.00
38		0.46				0.84					0.00
39		0.62									0.00
40			0.55		0.42						0.00
41			0.34		0.32						0.00
42		0.31	0.31		0.31						0.00
43											0.00

pieces of reasoning. Fragments representing similar physical reasoning steps are grouped; each group is then identified with the general reasoning used. These groups are called principles. For the current study, expert solutions were collected from the lead instructor and the research team. Table IV shows the resulting model for the FMCE. Study 1 introduced a taxonomy of these principles:

(a) Correlation Matrix



(b) Partial Correlation Matrix

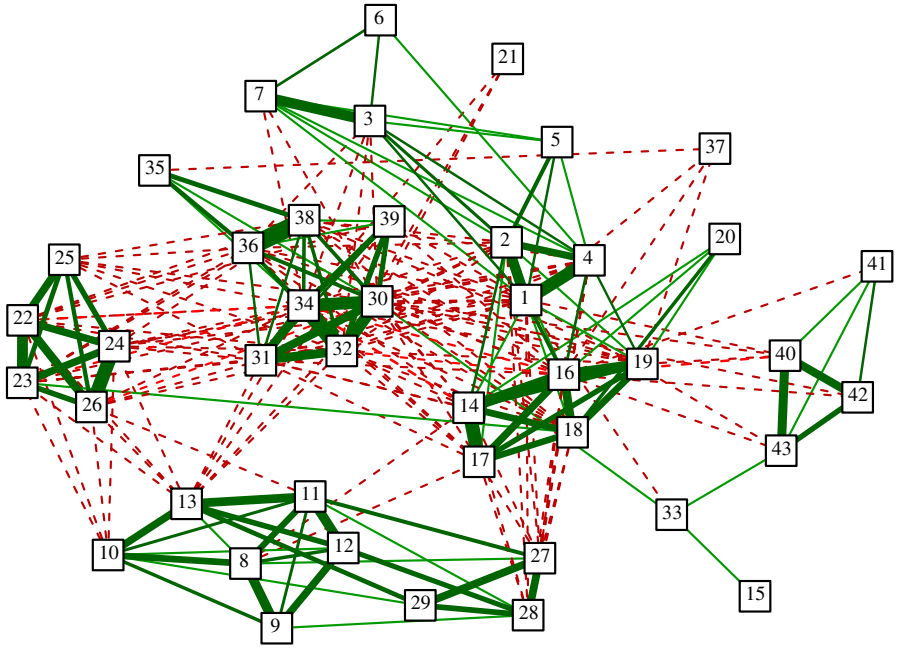


FIG. 1. (a) FMCE correlation matrix ($|r| > 0.3$) and (b) FMCE partial correlation matrix ($|r| > 0.1$). Line thickness represents the size of the correlation. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations.

Definitions (DF) Definitions of physical quantities. For example, the definition of velocity $\vec{v} = d\vec{r}/dt$.

Laws (L) Physical laws. For example, Newton's 1st law.

Facts (F) Facts about the universe that are not as general as a physical law. For example, a surface exerts a normal force on an object.

Results (R) Primary results derived from the laws, facts, and definitions specialized to some physical situation. For example, the three-dimensional constant acceleration kinematic equations.

Corollaries (C) Secondary results derived from laws, facts, results, and definitions. For example, the instantaneous velocity is tangent to the trajectory.

Lemmas (LM) A qualitative specialization of a principle to a subset of items. For example, "if the force is in the direction of motion, the object speeds up."

Reasoning Steps (RS) Reasoning not specifically related to physics. For example, reading a graph.

The FMCE models did not require any results or corollaries. Each principle in Table IV is labeled with an

TABLE IV. Theoretical model of Newtonian mechanics as tested by the FMCE. Principles in bold were included in the optimal model.

Label	Derived From	FMCE No.	Principle
			Kinematics
DF1		40–43	Definition of velocity ($\vec{v} = d\vec{r}/dt$)
DF2		22–26	Definition of acceleration ($\vec{a} = d\vec{v}/dt$)
			Dynamics
L1	L2	2, 5, 14, 15, 17	Newton's 1st law
L2		6	Newton's 2nd law
L3		30–39	Newton's 3rd law
LM1	L2, DF2	1, 4, 16, 19, 20	A force in the direction of motion causes objects to speed up.
LM2	L2, DF2	3, 7, 18, 20	A force opposite the direction of motion causes objects to slow down.
			Properties of forces
L4		11–13, 27–29	Objects near Earth's surface experience an approximately constant force of gravity toward the center of Earth.
LM3	L4	8–10	An object on an incline experiences a constant net force down and parallel to the incline.
F8		21	There is no contact force on an object after contact is lost.
			Other
RS1		14–20, 22–26, 40–43	Reading a graph.

abbreviation for the type of principle and a number. The items with labels printed in bold are the principles included in the optimal constrained MIRT model. When possible, the principles are labeled consistently with study 1. Study 1 identified seven facts used in the solution to the FCI; none were required in the FMCE solution. As such, the fact identified in the FMCE solution was labeled F8. Some principles, called secondary principles, can be derived from more fundamental principles. Each secondary principle is also labeled with the set of more fundamental principles from which it was derived.

