Multidimensional item response theory and the Force Concept Inventory

John Stewart,* Cabot Zabriskie, Seth DeVore, and Gay Stewart
Department of Physics and Astronomy, West Virginia University, Morgantown, West Virginia 26506, USA

(Received 8 March 2018; published 13 June 2018)

Research on the test structure of the Force Concept Inventory (FCI) has largely been performed with exploratory methods such as factor analysis and cluster analysis. Multidimensional Item Response Theory (MIRT) provides an alternative to traditional exploratory factor analysis which allows statistical testing to identify the optimal number of factors. Application of MIRT to a sample of $N = 4716$ FCI post-tests identified a 9-factor solution as optimal. Additional analysis showed that a substantial part of the identified factor structure resulted from the practice of using problem blocks and from pairs of similar questions. Applying MIRT to a reduced set of FCI items removing blocked items and repeated items produced a 6-factor solution; however, the factors still had little relation the general structure of Newtonian mechanics. A theoretical model of the FCI was constructed from expert solutions and fit to the FCI by constraining the MIRT parameter matrix to the theoretical model. Variations on the theoretical model were then explored to identify an optimal model. The optimal model supported the differentiation of Newton's 1st and 2nd law; of one-dimensional and three-dimensional kinematics; and of the principle of the addition of forces from Newton's 2nd law. The model suggested by the authors of the FCI was also fit; the optimal MIRT model was statistically superior.

DOI: 10.1103/PhysRevPhysEducRes.14.010137

I. INTRODUCTION

The Force Concept Inventory (FCI) was introduced 25 years ago and has become one of the most used and most studied instruments in physics education research (PER) [1]. Measurements using the instrument have been important in the recognition that traditional instruction was not sufficient for students to develop a conceptual understanding of Newton's laws [2]. Its success was followed by the development of numerous other conceptual instruments some of which found wide-spread use including the Force and Motion Conceptual Evaluation [3], the Conceptual Survey of Electricity and Magnetism [4], and the Brief Electricity and Magnetism Assessment [5]. These four instruments have in turn been used to help understand the effect of pedagogical innovations, the challenges of learning physics, and issues of inclusion in physics. The impact of these instruments has been immense; they have been used in a substantial subset of the studies done in PER. For a broad overview of PER including the role of conceptual inventories in PER, see Docktor and Mestre's recent synthesis [6].

A substantial number of studies have attempted to understand the overall structure of the FCI. These have included purely exploratory or descriptive methods such as factor analysis [7–9], module analysis [10], cluster analysis [11,12], item response theory [13–16], and item response curves [17,18]. The structure of student reasoning on the FCI has also been investigated by methods such as model analysis that require the input of a partial model of the concepts measured by the FCI [19]. Model analysis was later shown to be exact only in certain limiting cases [12]. For a summary of these exploratory and nonexploratory methods, see the review by Ding and Beichner [20].

The reliability and validity of the FCI have also been tested. The internal consistency of the FCI measured by Cronbach's alpha is quite strong [16,21]. The instrument has also demonstrated good test-retest reliability [22]. While some validity issues have been identified [16], these are minor compared to those reported for some other instruments [23].

The current study explored the factor structure of the FCI using Multidimensional Item Response Theory (MIRT). This method has previously been applied to the FCI [24] by Scott and Schumayer. MIRT, described in detail in Sec. II, provides statistical criteria for determining the optimal number of factors unlike traditional exploratory factor analysis (EFA). The current study applied MIRT to the FCI using a larger data set than the previous study collected under conditions where correct answering was more strongly incentivized, thus allowing a finer resolution of

*jcstewart1@mail.wvu.edu

the details of student thinking. MIRT also allows models to be constrained to eliminate factor loadings that should not theoretically occur. As such, it allows a more detailed exploration of the structure of an instrument than traditional EFA.

A. Factor analysis and the FCI

The authors of the FCI provided a detailed description of the physical concepts each item in the original instrument was designed to measure [1]. Soon after its publication, attempts to extract the suggested structure with EFA were unsuccessful leading to debate about what the instrument actually measured [7,25,26]. Huffman and Heller reported that, for a sample of 145 high school students, principle component analysis identified two factors: Newton's 3rd law and kinds of forces. For 750 university students, only one factor was identified: kinds of forces. This study selected the number of factors by requiring that each new factor explain at least 5%–10% additional variance. The difference in the number of factors identified between the Huffman and Heller study and other studies of the FCI may have resulted from the use of different criteria to identify the optimal number of factors. Methods to identify the optimal number of factors are discussed in Sec. II E.

Scott, Schumayer, and Gray applied EFA to the FCI post-test scores of a sample of 2150 students in a college algebra-based physics course [8]. The FCI was delivered electronically and students were given no special incentive for completion. They found a single factor explained a substantial portion of the variance, but concluded a five-factor model was optimal. Parallel analysis was used to select the optimal number of factors. The "knee" of their Scree plot suggested that two or three factors could also be considered optimal. In examining the loadings on the single factor, they discuss the possibility of very difficult items not being strongly correlated with the single-factor solution. The variance explained by the addition of each new factor is not reported, and, therefore, the number of factors selected cannot be compared with Huffman and Heller's solution [7].

Semak *et al.* explored the evolution of the structure of student thinking on the FCI using factor analysis [9] for 427 algebra- and calculus-based introductory physics students. They found the optimal solution had 5 factors on the pretest and 6 factors on the post-test. Parallel analysis was used to select the optimal number of factors; however, examination of the Scree plots from their study suggests one could have also identified one or two factors as optimal for both the pretest and post-test. This would have provided support for Huffman and Heller's model. We provide a comparison of the four reported factor structures in Sec. V.

Factor analysis has also been used to investigate other sets of physics problems. Ramlo [27] calculated the factor structure of the FMCE [3] finding 3 factors for the pretest; however, these factors contained a mixture of concepts and Ramlo concluded the pretest factor structure was undefined. Three factors were also found for the post-test with items covering similar conceptual topics largely loading onto the same factor. Ramlo used a Scree plot to identify an eigenvalue cutoff of 2.5 to determine the optimal number of factors.

B. Item Response Theory and the FCI

Item Response Theory (IRT) contains a broad set of statistical models which calculate the probability of a student with some overall proficiency or ability to answer individual items on a test correctly. Many different IRT models have been used to investigate the FCI including the Rasch model, the 2-parameter logistic (2PL) model, the 3PL model, and MIRT. These models are reviewed in Sec. II.

Many studies have investigated the FCI with IRT using a single ability parameter (unidimensional IRT). Wang and Bao employed the 3PL IRT model to investigate the FCI pretest for 2802 college students taking calculus-based physics [13]. They reported excellent model fit with all items showing reasonable difficulty parameters and no items with negative discrimination parameters. The 3PL model adds a parameter to the 2PL model to account for random guessing. The majority of the guessing parameters were less than the 20% random guessing would produce. The use of the 3PL model for distractor-driven instruments has been questioned [18].

Planinic, Ivanjek, and Susac performed a Rasch analysis of 1676 Croatian high school students who had completed an algebra-based physics class [14]. The Rasch model difficulty parameters were largely in agreement with the overall item average. This study is difficult to generalize because the overall score on the instrument (27.7%) was so low and the measurement was performed two and one-half years after instruction.

Osborn Popp, Meltzer, and Megowan-Romanowicz also used Rasch model IRT for a sample of 4775 high school students to investigate item fairness; all students had been taught using Modeling Instruction [15]. IRT using the Rasch model was used to determine if items within the FCI were of equal difficulty for men and women. They found that a number of items were significantly easier for male students and some for female students.

Traxler *et al.* [16] also investigated item fairness in the FCI with IRT using the 2PL model. They found that eight items were substantially biased toward men and and two toward women; they proposed a reduced 19-item instrument to eliminate all biased and poorly functioning items.

Han *et al.* used the 3PL IRT model as part of the process of evaluating the equivalence of two shorter versions of the FCI [28]. Traxler *et al.* [16] cautions that the gender unfair items were not evenly distributed between the shortened tests, and, therefore, the two shorter tests might have different performance results for men and women.