While most principles in Table IV are fundamental to Newtonian mechanics, a few require additional explanation. The reasoning step, RS1, was added because there is a strong delineation in the FMCE where a substantial subset of the items use graphs for the answer choices. As such, RS1 was added as an independent principle. Fact F8 (there is no contact force after contact is lost) was added because item 21 seems to explicitly probe student understanding of this fact. Lemma LM3 (force down an incline) is shown as directly derived from L4 (the law of gravitation); however, a complete derivation would involve introduction of the normal force and a resolution of the sum of the force of gravity and the normal force into a force down the plane. The FMCE does not include other items involving the normal force or a resolution of a sum of forces in differing combinations; therefore, these additional principles could not be resolved.

D. Model transformation plan

A confirmatory analysis proceeds by first fitting a theoretical model and then by exploring a small set of theoretically motivated transformations of that model. By starting with and relying on a theoretical model, confirmatory methods are less susceptible to mistaking random fluctuations in the data for real effects.

RS1 does not specifically involve applying reasoning unique to physics; we tried removing it from the model. The optimal model in study 1 (Tables III and VII [6]) contained no lemmas. The optimal models in study 2 contained fewer lemmas than the initial expert model. As such, lemmas were removed from the model by replacing each lemma with the more fundamental principles from which it was derived as identified in Table IV. Finally, Newton's 1st and 2nd laws are related. We attempted, first, to replace Newton's 1st law with Newton's 2nd law and the definition of acceleration. We, then, replaced Newton's 1st law with Newton's 2nd law alone. Finally, LM3 and L4 are related; we tried combining them. Table V shows the process of transforming the model.

E. The optimal model

The model transformation process and the superior models selected at each stage are shown in Table V. The starting point, model 1, was the theoretical model shown in Table IV.

To map the theoretical model onto the MIRT model, the MIRT parameter matrix is constrained so that only the principles that theoretically (based on Table IV) should be involved in the solution of the item are nonzero. The model is then transformed by replacing secondary principles with more fundamental principles according to the model transformation plan. For RS1, this involved removing the latent ability trait and the item discrimination associated with RS1 from the model. For the other transformations, the items that "loaded" on the principle to be removed were set to load on the principles from which it was derived. The term "load" is used in analogy to the factor loadings in EFA.

TABLE V. The MIRT model transformation process.

Transformed model	Transformation	Original model	AIC	BIC	RMSEA	TLI	CFI	Superior model
1			150 549	151 486	0.05	0.98	0.98	
2	Remove RS1.	1	152 101	152 935	0.05	0.98	0.98	1
3	Replace LM1 and LM2 with L2 and DF2.	1	150 827	151 808	0.05	0.98	0.98	1
4	Replace LM1 with L2 and DF2.	1	150 468	151 437	0.05	0.98	0.98	4
5	Replace LM2 with L2 and DF2.	1	150 617	151 580	0.05	0.98	0.98	1
6	Replace L1 with L2 and DF2.	4	150 033	151 034	0.05	0.98	0.98	6
7	Replace L1 with L2.	4	149 906	150 875	0.05	0.98	0.98	7
8	Replace LM3 with L4.	7	149 247	150 216	0.05	0.98	0.98	8

When a principle is removed from the model, the latent trait θ_k representing that principle is no longer used in the model. For example, in model 4, the latent trait associated with LM1 is removed, and all items where the discrimination associated with LM1 was allowed to be nonzero were adjusted so the discriminations of L1 and DF2 were nonzero.

The initial model, model 1, implementing the model of the instrument in Table IV was fit and, then, the model transformation plan was carried out. All models had exceptional fit statistics with RMSEA $= 0.05$, CFI $=$ 0.98, and TLI $= 0.98$. Removing RS1 (reading a graph) did not improve model fit producing very strong changes, by Raftery's criteria [50], in AIC and BIC. As such, RS1 was retained in the model. Replacing LM1 (force in the direction of motion causes an object to speed up) and LM2 (force opposite the direction of motion causes an object to slow down) with L2 (Newton's 2nd law) and DF2 (definition of acceleration) also produced very strong increases in both AIC and BIC over the full model (model 1); these changes were not retained. Next, each lemma was examined individually. Replacing only LM1 with L2 and DF2 did improve model fit (strong improvement in AIC and BIC); however, replacing only LM2 with L2 and DF2 did not improve model fit. Lemma LM2 (applying a force opposite the direction of motion causes speed to decrease) seems to be separate in student thinking from Newton's 2nd law (L2) and the definition of acceleration (DF2). As such, LM2 was retained as a separate principle, while LM1 was not, to form model 4. Model 4, including LM1 but not LM2, was then transformed to remove L1 (Newton's 1st law) in two alternate ways. Model 6 replaced L1 with L2 (Newton's 2nd law) and DF2 (definition of acceleration) and was a significant improvement over model 4. Model 7 transformed model 4 by replacing L1 with L2 alone and produced improved AIC and BIC (both strong changes); these changes were stronger than the changes between model 4 and 6. As such, model 7 was retained. As such, students' reasoning did not require different abilities with Newton's 1st and 2nd law. Finally, model 7 was transformed by replacing LM3 (motion down an incline) with L4 (the force of gravity) which improved model fit substantially (both strong changes in AIC and BIC). The principles included in the optimal model, model 8, are shown in bold in Table IV.