Scott and Schumayer [24] attempted to replicate the work of Scott, Schumayer, and Gray [8] on a related data

set using MIRT. They confirmed the 5-factor solution. Comparing the factor models of the two studies showed very good, but not perfect, agreement suggesting MIRT and EFA are complementary techniques. To select the optimal number of factors, AIC and BIC (described in Sec. II) were minimized.

IRT has also been used to explore other sets of physics problems. Lee *et al.* [29] used 2PL IRT to examine how the skill of physics students using an online homework system changed between their first and second attempts at a problem. Whether feedback was given on the first attempt and the type of feedback strongly influenced the change in student skill (IRT ability) between the first and second attempt.

Morris *et al.* [17] introduced an alternative to IRT (bearing a very similar name), item response curves (IRC), which was used to analyze the FCI. IRC analysis simplifies IRT analysis by using the overall test score as a surrogate for student ability, greatly reducing computational demands and allowing intuitive exploration of the effect of distractors. Using a sample of over 4500 students drawn from multiple institutions, a later study by Morris *et al.* [18] compared IRC analysis to the IRT analysis of Wang and Bao [13] and found excellent correlation between the difficulty parameters of the models.

C. The structure of knowledge

Most explorations of the structure of the FCI have focused on determining a general structure that represents the entire instrument in terms of a small number of factors or clusters. This reductionism is at odds with a large body of research suggesting students' knowledge of physics is complex and that students (novices) do not possess the strongly integrated view of physics of expert practitioners. Experts and novices categorize problems differently; novices by surface features and experts by deeper conceptual divisions [30,31]. One commonly accepted difference in the knowledge structure of experts is the hierarchical nature of the structure, with the most fundamental principles at the top and less fundamental concepts branching out from there [32–34]. This more deliberate structuring of knowledge allows experts to engage more efficiently in chunking of knowledge [35–37] for more expedient application of the correct physics principles when engaging in problem solving.

Conversely, novices lack this deliberate knowledge structure leading to less deliberate methods of problem solving. This review will follow the categorization of expert-novice research presented in Docktor and Mestre's extensive synthesis of PER [6]. One view regarding some of the novel ways that novices approach problems differently from experts is the "misconceptions" view. This view argues that students, through their life experiences, have developed theories regarding how the world works and that using these, often incorrect, theories leads to some of the common difficulties in physics problem solving [38–40]. Research into these misconceptions has shown that they are very difficult to overcome due in part to the time students have spent believing them to be true [41,42]. Another method of explaining the differences is the "ontological categories" view, which posits that students miscategorize their knowledge, storing it in incorrect broad categories (i.e., thinking of force as a thing that can be used up) [43–45]. Another popular theoretical framework is the "knowledge in pieces" view [46–48] wherein student understanding consists of a number of granular facts that are activated, either individually or in small groups, to synthesize a solution. Regardless of the theoretical framework used to describe it, novice knowledge and the associated problem solving techniques have been shown to be highly sensitive to the context of the problem and how it relates to problems they have seen in the past [49–51]. As such, the knowledge state of students may be better described by models of a granular knowledge structure instead of the integrated models implied by factor analysis or cluster analysis.

The current work will produce a fine-grained model of the information needed to solve FCI problems. This model is very similar to models produced by a paradigm of cognitive research into complex problem solving pioneered by Simon and Newell [52]. This paradigm and its history, which dominated research into problem solving for over 30 years, were summarized by Ohlsson [53]. The paradigm constructed computational models that replicated the problem solving sequence of human solvers; the sequence of the human solver was identified by coding extensive think-aloud transcripts. This method was applied to examine expert-novice differences in problem solving in kinematics and dynamics, as well as other fields [54,55]. Reif and Heller offered a related detailed model of problem solving in mechanics [35]; this model did not meet the test of being computationally functional, but was meant to be a complete model that could serve as a prescription of expert behavior. The model we will propose for the FCI shares many features with the computational models of Larkin *et al.* [55] and the model of Reif and Heller [35]. The work on complex problem solving focused primarily on quantitative solutions; however, the framework presented by Reif and Heller acknowledged the role of qualitative decisions in the solution process and suggested extensions to model qualitative reasoning.

D. Research questions

This study seeks to answer the following research questions.

- RQ1: What factor structure is extracted for the FCI by MIRT? Is this structure consistent with the results of other factor analysis?
- RQ2: Can parts of this factor structure be explained by factors other than the structure of student knowledge of Newtonian mechanics?
- RQ3: If blocked items and repeated reasoning groups are removed, is the resulting factor structure consistent with Newtonian mechanics?

RQ4: Can theoretically constrained MIRT produce a model of the physical constructs measured by the FCI? If so, what is the optimal model of the FCI for this student population?

RQ5: Does the structure proposed by the FCI's authors provide a superior description of the instrument to the optimal model identified by MIRT?

This work leaves two important areas of analysis for future research: the role of misconceptions and bias. The FCI was constructed so that the distractors represented common misconceptions. In the analysis in this paper, only the correctness of the responses was analyzed. MIRT could be extended to include factors representing common misconceptions to determine how the models presented in this work would be modified.

There is a substantial body of research indicating that some problems within the FCI are unfair to women, with a few unfair to men. These problems have often factored together in previous analysis [8,9] leading to the possibility that some factors are identified because of biases in the problems. Many biased problems were removed in the analysis in this study to remove spurious correlations; however, future research should investigate whether the factor structure identified is independent of gender. While this study will not focus on gender fairness, the reduced fair 19-item FCI proposed by Traxler *et al.* [16] will be examined using the optimal theoretical FCI model identified by MIRT. For a review of research into FCI item bias, see Traxler *et al.* [16]. For a review of the issue of gender disparities in conceptual inventories see Madsen, McKagan, and Sayre [56] or Henderson *et al.* [57]. For a general review of gender in physics see Traxler *et al.* [58].

II. METHODS

A. Force Concept Inventory

The FCI is a 30-item multiple-choice instrument that includes conceptual questions about Newton's laws, kinematics, and forces [1]. Each item has five possible responses. The incorrect responses were developed to include common misconceptions. The FCI contains some individual items and some items that are grouped into blocks which share a common stem. The FCI was revised after its introduction; this work will use the revised FCI published with Mazur's *Peer Instruction* [59] and available at PhysPort [60].

B. Sample

The sample of FCI post-test results was collected from a large, southern land-grant university with an enrollment of approximately 25 000 students. This university held a Carnegie Classification of "Highest Research Activity" for the period studied. The sample is comprised of 4716 complete post-test responses collected from the Spring 2002 semester to Fall 2012 semester (23.1% women). The demographics of the university in 2012 were 79% White, 5% African American, 6% Hispanic, and 3% or less of other groups. The 25th to 75th percentile range of the general student population's ACT scores was 23–29 [61]. This sample was also used in the analysis of Traxler *et al.* [16].

The sample was collected in the introductory calculus-based mechanics course serving future physical scientists and engineers. Students in the course were required to attend two 50-minute lectures and two 2-hour laboratories each week. The lectures were largely traditional with attendance monitored by a quiz given at the beginning and end of each session. The lab sessions featured a mixture of activities including teaching assistant (TA) led interactive demonstrations, small group problem solving, inquiry-based hands-on activities, and traditional experiments. The class had been revised previous to the beginning of data collection and was presented with few changes over the period studied. The class was managed by the same lead instructor for the period studied; this instructor taught 75% of the lecture sections and oversaw the instruction of the remaining sections.

C. Item Response Theory

Many IRT probability models have been constructed to model student responses to different test structures and testing situations [62]. One of the most intuitive and widely used is the two-parameter logistic (2PL) model. The 2PL model uses unidimensional IRT, which explicitly models the effect of the single latent trait of ability, θ_i, on the probability of a student, i, successfully answering a question. The 2PL model assumes that each item j has a difficulty b_j and discrimination a_j. The probability, π_{ij}, that student i will successfully answer item j is given by the logistic function

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

The 2PL model can be expanded to the 3PL model by including an additional parameter for each item to model random guessing. The 3PL model has been used in some studies of the FCI. A simplification of the 2PL model, called the Rasch model, has also been used to study the FCI; the Rasch model sets the discrimination of each item to 1, $a_j = 1$.