Table VI shows the item-level MIRT parameters for the optimal constrained model (model 8). Following study 1 and 2, an overall discrimination a_0 was added to each item to capture a general ability with Newtonian mechanics. The "principles" column presents the discrimination of each principle on each FMCE item as well as its standard error. For example, the discrimination of FMCE item 1 on Newton's 2nd law (L2) is 6.21 ± 0.04 and on the definition of acceleration (DF2) is 0.20 ± 0.01. These discriminations represent the additional discrimination of the item on the principle above the item's discrimination of a general ability with Newtonian mechanics measured by a_0. The table also presents the overall difficulty d of each item. The standard error of each parameter was calculated by bootstrapping with 1000 replications using the "boot" [54] package in the R software system.

F. Comparison with the FCI

The optimal model for the FMCE (model 8) is strikingly different from the optimal model for the FCI presented in study 1 (Table III [6]). While the FCI required 19 principles for its description, the FMCE required only 8. The principles retained in the optimal models of both the FCI and the FMCE are shown in Table VII; the number of items using each principle is also presented. The FCI analysis retained only the first item in a problem block to correct for spurious correlations produced by blocking; this was not possible in the FMCE where all but one item is blocked. This was done under the assumption that the students generally address the items in an instrument in the order given, and as such, the first item in an item block would not be affected by other items in the block. The correlations produced by item blocking or "chaining" have been explored by authors [55]. Recently, clusters of incorrect answers in item blocks have been identified in the FCI where the second item is the correct answer if the incorrect answer to the first item had been the correct answer [56]. If the additional FCI items were retained, 4 additional principles would be required. As such, measured

TABLE VI. Optimal MIRT model (model 8). The number in parenthesis is the discrimination a_{jk} for the principle on the item. a_0 is the discrimination for a factor loaded on all items and d is the difficulty of the item. Values are presented as the mean ± the standard error of the mean. Items students find more challenging have smaller values of d. In general, well-functioning items will have a positive overall discrimination a_0. Items measuring the principles identified in the expert model should have positive principle discriminations a_{jk} that are substantially different from zero. Bolded item numbers represent the first item in an item block.

FMCE No.	Principles	a_0	d
1	L2(6.21 ± 0.04) DF2(0.20 ± 0.01)	8.54 ± 0.05	−3.49 ± 0.02
2	L2(1.79 ± 0.00)	3.53 ± 0.01	−1.68 ± 0.00
3	LM2(1.90 ± 0.01)	4.22 ± 0.01	−0.40 ± 0.00
4	L2(2.25 ± 0.01) DF2(0.03 ± 0.00)	3.56 ± 0.01	−1.65 ± 0.00
5	L2(0.52 ± 0.00)	1.74 ± 0.00	0.23 ± 0.00
6	L2(0.24 ± 0.00)	2.14 ± 0.00	−2.11 ± 0.00
7	LM2(1.67 ± 0.01)	3.66 ± 0.01	0.01 ± 0.00
8	L4(2.01 ± 0.01)	3.48 ± 0.01	−3.17 ± 0.01
9	L4(1.30 ± 0.00)	2.23 ± 0.00	−1.63 ± 0.00
10	L4(1.10 ± 0.00)	1.86 ± 0.00	−0.41 ± 0.00
11	L4(2.39 ± 0.01)	3.93 ± 0.01	−1.07 ± 0.00
12	L4(1.75 ± 0.00)	2.84 ± 0.00	−0.50 ± 0.00
13	L4(1.82 ± 0.01)	3.00 ± 0.01	0.51 ± 0.00
14	L2(2.04 ± 0.01) RS1(4.04 ± 0.02)	6.07 ± 0.03	−2.84 ± 0.01
15	L2(−0.27 ± 0.00) RS1(−0.02 ± 0.00)	0.78 ± 0.00	3.58 ± 0.00
16	L2(1.56 ± 0.01) DF2(−0.56 ± 0.01) RS1(2.62 ± 0.01)	4.51 ± 0.01	−1.49 ± 0.01
17	L2(1.38 ± 0.01) RS1(3.34 ± 0.01)	4.22 ± 0.01	−3.52 ± 0.01
18	LM2(0.19 ± 0.00) RS1(1.62 ± 0.00)	3.36 ± 0.01	−1.78 ± 0.00
19	L2(0.77 ± 0.00) DF2(−0.41 ± 0.00) RS1(1.38 ± 0.01)	2.83 ± 0.00	−1.72 ± 0.00
20	L2(0.26 ± 0.00) LM2(0.01 ± 0.00) DF2(−0.15 ± 0.00) RS1(0.41 ± 0.00)	1.61 ± 0.00	−0.90 ± 0.00
21	F8(0.13 ± 0.00)	2.35 ± 0.00	−0.88 ± 0.00
22	DF2(1.85 ± 0.01) RS1(1.28 ± 0.01)	4.15 ± 0.01	2.75 ± 0.01
23	DF2(1.36 ± 0.00) RS1(0.98 ± 0.00)	3.34 ± 0.01	0.95 ± 0.00
24	DF2(2.20 ± 0.01) RS1(1.67 ± 0.01)	4.43 ± 0.01	2.51 ± 0.01
25	DF2(0.86 ± 0.00) RS1(0.77 ± 0.00)	2.48 ± 0.00	0.37 ± 0.00
26	DF2(2.13 ± 0.01) RS1(1.50 ± 0.01)	4.33 ± 0.01	2.93 ± 0.01
27	L4(0.70 ± 0.00)	3.07 ± 0.00	0.71 ± 0.00
28	L4(0.64 ± 0.00)	2.36 ± 0.00	0.44 ± 0.00
29	L4(0.73 ± 0.00)	2.89 ± 0.00	1.16 ± 0.00
30	L3(2.91 ± 0.00)	1.86 ± 0.00	1.90 ± 0.00
31	L3(1.88 ± 0.00)	1.65 ± 0.00	1.79 ± 0.00
32	L3(5.56 ± 0.01)	4.43 ± 0.01	3.78 ± 0.01
33	L3(0.77 ± 0.00)	1.16 ± 0.00	4.07 ± 0.00
34	L3(6.15 ± 0.02)	5.11 ± 0.02	3.73 ± 0.01
35	L3(0.74 ± 0.00)	0.79 ± 0.00	0.77 ± 0.00
36	L3(2.42 ± 0.00)	1.56 ± 0.00	−1.81 ± 0.00
37	L3(0.37 ± 0.00)	0.97 ± 0.00	1.57 ± 0.00
38	L3(2.40 ± 0.00)	1.50 ± 0.00	−1.70 ± 0.00
39	L3(1.35 ± 0.00)	1.60 ± 0.00	1.86 ± 0.00
40	DF1(3.20 ± 0.02) RS1(0.16 ± 0.01)	3.74 ± 0.02	9.27 ± 0.05
41	DF1(0.46 ± 0.00) RS1(−0.01 ± 0.00)	1.18 ± 0.00	1.43 ± 0.00
42	DF1(1.40 ± 0.00) RS1(0.11 ± 0.00)	2.08 ± 0.00	3.56 ± 0.01
43	DF1(1.53 ± 0.00) RS1(0.13 ± 0.00)	1.47 ± 0.00	4.97 ± 0.01