The assumption of unidimensional IRT, that a single ability parameter captures the students' facility with the test material, may be correct for some instruments but it seems unlikely for the FCI, which measures a number of different facets of Newton's laws and kinematics. Multidimensional IRT (MIRT) extends unidimensional IRT by estimating multiple ability traits for each student. Mathematically, the student ability θ_i, which is a scalar in unidimensional IRT becomes a vector, $\boldsymbol{\theta}_i$, in MIRT. If k ability traits are estimated for each item, each trait is associated with its

own item discrimination a_{jk}, making the discrimination a vector, $\boldsymbol{a}_j$.

Multiple MIRT models exist; the most common MIRT model is called the *compensatory* model where the probability of a particular response is determined by a linear combination of $\boldsymbol{\theta}_i$ components where large components of $\boldsymbol{\theta}_i$ will compensate for the smaller components of $\boldsymbol{\theta}_i$. This 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 would be the product $-a_j b_j$ in the 2PL model and the product $\boldsymbol{a}_j \cdot \boldsymbol{\theta}_i$ is a dot product of two vectors. The parameter d_j is related to the difficulty of the item. While this MIRT model estimates multiple discrimination parameters for each item, it estimates only one d_j parameter. This is not optimal; it would be beneficial to know the difficulty of the item by individual trait. Noncompensatory MIRT would extract the difficulty of each item; however, this doubles the number of parameters estimated. We attempted to apply noncompensatory MIRT to the FCI but the models did not converge.

D. Model fit statistics

Unlike traditional factor analysis which identifies factors as eigenvectors of the correlation matrix, IRT introduces a statistical model, which is then fit to the observations. The model is used to calculate the likelihood function L, which represents the probability the observation occurred given the probability model. Maximum likelihood estimation techniques are used to search the parameter space to select a set of parameters which maximize L, the set of parameters which make the observed results most likely. This form of estimation can be used for a wide set of models and a number of general model fit statistics have been developed. We will report the Akaike information criterion (AIC), $\mathrm{AIC} = 2k - 2\ln(L)$, and the Bayesian information criterion (BIC), $\mathrm{BIC} = k\ln(n) - 2\ln(L)$, where n is the sample size and k the number of parameters estimated. The optimal model minimizes both quantities. AIC estimates the relative information lost when using a model to approximate the "true" model for a random sample drawn from the same population as the sample to which the model was fit, the out-of-sample information loss [63,64]. The second term in AIC is the in-sample information loss; the first term corrects for overfitting as the number of parameters increases. BIC has a similar interpretation but more strongly penalizes the addition of parameters. Smaller AIC or BIC represent less lost information.

Effect size standards for differences in AIC and BIC have not been formalized, but the magnitude of differences can be understood through the relation with L. A difference between the AIC or BIC of two models 1 and 2 with the same n and k is $\Delta_{12} = \mathrm{AIC}_1 - \mathrm{AIC}_2 = \mathrm{BIC}_1 - \mathrm{BIC}_2 = -2[\ln(L_1) - \ln(L_2)] = -2\ln(L_1/L_2)$. As such, the likelihood ratio between the two models follows $L_1/L_2 = e^{-\Delta_{12}/2}$. A two point decrease in either AIC or BIC then results in a model that is e times more likely. According to Burnham and Anderson, if the difference in AIC between two models is greater than 2, then there is little evidence that the two models are the same and, therefore, the one with lower AIC should be selected [63]. BIC follows a similar rule to AIC with a difference of 2 or more between the BIC of two models indicating a significant difference between them [65]. Raftery further classified differences in BIC as $\Delta_{12} \leq 2$ as "weak," $2 < \Delta_{12} \leq 6$ as "positive," $6 < \Delta_{12} \leq 10$ as "strong," and $\Delta_{12} > 10$ as "very strong" [65]. Because of the similarity of the statistics, we will also adopt these conventions for AIC. A very strong change of 10 in AIC or BIC results in a model that is $e^5 = 148$ times more probable.

A substantial number of additional fit statistics have been developed for maximum likelihood models. We report the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI). Hu and Bentler suggest using multiple indices to evaluate model fit [66]. Acceptable model fit is characterized by RMSEA < 0.05 and CFI > 0.96 or TLI > 0.96. For a summary of fit statistics, see Eaton and Willoughby [67].

Nested models were compared using a likelihood ratio test. If two models with likelihood functions L_1 and L_2 differ by k parameters where model 1 has fewer parameters, then the test statistic $\chi^2 = -2\ln(L_1/L_2)$ has a chi-squared distribution with k degrees of freedom and can be used to test whether the models are significantly different [68].

E. Additional analyses

While MIRT allows statistical selection of the optimal number of factors, traditional EFA uses a number of nonstatistical criteria. Factor selection often begins by an examination of the "Scree" plot which plots the eigenvalue of the correlation or covariance matrix corresponding to the factor against the factor number; the eigenvalue is related to the variance explained by the factor. An example of a Scree plot is shown in the Supplemental Material [69]. One then identifies the "knee" of the Scree plot, the point of greatest curvature. The number of factors corresponding to the knee is the optimal number of factors. For factor numbers greater than the location of the knee, each additional factor explains substantially less variance than the factors already extracted.

Many additional methods have been developed and often yield contradictory results. The sum of the eigenvalues of the correlation or covariance matrix is equal to the trace of the matrix; therefore, an eigenvalue that is less than the mean correlation or covariance represents a factor that explains less variance than an individual item. The optimal number of factors can then be extracted as the last factor

with eigenvalue above the mean. Parallel analysis computes the eigenvalues of a random correlation matrix; the optimal factor number is the last factor with eigenvalue greater than the parallel analysis eigenvalue.

Partial correlation matrices will be reported. The partial correlation matrix for the dichotomous scores on individual FCI items was calculated by regressing the total FCI score on the individual item score using a general linear model. The correlations of the residuals of these regressions form the partial correlation matrix.

The mean and standard deviation of MIRT parameters, $\boldsymbol{a}_j$ and d_j, were calculated by bootstrapping. Bootstrapping is a statistical technique that allows the calculation of the average, standard deviation, and confidence interval for a data set without assuming a statistical model. This is done by forming sub-samples of the data with replacement and recalculating the desired parameter for each sub-sample. For this work, 200 subsamples were used; this calculation required one week of computational time on a modern personal computer.

All statistical analyses were carried out in the R software package [70]. MIRT was performed with the "mirt" package [71]. Nested MIRT models were compared using the anova function which performs a likelihood ratio test. This work used correlation analysis to investigate the origin of the factor structures extracted. The correlation matrix was presented in a visualization rendered by the "qgraph" package [72]. Partial correlation matrices were constructed by using the "glm" function to regress the total FCI score on the dichotomous scores of the individual items. Factor analysis was carried out with the "factanal" function in the "stats" package. The "nFactors" package was used to generate the Scree plot and to perform parallel analysis. Bootstrapping was performed with the "boot" package [73,74].

F. Supplemental Material

See Supplemental Material [69] for traditional factor analysis including the Scree plot, 3- and 5-factor MIRT models, and the constrained MIRT model without the factor loading on all items [69].

III. RESULTS—EXPLORATORY ANALYSES

The FCI was first examined with MIRT without employing a theoretical model, thus performing an exploratory factor analysis (EFA). Exploratory methods extract models from data without the constraints of an imposed model. Correlation analysis was then used to understand the resulting factor structure. Expert solutions of the FCI were then used to construct a theoretical model of mechanics which allowed further exploration of the correlation structure. In Sec. IV, the work shifts to a confirmatory analysis using MIRT to explore how the theoretical model mapped onto student responses to the FCI. Finally, the model proposed by the FCI authors was fit and compared to the optimal model identified in this work.

A. Exploratory factor analysis

MIRT was used to perform EFA on the FCI. Models with progressively more factors were fit and compared using a likelihood ratio test which computes a chi-squared statistic. A 9-factor model improved model fit over an 8-factor model [$\chi^2(22) = 53.44$, $p < 0.001$] and explained 56% of the variance in the item scores. The last factor added explained 3.6% additional variance. The 10-factor model did not significantly improve model fit. The 9-factor model showed a very strong improvement in AIC and BIC on both the 8-factor and 10-factor model using Raftery's classification [65]. The 9-factor model (varimax rotation) is shown in Table I. Factors are reported as columns and labeled "FC." The table also identifies the FCI problem

TABLE I. Factor loadings for exploratory factor analysis with Multidimensional IRT (varimax rotation). Loadings greater than 0.7 are highlighted in dark gray. Loadings between 0.5 and 0.7 are highlighted in light gray.

#	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	d
1									0.78	8.08
2									0.42	0.90
3		-0.54								3.36
4			0.88							1.38
					Block 5-6					
5						-0.71				0.63
6					0.78					4.81
7					0.64					2.81
					Block 8-11					
8		-0.56			0.35					3.38
9		-0.63								2.18
10		-0.53		-0.32						4.14
11		-0.58								1.81
12					0.33			-0.44		3.40
13		-0.63				-0.41				3.40
14	-0.35							-0.47		1.01
					Block 15-16					
15		-0.52	0.64							0.91
16		-0.39	0.35	-0.43						3.89
17				-0.74						0.37
18						-0.81				0.70
19		-0.58								2.73
20										0.79
					Block 21-24					
21	-0.74									−0.18
22	-0.84									0.83
23	-0.48				0.37		-0.40			2.10
24	-0.39						-0.50			3.96
					Block 25-27					
25				-0.86						0.57
26	-0.39			-0.61						−1.34
27	-0.36									1.52
28			0.74							1.91
29										1.63
30										0.67

blocks. The table reports d; d is related to the overall difficulty of the item [Eq. (2)]. Easier items have larger d. Loadings greater the 0.7 are highlighted in dark gray. Loadings between 0.5 and 0.7 are highlighted in light gray.

While the 9-factor model was statistically superior, the model fit statistics shown in Table II did not provide a clear identification of the number of factors. While the 9-factor model is statistically significantly better than all other models, there was not a significant improvement from the 6-factor model to the 7-factor model [$\chi^2(24) = 32.79$, $p = 0.109$]. The 9-factor model was a significant improvement over the 6-factor model [$\chi^2(69) = 196$, $p < 0.001$]. The 5-factor model had superior RMSEA, CFI, and TLI statistics. While the 9-factor model minimized AIC, the 6-factor model minimized BIC. The knee in the Scree plot calculated using traditional EFA, presented in the Supplemental Material [69], suggests 3 to 4 factors. As such, after 3 factors are extracted, it is difficult to make a definitive case for the number of factors. We selected the 9-factor model because it was the model identified as optimal using the likelihood ratio test, minimized AIC, and provided the greatest resolution of the structure of the instrument. Three- and 5-factor MIRT models are presented in the Supplemental Material [69].

Traditional EFA was also performed. For this analysis, the criterion that the eigenvalue be greater than the mean eigenvalue suggested a 7-factor solution, parallel analysis suggested a 6-factor solution, while an examination of the knee in the Scree plot suggested 3–4 factors. Like other published Scree plots, there was a rapid decline from 1 to 3 factors followed by a long tail where additional factors each explained 2%–4% additional variance. If Huffman and Heller's criteria for the retained factors, which were required to explain 5%–10% of the variance, was used [7], only two factors would have been retained. The 5-factor EFA solution is presented in the Supplemental Material [69]. The 5-factor solution was very similar to other published solutions with many items loading on the first two factors as was also observed by Scott, Schumayer, and Gray [8]. Exploratory methods, such as factor analysis or cluster analysis, can identify structures correlated by unexpected features. The items in the first two factors in either the 5-factor EFA model in the Supplemental Material [69] or in Scott, Schumayer, and Gray do not seem strongly related by the physical principles they test, which opens the possibility that some other feature is causing the correlations which cause groups of items to be identified as factors.

TABLE II. MIRT fit statistics.

Factors	AIC	BIC	RMSEA	TLI	CFI
1	132 042	132 430	0.071	0.83	0.84
2	128 805	129 379	0.047	0.92	0.94
3	127 863	128 619	0.042	0.94	0.95
4	127 223	128 153	0.038	0.95	0.96
5	126 553	127 651	0.032	0.97	0.98
6	126 239	127 498	0.066	0.85	0.91
7	126 254	127 668	0.071	0.83	0.91
8	126 192	127 755	0.067	0.85	0.92
9	126 180	127 885	0.060	0.88	0.94
10	126 214	128 055	0.066	0.86	0.94

B. Correlation analysis

Factor analysis accomplished either traditionally or through MIRT identifies combinations of items which vary together. Covariation of individual items can also be examined through correlation analysis. The full FCI correlation matrix contains 900 entries making it difficult to interpret; however, numerous visualizations of the correlation matrix have been created. Figure 1 presents one such visualization of the FCI correlation matrix created with the R qgraph package. Solid lines (green) represent positive correlations and dashed lines (red) represent negative correlations. Only correlations greater than 0.3 (Cohen's criteria for medium effect size) are shown. No pair of questions was negatively correlated with $|r| > 0.3$ where r

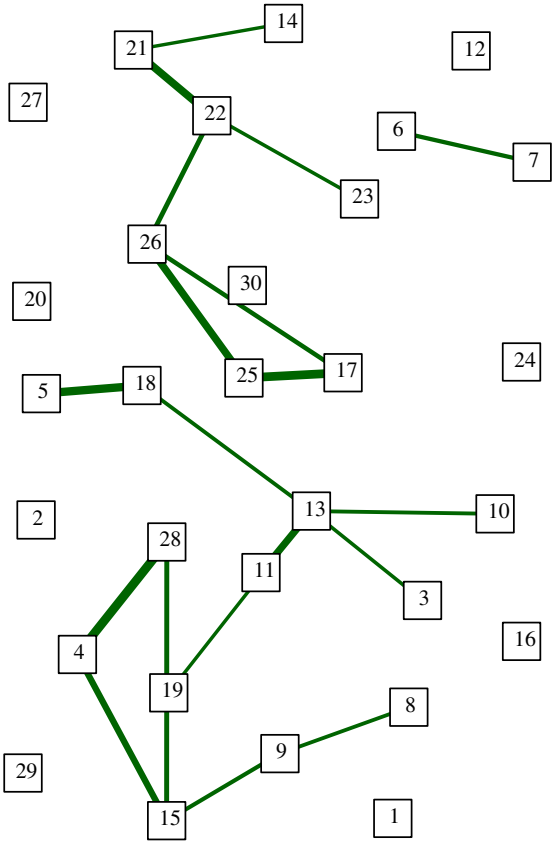

FIG. 1. Correlation matrix for all FCI items. Lines represent correlations with $|r| > 0.3$. Line thickness represents the size of the correlation. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations. No negative correlations were present.

is the correlation coefficient and, therefore, there are no dashed lines in the figure. The placement of nodes is calculated to be visually appealing and does not convey additional information; only the connections between nodes are important. Other visualizations are also useful; an alternate visualization created with the corrplot package is provided in the Supplemental Material [69].

There are many potential sources of the correlations shown in Fig. 1. Groups of highly correlated items often form the elements of a factor with the highest loading; in some sense they "nucleate" the factor. Some correlations may arise because two items require similar physical principles for their solution or that they elicit the same misconception. In previous factor analysis, only these explanations of the factor structure have been considered.

The FCI contains 4 groups of problems where each item in the group shares a common stem; we will call these groups "problem blocks." The problem blocks have been identified in Table I. One additional group of items 25–27 does not share the same stem, but items 26 and 27 explicitly refer to item 25. While blocking the problems may shorten the reading time for the student, it can also generate correlations between items that are not the result of the physical properties required for their solution. If a student misinterprets the stem, then this error will affect the solution of each problem in the block. An error in an earlier item in a block can cause errors in later items. Examination of Table I shows that many of the largest factor loadings in individual factors occur for problems in the same block; likewise, in Fig. 1 many of the most strongly correlated item pairs are part of problem blocks. An examination of the physical principles required to solve the strongly correlated blocked problems does not suggest the level of commonality demonstrated by the factor or correlation structure. As such, it seems likely that at least some of the factor and correlation structure results for the decision to use groups of problems with a common stem.