by the number of principles required by experts to solve the instrument, the FCI has a much more thorough coverage of Newtonian mechanics. The difference in coverage between the two instruments was also noted by Thornton *et al.* [8]. The distribution of principles within the instruments is also very different. Of the 20 FCI items analyzed (keeping only the first item in a block), only 4 required a single principle. Of the 43 FMCE items analyzed, 25 required only one principle. An additional 9 items required similar combinations of principles. As such, the items in FCI are much more interconnected by principles they share than the items in the FMCE.

Table VII indicates that neither instrument directly used R2; this principle was retained because principle C4 is

TABLE VII. Comparison of the optimal model of Newtonian mechanics as tested by the FCI and the FMCE. The number of items in each instrument using the principle is also presented.

Label	No. FMCE items	No. FCI items	Principle
			Kinematics
DF1	4	3	Definition of velocity ($\vec{v} = d\vec{r}/dt$)
DF2	10	1	Definition of acceleration ($\vec{a} = d\vec{v}/dt$)
R1	···	4	Trajectory $\vec{a}$ = constant $[\vec{r}(t) = \vec{r}_0 + \vec{v}_0 t + \frac{1}{2}\vec{a}t^2]$
R2	···	···	Velocity $\vec{a}$ = constant $[\vec{v}(t) = \vec{v}_0 t + \vec{a}t]$
C1	···	1	Instantaneous velocity is tangent to the trajectory.
C2	···	2	Objects moving in a curved trajectory will experience centripetal acceleration.
C3	···	2	1D trajectory a = constant, $[x(t) = x_0 + v_0 t + \frac{1}{2}at^2]$
C4	···	1	1D velocity a = constant, $[v(t) = v_0 + at]$
			Dynamics
DF3	···	2	The net force is the vector sum of the forces (forces add as vectors).
L1	···	4	Newton's 1st law
L2	11	4	Newton's 2nd law
L3	10	3	Newton's 3rd law
LM2	4	···	A force opposite the direction of motion causes objects to slow down.
			Properties of forces
L4	9	11	Objects near Earth's surface experience a constant downward force or acceleration of gravity toward the center of Earth.
F1	···	1	An object in contact with a surface experiences a normal force.
F2	···	3	An object does not necessarily experience a force in the direction of motion.
F3	···	2	Air pressure does not exert a net downward force.
F4	···	1	The wind can exert a force on an object.
F5	···	2	Air resistance is negligible for a compact object moving a short distance.
F6	···	1	The force of gravity is approximately constant near Earth's surface.
F8	1	···	There is no contact force on an object after contact is lost.
			Other
RS1	16	···	Reading a graph.

derived from it. The FCI contains a lemma related to FMCE lemma LM2; however, the lemma was stated in terms of acceleration, not force. This lemma was not retained in the optimal model in the FCI. The table also indicates that the FMCE does not require Newton's 1st law; this resulted from Newton's 1st and 2nd law being combined in the optimal model for the FMCE. Newton's 1st and 2nd law were not combined in the optimal model for the FCI.

Figure 2 shows a visualization of principle structure of the optimal models for both the FCI (study 1) and the FMCE. The nodes represent principles, edges represent how many times two principles are used in the same item. For example, in the FMCE, five items use both L2 and DF2. The number near the curve connecting the same node represents the number of times the principle is the only principle for an item. Facts and reasoning steps were not included in the figure. For the full figure including facts see the Supplemental Material [57]. For the FMCE, the number with a star indicates the number of items using only the labeled principle with RS1 (reading a graph). The FCI includes 6 facts and the FMCE 1 fact; both instruments include one reasoning step. The reasoning step in the FCI is used in only one item and was not included in the optimal model; however, the reasoning step in the FMCE is used in a substantial subset of the items. The principles shared by both instruments were named consistently in both the current work and study 1; Table VII presents a list of the principles and their labels.

Figure 2 shows that most FCI items included multiple principles while most FMCE questions were designed to test one particular principle. Further, different FCI items employed different combinations of principles producing a connected network; the FMCE often either used a single principle or the same combination of principles producing a disconnected network. This implies, as the FCI authors intended [2], the FCI measures an integrated Newtonian force concept. The FMCE measures dimensions of this force concept, but these dimensions are far less integrated.

G. FMCE subscales

A fundamental challenge in applying the FCI to understand learning is its lack of a consistent subscale structure demonstrated by multiple studies identifying an inconsistent and often unintelligible factor structure [17,20,21]. The integrated network of principles shown in Fig. 2(a) serves to explain why consistent subscales of the instrument measuring identifiable dimensions of Newtonian reasoning have not been extracted. This means that, while the FCI measures an integrated force concept, it can provide little

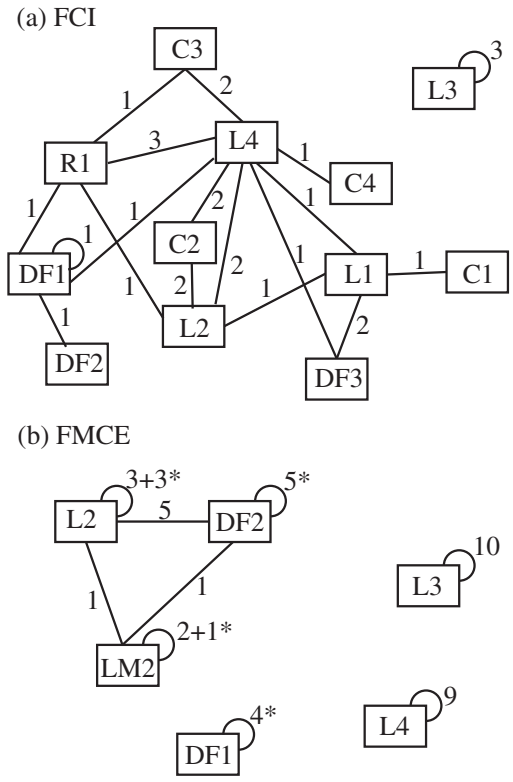


FIG. 2. Comparison of the network of principles of the (a) FCI and the (b) FMCE. The nodes represent principles, edges represent how many times two principles are used in the same item. The number near the curve connecting the same node represents the number of times the principle is the only principle for that item. For the FMCE, the number with a star represents the number of items using only the principle with reasoning step RS1 (reading a graph).

information about the details of that force concept. The FCI cannot tell you if a student's difficulty lies with Newton's laws or kinematics.

The repetition of single principles or the same combination of principles across many items in the FMCE suggests that the instrument should have identifiable subscales. This is supported by the 5-factor exploratory model where factors generally loaded most heavily within item blocks; the single-principle or repeated-principle items also generally were restricted to the same factor. Table VIII presents possible subscales first using the set of principles in the original model (model 1), then principle combinations suggested by the optimal constrained model (model 8). For subscales combining multiple principles, a name for the combined scale is suggested. The average and standard deviation for each set of items is presented. For example, the average of FMCE items 40–43, testing the definition of velocity (DF1), is 0.85 ± 0.2. Cronbach's alpha, α, which measures the internal consistency of the subscale, is also presented. Cronbach's alpha of 0.7 is considered adequate for low stakes testing, while 0.9 is required for high stakes testing [58]. All but one of the suggested subscales demonstrated adequate internal consistency. The last two rows combine principles as suggested by the optimal model and have excellent alphas. Alpha generally grows as the number of items increases and, therefore, the increased alpha might result from the larger number of items in the combined subscales.

The subscale structure suggests that, rather than measuring an integrated force concept, the FMCE might be more productively employed to measure the details of that force concept. For example, for the class studied additional instructional resources might productively be directed toward understanding motion under gravity.

The set of subscales with acceptable alpha values opens the possibility that these subscales could be used as a model for the instrument. A confirmatory factor analysis was performed with these five subscales producing fairly poor fit statistics (CFI = 0.80, TLI = 0.79, and RMSEA = 0.080). The optimal model 8 found by MIRT was a

TABLE VIII. FMCE subscales. Average score presented as mean ± standard deviation; Cronbach's α provides a measure of internal consistency.

Principle	FMCE No.	Average	α	Description
DF1	40–43	0.85 ± 0.2	0.66	Definition of velocity
DF2	22–26	0.61 ± 0.4	0.90	Definition of acceleration
L1	2, 5, 14, 15, 17	0.50 ± 0.3	0.78	Newton's 1st law
L3	30–39	0.63 ± 0.3	0.84	Newton's 3rd law
L4	11–13, 27–29	0.49 ± 0.4	0.88	Objects near Earth's surface experience an approximately constant force of gravity toward the center of Earth.
LM1	1, 4, 16, 19, 20	0.36 ± 0.4	0.88	A force in the direction of motion causes objects to speed up.
LM2	3, 7, 18, 20	0.40 ± 0.4	0.80	A force opposite the direction of motion causes objects to slow down.
LM3	8–10	0.32 ± 0.4	0.83	An object on an incline experiences a constant net force down and parallel to the incline.
L1/L2/LM1/LM2	1–7, 14–20	0.42 ± 0.3	0.93	Newton's 1st and 2nd law
L4/LM3	8–13, 27–29	0.44 ± 0.4	0.91	Motion under gravity

substantially superior model. Various additional combinations of subscales were also modeled using CFA including subscales which eliminated the problematic items flagged by Thornton *et al.*; none produced CFI or TLI about 0.85.