A second possible source of correlations not related to underlying physical principles is correlation through total test score. The FCI has repeatedly been shown to be an instrument with high internal consistency as measured by Cronbach's alpha [16,21]. All correlations with $|r| > 0.3$ are positive in Fig. 1. Two problems could be correlated because either only the strongest students answer them correctly or only the weakest students answer them incorrectly; they are correlated through the total test score. To remove this effect, a partial correlation matrix controlling for total test score was calculated as shown in Fig. 2. Examination of Fig. 2 shows that the problem blocks $\{8, 9\}$, $\{21, 22, 23, 24\}$, and $\{25, 26\}$ still stand out as highly correlated. Four other groups of questions emerge as correlated $\{5, 18\}$, $\{6, 7\}$, $\{17, 25\}$, and $\{4, 15, 28\}$. To understand these groups, we construct a model of the solution to the FCI in the next section.

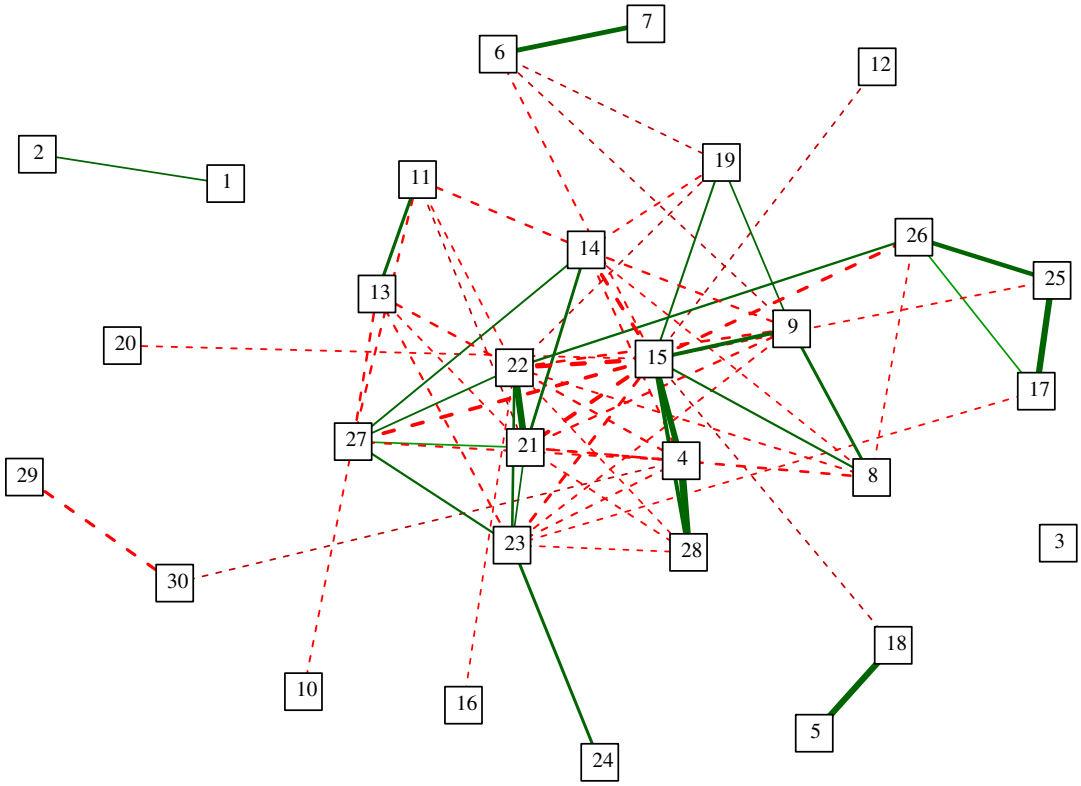

FIG. 2. Partial correlation matrix for all FCI problems correcting for total FCI score (only $|r| > 0.1$ shown). Line thickness represents the size of the correlation. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations.

C. A theoretical framework

Hundreds of physicists have offered models of the structure of introductory mechanics either through the production of textbooks, scientific papers, or in their solution of introductory mechanics problems. We sought to produce one such model that synthesized the structure of introductory mechanics commonly presented in textbooks with the statements found in expert solutions of FCI problems. This resulted in a set of statements about introductory mechanics shown in Table III; the statements will be called "principles" following Larkin *et al.* [55]. The principles were classified as definitions (DF), laws (L),

TABLE III. Theoretical model of Newtonian mechanics as tested by the FCI. Principles in bold were included in the optimal model 3 fitting the reduced FCI.

Label	Derived from	FCI No.	Principle
Kinematics			
DF1		19, 20	Definition of velocity ($\vec{v} = d\vec{r}/dt$).
DF2			Definition of acceleration ($\vec{a} = d\vec{v}/dt$).
R1			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	DF1	6, 7	Instantaneous velocity is tangent to the trajectory.
C2	DF2	5, 18	Objects moving in a curved trajectory will experience centripetal acceleration.
C3	R1		1D trajectory a = constant, ($x(t) = x_0 + v_0 t + \frac{1}{2}at^2$).
C4	R2		1D velocity a = constant, ($v(t) = v_0 + at$).
LM1	DF1	14	If two objects move together, they have the same initial velocity when separated.
LM2	R1	2	Motion may be separated along orthogonal axes.
LM3	C3	2	If motion is one-dimensional and $a = 0$, then $d = vt$.
LM4	R2	3, 22, 26	Objects under constant acceleration with $\vec{a}$ parallel to $\vec{v}$ speed up.
LM5	R2	27	Objects under constant acceleration with $\vec{a}$ opposite to $\vec{v}$ slow down.
LM6	R1	12, 14, 21	Objects under constant acceleration with some initial velocity perpendicular to the acceleration travel in a parabolic arc.
LM7	DF2	20	If velocity is constant, then acceleration is zero.
LM11	C3	1, 2	If the accelerations and initial velocities are equal, objects travel the same distance in the same time.
Dynamics			
DF3		26	The net force is the vector sum of the forces (forces add as vectors).
L1		6, 7, 8, 10, 17, 23, 24, 25	Newton's 1st law.
L2		5, 18, 26, 27	Newton's 2nd law.
L3		4, 15, 16, 28	Newton's 3rd law.
LM8	L2	21	Constant force produces constant acceleration.
LM9	L2	8, 21	If the force only has one component, an object accelerates in that direction.
LM10	DF3	17, 25	If the net force is zero and only two forces are exerted on the object, they must be equal but opposite.
Properties of forces			
L4		1, 2, 3, 5, 11, 12, 13, 14, 17, 18, 29, 30	Objects near the earth's surface experience a constant downward force/acceleration of gravity.
F1		11, 29	An object in contact with a surface experiences a normal force.
F2		11, 13, 18, 30	An object does not necessarily experience a force in the direction of motion.
F3		3, 29	Air pressure does not exert a net downward force.
F4		30	The wind can exert a force on an object.
F5		1, 3	Air resistance is negligible for a compact object moving a short distance.
F6		3	The force of gravity is approximately constant near the earth's surface.
F7		27	Objects that slide across a surface experience a force of friction opposite motion.
Other			
DF4			Magnitude of vector $\left(\lvert\vec{A}\rvert = \sqrt{A_x^2 + A_y^2 + A_z^2}\right)$
C5	DF4	9	Triangle inequality
RS1		19	If one quantity is constant and another quantity is smaller at one time and larger at another time, then the two quantities must be equal at some time.

corollaries (C), results (R), facts (F), lemmas (LM), and reasoning (RS). Corollaries could be derived from laws, results, and definitions but required some nontrivial reasoning. A result, such as the constant acceleration kinematic equations, was derived as a special case of the laws and definitions. Knowledge of how the universe worked that did not raise to the level of a law were called facts. Expert solutions often contained specializations of the physical laws and definitions to the individual problem; these special cases were called lemmas. The FCI contains one item (item 19) which required a unique piece of reasoning (RS1) in multiple expert solutions. To solve the problem, one must argue if one quantity is constant and another begins smaller than the first quantity and ends larger than that quantity, then the quantities must be equal at some point. Many of the principles in Table III are consistent with principles used in models of physics problem solving proposed by Larkin *et al.* [55] and Reif and Heller [35]. The principles can be divided into two broad classes: *core* principles including the definitions, laws, facts, corollaries, and results and *supplementary* principles including the lemmas and reasoning. Without the core principles, the description of Newtonian mechanics is incomplete; supplementary principles specialize core principles to specific situations or provide specific patterns of reasoning.