The L1/L2/LM1/LM2 subscale was identified ignoring RS1 which does not involve reasoning specific to physics. If RS1 were used to split the group, it would divide the items in this group into the force sled and force graph subgroups suggested by Thornton and Sokoloff [1].

IV. DISCUSSION

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

RQ1: What is the optimal model of the FMCE identified using exploratory factor analysis? To what extent does the blocking of items explain the factor structure? An EFA using MIRT showed that a 10-factor solution minimized AIC, a 9-factor model minimized BIC, and a 5-factor model had superior RMSEA, CFI, and TLI. For both the 5-factor and 10-factor model, most items in each item block had their highest factor loadings in the same factor. Unlike the FCI, this produced a set of factors containing items generally representing the same physical principle. While the 10-factor model separated more thoroughly into item blocks; the 5-factor model generally combined items identified as physically similar by the optimal constrained MIRT model in Table VI. It seems likely that the 10-factor solution was identified as optimal because it most closely matched the blocked structure of the instrument. It also seems likely that the 5-factor model had excellent fit statistics because it most closely captured the groups of items associated with the same principle or sets of principles. While blocking seems to also be causing correlations not supported by the general physical principles tested in the FMCE, it seems to be much less important than in the FCI.

The factor structure extracted was dramatically different than the only other reported factor structure of the instrument. Ramlo reported only 3 factors [11]. The 3-factor solution is curious. The blocked structure of the instrument strongly suggests the 10-factor solution (9 blocks and 1 unblocked item). The correlation matrix and confirmatory MIRT analysis strongly suggest 5 factors. It is difficult to construct a theoretical reason to support 3 factors; it seems quite likely that the sample size in Ramlo's study ($N = 146$) was insufficient to resolve the full factor structure.

In their introduction of the instrument, Thornton and Sokoloff [1] discuss subgroups of items such as the force sled items (items 1–7) or the coin toss items (items 11–13), but do not discuss the overall blocked structure of the instrument. Superficially, because the FMCE is divided into 9 blocks, where the items in each block measures a similar physical concept, and one unblocked item, the instrument seems to be designed to produce 10 factors. The 10-factor model minimized AIC in this study; however, its other fit statistics were fairly poor. Also, some of the factors in the 10-factor model failed to load strongly on any set of items. The 5-factor model, which generally combined items into subgroups based on common sets of principles identified by constrained MIRT, had excellent fit statistics. The EFA also provides support for the identification of 5, 6, 15, 33, 35, and 37 as potentially problematic; all had relatively low factor loadings in the 5-factor model. The factor loading of item 39 does not suggest it is problematic.

The correlation analysis partially supports the identification of problematic items that were removed from the modified scoring proposed by Thornton *et al.* [8]. In Fig. 1(a), items 15, 33, 35, and 37 are weakly correlated with their items blocks; however, items 5, 6, and 39 are well connected to their subgroups. In Fig. 1(b), item 15 is only correlated with item 33 which is not strongly correlated to the other Newton's 3rd law items. Item 37 is negatively correlated with all items that it connects to in the diagram. Again, items 5, 6, and 39 do not appear problematic. This pattern of correlations was unexpected based both upon the blocks containing the items and the theoretical model of the solution of the items. These correlations are consistent with Thornton's *et al.* [8] observation that these items are often answered correctly by students who do not have an understanding of Newtonian mechanics, unlike items in the instrument not identified as problematic.

RQ2: What is the optimal model of the FMCE using constrained MIRT? The principles retained in the optimal constrained MIRT model are bolded in Table IV. As in both study 1 [6] and study 2, only some of the lemmas found in the expert solutions were included in the optimal model. Unlike in study 1, the optimal model of the FMCE combined Newton's 1st and 2nd laws. While this may be an artifact of instruction at the two institutions, it could also be the result of Newton's 1st and 2nd law items appearing in the same item blocks.

Table VI shows the difficulty and discrimination parameters of the optimal MIRT model. The number in parenthesis next to the principle label is the discrimination for the principle. The items show a broad range for this parameter. Many items measure only a single principle. Section III G describes how these items can be formed into subscales. Some items requiring multiple principles only discriminate strongly on one of the principles; these items should be good measures of the strongly discriminating principle. For example, items 1 and 4 require two principles but discriminate on one principle more strongly than the other; items 1 and 4 should be good measures of Newton's 2nd law (L2). Items 22 through 26 require multiple principles with commensurate and large discriminations. These items measure multiple principles at the same time, but cannot differentiate between the principles. Item 20 had small discrimination values for all principles and, therefore, does not contribute additional information about these principles. Item 15 had negative discrimination; this may indicate

the item is problematic. Item 15 was one of the items Thornton *et al.* removed in their revised scoring [8].

The optimal model can also be used to understand previous research into the FMCE. Many of the items removed by Thornton *et al.* [8] had relatively weak principle discrimination 5, 6, 33, 35, and 37; as above, item 15 had negative discriminations. The clusters identified by Thornton *et al.* 8_10, 11_13, and 27_29 all involve a single principle L4 (the law of gravitation) supporting scoring them as a single item. The optimal model also supports Smith and Wittmann's proposal to combine items 8–10, 11–13, and 27–29 to form a single subgroup [9]. The optimal model also strongly supports Smith and Wittmann's identification of items 40–43 as a subgroup. The analysis which retains RS1 (reading a graph) also supports both Thornton and Sokoloff [1] and Smith and Wittmann's [9] division of the Newton's 1st and 2nd law items into distinct subgroups where one involves reading graphs and the other does not.

RQ3: How do the number of principles and connectivity of the principles in the optimal MIRT models of the FMCE and the FCI compare? Examining either the initial or optimal models in the present study (Table IV and Table VI), the corresponding models in study 1 (Table III and VII [6]), or the summary provided in Table VII show that the two instruments have substantially different coverage. The initial expert model of the FCI contained 34 principles while the optimal model required 19 principles. These 19 principles were distributed over 20 items. Study 1 analyzed only a subset of the items in the FCI to remove the effect of item blocking. The initial expert model of the FMCE contained only 11 principles while the optimal model contained 8 principles. These 8 principles were spread over 43 items.

Figure 2 shows the sharp contrast between the principle networks of the FCI and the FMCE. The FCI's network is generally connected while the FMCE network in generally disconnected. The correlation and partial correlation matrices of the two instruments are also dramatically different. These observations are related to the very different use of principles in the two instruments. Most FCI items use multiple principles and very few combinations of principles are repeated. This leads to the generally connected FCI network in Fig. 2(a) and serves to explain the sparsely connected correlation and partial correlation matrices reported in study 1 (Figs. 1 and 2 [6]). The majority of the items in the FMCE use a single principle; many other items repeat combinations of principles leading to the generally disconnected FMCE network in Fig. 2(b). This practice serves to explain the strong connections in the correlation matrix and the islands of connected items in the partial correlation matrix of the FMCE. Of the 20 FCI items analyzed in study 1, only 4 use only a single principle, 20%. Of the 43 FMCE items analyzed in the current study, 25 use only a single principle, 58%. If RS1 is removed, an additional 13 items use only one principle, or 38 of the 43 items, 88%.

The interconnected nature of the FCI serves to explain the failure of EFA to extract a factor structure that combined items which theoretically tested the same underlying concept [17–19]. There simply are not substantial groups of items in the FCI that test the same underlying concept or concepts. The only repeated set of single-principle items in the FCI, which measure Newton's 3rd law, do consistently factor together in most exploratory studies [6,20,21]. Because the FMCE has many groups of items either testing the same principle or repeated groups of principles, it was possible to extract a subscale structure. Table VIII shows possible subscales either extracted from the initial expert model or from the optimal model. Except for the subscale representing DF1 (the definition of velocity), all have acceptable internal consistency for low stakes testing [58]. The items in the DF1 subscale were proposed as an additional subgroup (40–43) by Smith and Wittmann [9]. The low internal consistency of this subgroup suggests that multiple other combinations might provide more reliable measures of the facets of student knowledge of mechanics.

V. IMPLICATIONS

This work showed a sharp contrast between the coverage and connectedness of the FMCE and the FCI. While the FCI measures an integrated force concept, the FMCE uses the repetition of single principles or the same combination of principles across many items to repeatedly measure facets of that force concept. This allowed the identification of subscales within the FMCE. Only a fraction of the principles used in the FCI are represented in these subscales. This suggests that, rather than providing symmetric information about the understanding of Newtonian mechanics, the FCI and the FMCE provide complementary information. The FCI measures an overall Newtonian force concept with stronger coverage than the FMCE; the FMCE measures subdimensions of this force concept. The FCI can provide instructors a broad measure of the overall conceptual understanding of their students; the FMCE can allow instructors to identify individual concepts where students need improvement. This partially alleviates one of the primary weakness of the FCI, the lack of a well-defined subscale structure. While the subscale internal consistencies were adequate, the low model fit of the CFA suggests the instrument requires further refinement to have a well-defined subscale structure.

Recent research by Traxler *et al.* [25] provided compelling evidence that some items within the FCI are unfair to either men or women. They proposed a 19-item version of the instrument to eliminate validity and fairness problems; instructors interested in using both the FCI and FMCE should use this reduced instrument. Henderson *et al.* repeated the analysis for the FMCE and found few unfair items [16].

Beyond the specific results reported, this study as well as studies 1 and 2 demonstrated the additional insights that

can be derived from evaluations of physical knowledge using the fine-grained, expert-derived models presented in these three works.

VI. LIMITATIONS

This work was performed with a single sample drawn from a single institution. Additional samples should be tested to determine if the conclusions are general. The theoretical model presented in Table IV was constructed from a sample of expert solutions at a single institution. Other models should be constructed and explored. Any researcher who wishes to explore an alternate model may request the data used in this study.

VII. FUTURE WORK

The constrained MIRT method will be applied again to both the FCI and the FMCE disaggregating the samples by gender to attempt to understand the gender differences identified in the FCI by Traxler *et al.* [25] and the smaller differences identified in the FMCE by Henderson *et al.* [16]. Models will also be extended to include common misconceptions to explore the competition of expert and naive reasoning.