To map out the subset of Newtonian mechanics tested by the FCI, a careful solution of the FCI was collected from the lead instructor who oversaw the course studied. Solutions were also collected from faculty and graduate students in the research team. These solutions were decomposed to the sentence level and each sentence classified. These statements did not contain many of the core principles shown in Table III. Many lemmas, however, provided specializations of more general core principles not found in the expert solutions. The core principles were introduced based on the project team's understanding of Newtonian physics. For example, the expert solutions contained lemma LM7 ("If velocity is constant, then acceleration is zero"), but did not contain the more general definition DF2 ("$\vec{a} = d\vec{v}/dt$"), so DF2 was added to the model. A core principle was introduced for each lemma; often many lemmas were derived from a single core principle. With only a small sample, it became obvious that a complete set of supplementary principles would be very long and not particularly useful, but that the existing lemmas fit well into a well-established structure of Newtonian mechanics involving the core principles. The model of Newtonian mechanics as measured by the FCI produced by this process is shown in Table III. The table also shows the core principle from which a supplementary principle can be derived and the FCI items whose solution requires the principle.

The model in Table III represents a preliminary model for understanding solutions of the FCI. It does not contain any representation of student misconceptions. The set of lemmas would almost certainly change somewhat if a different set of expert solutions were used. Some parts of the core model would be agreed upon by most experts: DF1, DF2, L2, and L3, for example. However, it is doubtful that a group of experts would agree on all elements. For example, it might be argued that Newton's 1st law is unnecessary because it can be derived from Newton's 2nd law and kinematics. Also, it might be argued that separate principles for one-dimensional kinematics (C3 and C4) and three-dimensional kinematics (R1 and R2) are unnecessary. In Sec. IV, MIRT is used to explore possible changes to the model and identify the model which most strongly captures the Newtonian thinking of this student population.

We note that the fact F2 might be considered to specifically address the motion-implies-force misconception. It was present in most expert solutions to eliminate specific distractors. We will find that its inclusion improves the model and future work may identify other facts that allow common misconceptions to be added to FCI models.

There were some additional minor decisions made to produce the model in Table III. Item 17 has a distractor that requires the application of F3 (net downward force of the air); no expert solution included this principle and it was not included in the model of item 18. LM4 and LM5 were written for general three-dimensional motion and are marked as derived from R2. Items 26 and 27, which use these lemmas, are one-dimensional problems. As the lemmas are folded into the principles they are derived from to produce model 3, the items using these lemmas will be appropriately distributed to one- or three-dimensional kinematic principles. Item 18 was coded with a centripetal acceleration implying a force in the direction of the tension force; this item could have also been coded by introducing the tension force as an additional fact. The correlation with item 5 and the lack of any additional items using a tension force caused the selection of this coding. Law L4 and fact F6 both involve a constant force of gravity near Earth's surface. Fact F6 was introduced because FCI item 3 seems to require the student to explicitly reason that the force of gravity does not change much over the height of a single-story building.

D. Reduced exploratory factor analysis

The theoretical model in Table III provides an explanation for some of the remaining strong correlations in Fig. 2 which were not explained by the block structure of the FCI. Items 4, 15, and 28 all require only L3 (Newton's 3rd law) for their solution. Items 17 and 25 share both L1 and LM10, items 5 and 18 share L2, L4, and C2, and items 6 and 7 share C1 and L1. Item 16 also only requires Newton's 3rd law; however, this item was not as strongly correlated with the other Newton's 3rd law items in Fig. 2. While Newton's 3rd law plays a central role in Newtonian mechanics and, therefore, one would expect it to be repeated multiple times in the FCI, the repetition of the other combinations of

principles is difficult to support theoretically as combinations somehow central to mechanics and thus deserving special focus. The FCI authors did not discuss the choice to include the problem pairs $\{5, 18\}$, $\{6, 7\}$, and $\{17, 25\}$ and, therefore, it seems likely the inclusion of these pairs of very similar problems was accidental. The inclusion of these problems does not affect the ability of the instrument to measure an overall force concept beyond the reduction of the breadth of the instrument; however, the repetition of these problems does impact the correlation and exploratory factor structure. Figure 2 shows the scores on these pairs of problems are highly correlated and these pairs make up the strongest loading in factors FC4, FC5, and FC6 in Table I. It seems likely that the strong correlations of the pairs was part of the reason these factors were extracted and that the factor structure could be significantly modified by removing one problem of each pair and inserting problems that repeated a different set of principles. As such, any general conclusion drawn from the existence factors FC4, FC5, or FC6 about the structure of knowledge of Newtonian mechanics is suspect.

These factors based largely on pairs of questions also serve to explain the relatively universal structure of the Scree plots reported in this and other works. The Scree plots reported all decreased strongly from 1 to 3 factors and then the amount of variance explained by additional factors diminishes rapidly. If a factor is mostly capturing the covariance of two items, the amount of variance it can explain will be small.

With these observations, much of the original factor structure identified by EFA appears to be a result either of the block structure of the FCI or of repeated problems with very similar solution structure. Removing all but the first problem in each problem block and the second of the repeated problem pairs produces a reduced 18-item instrument. Because item 6 was removed due to blocking, item 7 was retained. The Newton's 3rd law items were also retained because of the centrality of this principle to Newtonian physics. The optimal MIRT model for this set of problems is shown in Table IV; 6 factors were optimal.

Examination of Table IV shows some factors that map onto the theoretical model of mechanics. The problems have been placed in a Venn diagram in Fig. 3 based on the general classification in Table III. All FCI items have been included in the diagram. Items removed to eliminate blocking are bolded. Unfair items identified by Traxler *et al.* [16] are underlined; these will be discussed later. Few factors contain loadings that are localized to individual regions of the Venn diagram. There are also loadings that cannot be supported theoretically. Factor FC3 contains the Newton's 3rd law items, but it also loads on items 1 and 8 which have nothing to do with Newton's 3rd law. Likewise, item 15, which requires only Newton's 3rd law for its solution, also loads strongly on FC2. It is also difficult to understand why item 17 (force in elevator) and item 20 (blocks moving at different speeds) form factor FC1. It is unclear if correlations through the overall difficulty of the item could explain some of the unexpected structure.

TABLE IV. Exploratory factor analysis for the reduced FCI (varimax rotation). Only loadings greater than 0.3 are shown. Loadings greater than 0.7 are highlighted in dark gray. Loadings between 0.5 and 0.7 are highlighted in light gray.

FCI #	FC1	FC2	FC3	FC4	FC5	FC6	d
1			0.31			0.75	6.25
2						0.65	0.93
3		-0.60					3.35
4			0.86				1.15
5		-0.34		-0.32			0.32
7				-0.33	-0.43		2.28
8		-0.48	0.33		-0.36		3.06
12				-0.55			2.82
13		-0.70					3.02
14				-0.67			0.65
15		-0.47	0.64				1.04
17	-0.33						0.18
19		-0.68					2.90
20	-0.41						0.76
21				-0.62			−0.47
28			0.81				1.92
29							1.67
30		-0.35		-0.33			0.51

IV. RESULTS—CONFIRMATORY ANALYSES

The exploratory analysis of the previous section failed to extract a factor structure that was understandable within a theoretical model of Newtonian mechanics (Table III). For over 50 years, social scientists have argued that research should not rely purely on exploratory techniques but rather that having a robust theoretical framework is paramount to the determination of model validity [75]. According to

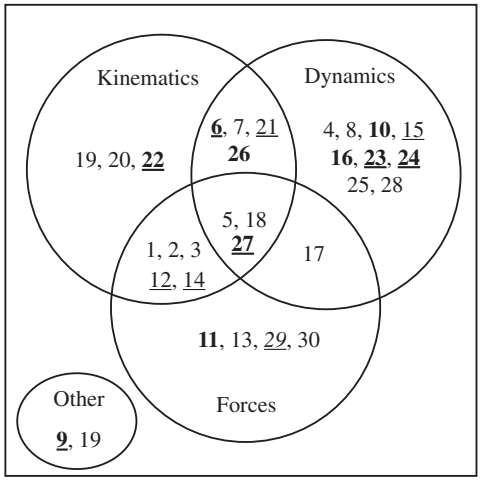

FIG. 3. Venn diagram of the distribution of problems in the FCI. Items in bold are the blocked items removed from the analysis. Underlined items are items identified as unfair to men or women by Traxler *et al.* [16] (Item 29 was identified as fair but unreliable).

Cronbach and Meele, there is no validity without an articulated theory and it is, therefore, inappropriate to use only exploratory techniques, such as EFA, on an instrument. Furthermore, EFA results provide only information about the data itself and should not be construed as providing genuine answers or solutions without a theoretical core [76]. Exploratory methods generally identify some structure, and without a framework that structure may simply be the result of random fluctuations in the data.

Confirmatory analysis instead proceeds from the previously articulated theoretical model and explores how that model can be used to understand the data. Often confirmatory analysis starts with fitting the full theoretical model and then examines a small number of theoretically motivated modifications to the model. The theoretical model of Newtonian mechanics presented in Table III was used as the starting point for a confirmatory analysis of the FCI. MIRT allows the exploration of this model by constraining the MIRT parameter matrix to the model. This is analogous to a confirmatory factor analysis (CFA), where the analysis proceeds from the theoretical model and determines how well the data fit the model. Constrained MIRT is not fully equivalent to CFA because they proceed from different underlying statistical models, but the method of exploring related models is equivalent.

A. Constrained MIRT

MIRT allows the exploration of student thinking about Newtonian mechanics by constraining the parameter matrix to a model. The $\boldsymbol{a}_j$ parameter matrix can be constrained so that parameters that should not theoretically affect a factor are zero. For example, if the model of Newtonian mechanics in Table III was used as the basis for a constrained MIRT model, then the factor representing DF1, a_{DF1}, could be constrained to be zero except for items 19 and 20. For this analysis, only the first problem in a problem block was retained as before; groups of similar problems $\{5, 18\}$, $\{6, 7\}$, $\{17, 25\}$, and $\{4, 15, 28\}$ were also retained. Because constrained MIRT is not exploratory, the correlations of these items will not unduly influence the analysis. The 20-item problem set analyzed in this section was then the following: 1, 2, 3, 4, 5, 7, 8, 12, 13, 14, 15, 17, 18, 19, 20, 21, 25, 28, 29, and 30.

The starting model for the confirmatory analysis included all the principles introduced in Table III which were not eliminated by removing blocked items. F7 and C5 were eliminated when blocked items were removed. The FCI has strong internal consistency and most items are positively correlated. To separate a general facility with Newtonian mechanics from a specific facility with one of the principles, an additional factor was added that loaded on all items. The fit statistics of this model, model 1, are shown in Table V. Because the parameter matrix was so sparse, fit parameters such as CFI, RMSEA, and TLI could not be calculated. Fit statistics in Table V apply to the transformed model number, as such, model 1 is transformed model 1. While some model fit measures were not available, model fit can be examined by the amount of AIC and BIC changes between models and ultimately from bootstrapping, which will show most parameters in the best fitting model have standard deviations that suggest the parameters are significantly different from zero.

After the full model is fit, confirmatory analysis examines theoretically motivated simplifications of the full model. Each transformation in Table V modified the original model to the transformed model. A likelihood ratio test determined whether the models were statistically different. Model 4 did not change the number of degrees of freedom from model 3; therefore, a chi-squared test could not be performed; however, AIC and BIC could be compared. Some transformations removed a principle from a previous model; other transformations combined two principles. For example, in model 5 all items that loaded on either L1 or L2 were set to load on only L2. These models do not exhaust the set of available models, but represented a set of models where a theoretical case could be made for each change.

Model 2 tested a fundamental question about the granularity of student knowledge of the FCI. The set of possible supplementary principles (reasoning and lemmas) is quite

TABLE V. Hierarchical MIRT modeling. The χ^2 test determines whether the models are statistically different; if so, it measures the improvement of the superior model over the inferior model.

Transformed model	Transformation	Original model	AIC	BIC	Chi-squared test	Superior model
1			91 067	91 668		
2	Remove all lemmas.	1	90 943	91 518	$\chi^2(4) = 116,\ p < 0.001$	2
3	Remove RS1.	2	90 920	91 488	$\chi^2(1) = 21,\ p < 0.001$	3
4	Combine DF3 with L2.	3	90 942	91 510		3
5	Combine L1 with L2.	3	90 929	91 491	$\chi^2(1) = 11,\ p = 0.001$	3
6	Combine C3 with R1; C4 with R2.	3	90 991	91 553	$\chi^2(1) = 73,\ p < 0.001$	3
7	Remove F2.	3	90 941	91 490	$\chi^2(3) = 27,\ p < 0.001$	3
8	Remove F5.	3	90 944	91 499	$\chi^2(2) = 28,\ p < 0.001$	3
9	Remove F6.	3	90 929	91 491	$\chi^2(1) = 11,\ p = 0.001$	3
10	Replace L1 with L2 and DF2.	3	90 933	91 521	$\chi^2(3) = 7.5,\ p = 0.058$	3

large while the set more general core principles (laws, facts, definitions, corollaries, and results) are substantially smaller. Each lemma represented a qualitative interpretation or a special case of a core principle. To determine if the lemmas were important to the understanding of the pattern of answers, model 2 was constructed which removed all lemmas and replaced them with the core principle from which they were derived. Model 2 was a significant improvement over model 1 with very strong changes in AIC and BIC and, therefore, the answering pattern for this sample could be understood without the lemmas. Student thinking about the FCI is better understood in terms of a short list of core principles rather than the extensive lists of qualitative lemmas derived from the core principle. This provides important insight into the number of principles needed to understand student Newtonian thinking while also substantially simplifying the research effort. The model without the lemmas could have been produced by any physics graduate student and should be much less dependent on the experts providing the solutions.

Confirmatory exploration continued by testing a sequence of models either motivated by questions that arose about what part of the core principles the FCI measured or questions about relations between the core principles. For each step, the difference in AIC and BIC between the better fitting model and the less well fitting model are reported. Model 3 removed the crossing reasoning step RS1 from model 2; this improved model fit (very strong change in AIC, strong change in BIC). RS1 was used only in a subset of expert responses; other experts simply observed that two of the interval lengths were comparable. Model 4 explored whether the vector addition of forces could be viewed as a part of Newton's 2nd law by combining L2 and DF3; model 3 was a significant improvement over model 4 (very strong change in AIC and BIC). These students answer Newton's 2nd law questions and addition of forces questions with different facility. Combining Newton's 1st law (L1) and Newton's 2nd law (L2) to form model 5 from model 3 also did not improve model fit over model 3 (strong change in AIC, very strong change in BIC). A second model that eliminated Newton's 1st law from model 3, model 10, replaced L1 with L2 (Newton's 2nd law) and DF2 (the definition of acceleration). This model was not statistically superior to model 3 and the model increased both AIC (very strong) and BIC (very strong). As such, L1 was retained as a separate entity. Combining C3 and C4 representing one-dimensional kinematics into R1 and R2 representing three-dimensional kinematics to form model 6 did not improve model fit over model 3 (very strong change in AIC and BIC). Fact F2 (there is not necessarily a force in the direction of motion) addresses a common misconception; removing F2 from model 3 to form model 7 did not improve model fit (very strong change in AIC and BIC). Finally, facts F5 (air resistance is negligible) and F6 (gravity is approximately constant) are additional pieces of information about mechanics; however, their use was only required to eliminate distractors and they were not used by some experts who solved the problem without considering the distractors. Neither model 8 which eliminated F5 from model 3 (very strong change in AIC and BIC) nor model 9 (strong change in AIC, very strong change in BIC) which eliminated F6 from model 3 improved model fit. As such, model 3, which contains only core principles, all of Newton's 3 laws with a separate definition of the addition of forces, leaves 1D and 3D kinematics separate, and contains facts 1–6, represented the best model of students' responses to the FCI. Interestingly, model 3 is probably closest to the model presented in traditional textbooks. Model 3 was also the model which minimized both AIC and BIC.

Model 3 with the transformations applied is presented in Table VI. The discrimination parameters for model 3 are presented in Table VII. For this model, the a_0 coefficient represents the factor that was loaded on all items representing the overall discrimination of the item. To allow comparison with the more intuitive 2PL model, an effective difficulty, b_j,

TABLE VI. Principles included in the optimal model of the FCI, model 3. Items in bold are the blocked items removed from the analysis. Underlined items are items identified as unfair to men or women by Traxler *et al.* [16] (Item 29 was identified as fair but unreliable).