VIII. CONCLUSIONS

This work examined the structure of the Force and Motion Conceptual Evaluation with Multidimensional Item Response Theory. Exploratory analysis identified 5-factor, 9-factor, and 10-factor solutions as optimal on some fit statistics. The loadings of both the 5-factor and 10-factor solution were generally consistent with the blocked structure of the instrument. Problems identified in the 10-factor model, as well as the superior fit statistics of the 5-factor model, suggest the 5-factor model as the best exploratory model. A confirmatory analysis using MIRT constrained to a theoretical model was also employed to determine the optimal model of the FMCE. The optimal model contained only 8 principles of mechanics compared to 19 principles in the optimal model for the FCI; the FCI has much broader coverage of mechanics than the FMCE. The distribution of principles in the two instruments was also dramatically different. Very few items in the FCI required only a single principle for their solution while the majority of FMCE items could be solved with a single principle. This repetition of single principles and combinations of principles allowed the identification of subscales within the FMCE; most subscales identified had Cronbach's alpha of at least the 0.7 required for low stakes testing.

ACKNOWLEDGMENTS

Data collection for this work was supported by National Science Foundation Grants No. EPS-1003907 and No. ECR-1561517.

[1] 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).
[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] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. Phys. Educ. Res. **10**, 020119 (2014).
[5] 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).
[6] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010137 (2018).
[7] C. Zabriskie and J. Stewart, Multidimensional item response theory and the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **15**, 020107 (2019).
[8] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **5**, 010105 (2009).
[9] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **4**, 020101 (2008).
[10] T. I. Smith, M. C. Wittmann, and T. Carter, Applying model analysis to a resource-based analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **10**, 020102 (2014).
[11] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, Am. J. Phys. **76**, 882 (2008).
[12] R. M. Talbot, Taking an item-level approach to measuring change with the Force and Motion Conceptual Evaluation: An application of item response theory, School Sci. Math. **113**, 356 (2013).
[13] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, Mason, OH, 1986).

[14] T. I. Smith, K. A. Gray, K. J. Louis, B. J. Ricci, and N. J. Wright, Showing the dynamics of student thinking as measured by the FMCE, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by L. Ding, A. Traxler, and Y. Cao (AIP, New York, 2018) p. 380.
[15] K. J. Louis, B. J. Ricci, and T. I. Smith, Determining a hierarchy of correctness through student transitions on the FMCE, in *Proceedings of the 2018 Physics Education Research Conference Proceedings, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2019).
[16] 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).
[17] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure? Phys. Teach. **33**, 138 (1995).
[18] 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).
[19] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[20] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. Phys. Educ. Res. **8**, 020105 (2012).
[21] 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).
[22] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[23] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **6**, 010103 (2010).
[24] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).
[25] 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).
[26] 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 *2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).
[27] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. Phys. Educ. Res. **11**, 020134 (2015).
[28] A. Newell and H. A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, NJ, 1972).
[29] 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).
[30] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, Science **208**, 1335 (1980).
[31] J. H. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Models of competence in solving physics problems, Cogn. Sci. **4**, 317 (1980).
[32] F. Reif and J. I. Heller, Knowledge structure and problem solving in physics, Educ. Psychol. **17**, 102 (1982).
[33] J. Clement, Students' preconceptions in introductory mechanics, Am. J. Phys. **50**, 66 (1982).
[34] L. C. McDermott, Research on conceptual understanding in mechanics, Phys. Today **37**, 24 (1984).
[35] G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, Accommodation of a scientific conception: Toward a theory of conceptual change, Sci. Educ. **66**, 211 (1982).
[36] A. A. DiSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[37] A. A. Disessa and B. L. Sherin, What changes in conceptual change?, Int. J. Sci. Educ. **20**, 1155 (1998).
[38] D. Hammer, Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions, J. Learn. Sci. **5**, 97 (1996).
[39] M. T. H. Chi and J. D. Slotta, The ontological coherence of intuitive physics, Cognit. Instr. **10**, 249 (1993).
[40] 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).
[41] 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).
[42] D. Hammer, More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research, Am. J. Phys. **64**, 1316 (1996).
[43] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).
[44] S. DeVore, J. Stewart, and G. Stewart, Examining the effects of testwiseness in conceptual physics evaluations, Phys. Rev. Phys. Educ. Res. **12**, 020138 (2016).
[45] Physport, https://www.physport.org. Accessed 8/8/2017.
[46] US News & World Report: Education, US News and World Report, Washington, DC, https://premium.usnews.com/best-colleges. Accessed 4/30/2017.
[47] 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.
[48] 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).
[49] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach* (Springer-Verlag, New York, NY, 2003).
[50] A. E. Raftery, Bayesian model selection in social research, Sociol. Methodol. **25**, 111 (1995).
[51] L. J. Cronbach and P. E. Meehl, Construct validity in psychological tests., Psychol. Bull. **52**, 281 (1955).
[52] L. A. Clark and D. Watson, Constructing validity: Basic issues in objective scale development., Psychol. Assess. **7**, 309 (1995).
[53] 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).

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

[55] W. M. Yen, Scaling performance assessments: Strategies for managing local item dependence, J. Educ. Measure. **30**, 187 (1993).

[56] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler, Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis, Phys. Rev. Phys. Educ. Res. **15**, 020122 (2019).

[57] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.020141 for a comparison of the FCI and FMCE including the facts identified in each instrument.

[58] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).