Principle	Derived from	FCI No.
Kinematics		
DF1		14, 19, 20
DF2		20
R1		2, 12, 14, 21
R2		3, **22**
C1	DF1	**6**, 7
C2	DF2	5, 18
C3	R1	1, 2
C4	R2	**26**, **27**
Dynamics		
DF3		17, 25, **26**
L1		**6**, 7, 8, **10**, 17, **23**, **24**, 25
L2		5, 8, 18, 21, **26**, **27**
L3		4, 15, **16**, 28
Properties of forces		
L4		1, 2, 3, 5, **11**, 12, 13, 14, 17, 18, *29*, 30
F1		**11**, *29*
F2		**11**, 13, 18, 30
F3		3, *29*
F4		30
F5		1, 3
F6		3
F7		**27**
Other		
C5		**9**

TABLE VII. Optimal MIRT model 3. 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 b is the difficulty of the item.

FCI No.	Principles	a_0	b
1	C3(1.04 ± 0.30) L4(−0.30 ± 0.19) F5(0.09 ± 0.11)	2.25 ± 0.44	−3.30 ± 0.22
2	R1(0.06 ± 0.05) C3(0.48 ± 0.12) L4(0.09 ± 0.05)	1.05 ± 0.07	−0.86 ± 0.06
3	R2(0.13 ± 0.09) L4(0.02 ± 0.12) F3(0.14 ± 0.09) F5(0.13 ± 0.09) F6(0.15 ± 0.09)	1.65 ± 0.19	−2.50 ± 0.12
4	L3(2.37 ± 0.29)	1.88 ± 0.19	−0.72 ± 0.05
5	C2(0.64 ± 0.15) L2(0.51 ± 0.10) L4(0.50 ± 0.11)	1.49 ± 0.13	−0.38 ± 0.05
7	C1(0.16 ± 0.09) L1(0.01 ± 0.05)	0.64 ± 0.06	−3.42 ± 0.26
8	L1(−0.27 ± 0.08) L2(−0.30 ± 0.09)	1.41 ± 0.12	−2.18 ± 0.09
12	R1(0.55 ± 0.08) L4(0.18 ± 0.07)	0.75 ± 0.07	−3.73 ± 0.31
13	L4(0.29 ± 0.10) F2(0.27 ± 0.09)	2.36 ± 0.17	−1.34 ± 0.05
14	DF1(0.22 ± 0.08) R1(1.03 ± 0.15) L4(0.31 ± 0.08)	0.78 ± 0.07	−0.99 ± 0.09
15	L3(0.79 ± 0.05)	0.87 ± 0.06	−0.78 ± 0.07
17	DF3(0.70 ± 0.13) L1(0.60 ± 0.12) L4(0.14 ± 0.06)	1.64 ± 0.13	−0.18 ± 0.04
18	C2(0.65 ± 0.14) L2(0.58 ± 0.11) L4(0.50 ± 0.11) F2(0.27 ± 0.09)	1.71 ± 0.13	−0.31 ± 0.04
19	DF1(0.16 ± 0.08)	1.28 ± 0.08	−2.04 ± 0.09
20	DF1(0.44 ± 0.12) DF2(0.23 ± 0.10)	1.12 ± 0.09	−0.83 ± 0.05
21	R1(0.82 ± 0.10) L2(0.28 ± 0.07)	0.80 ± 0.07	0.62 ± 0.07
25	DF3(0.70 ± 0.13) L1(0.60 ± 0.13)	1.91 ± 0.16	−0.13 ± 0.03
28	L3(1.28 ± 0.09)	1.70 ± 0.09	−0.98 ± 0.05
29	L4(−0.10 ± 0.09) F1(0.09 ± 0.06) F3(0.09 ± 0.06)	0.17 ± 0.06	−12.12 ± 6.28
30	L4(0.29 ± 0.09) F2(0.19 ± 0.07) F4(0.24 ± 0.10)	1.11 ± 0.08	−0.57 ± 0.05

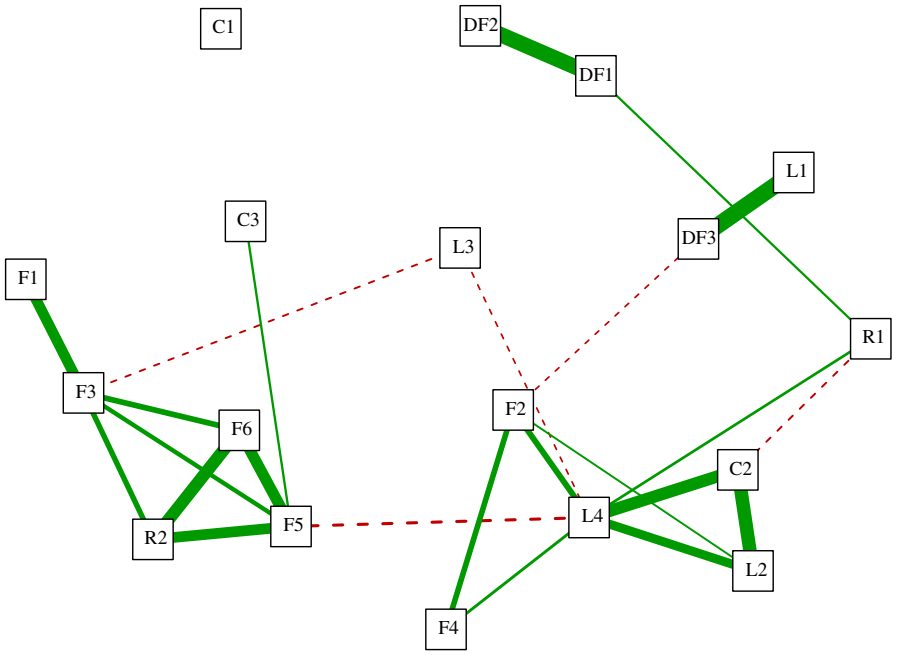

FIG. 4. Correlation matrix of student ability using model 3. Lines represent correlations with $|r| > 0.15$. Line thickness represents the size of the correlation. Solid (green) lines represent positive correlations; dashed (red) lines negative correlations.

is calculated $b_j = -d_j/a_{0j}$. The larger b_j the lower the probability the students will answer the item correctly; the 2PL probability function is shown in Eq. (1). The mirt package does not report normalized latent variables; the standard deviation of each latent variable has been absorbed into the a_j coefficient. Therefore, the a_j coefficient represents the change in log odds if the latent trait increases by 1 standard deviation.

Table VII presents the discrimination of each principle on each FCI item as well as the standard deviation of each item. For example, the discrimination of FCI item 4 on Newton's 3rd law (L3) is 2.37 ± 0.29; a higher discrimination than the other Newton's 3rd law items. The analysis also allows the determination of the relative discrimination of items that test multiple principles. For example, item 21 provides much better discrimination of student knowledge of three-dimensional motion (R1) than Newton's 2nd law (L2). As such, Table VII provides an exceptionally detailed model of what each FCI item measures.

Some alternate forms of the constrained analysis were also performed. The optimal model in Table VII included one factor that loaded on all problems; a factor capturing a student's overall facility with conceptual Newtonian mechanics. The model with this factor (AIC = 90 920, BIC = 91 488) was a significant improvement over the model without this overall factor (AIC = 94 442, BIC = 94 881) [$\chi^2(20) = 3562$, $p < 0.001$] with a very strong change in AIC and BIC. The model with this factor also had superior behavior in tests that compared model 3 to models where additional principles that damaged the model had been introduced. For example, the addition of L3 (Newton's 3rd law) to item 1 produced a significantly less well fitting model with the overall factor, but not without it. The model without this overall factor is presented in the Supplemental Material [69].

MIRT can also be used to estimate the ability of each student to answer each item. The correlations of these abilities are presented in Fig. 4. Because one factor was loaded onto all items, these abilities represent that difference between the student's general ability to solve a conceptual mechanics question and his or her ability to apply a specific principle. For students with a fully developed expert understanding of mechanics, we would expect their ability to apply each principle to be equal, and therefore their difference in ability to be zero. Figure 4 shows multiple principles with large correlations and large differences in the strength of the correlation between different items. From this diagram, we can infer that the students studied have differing but correlated abilities with concepts of velocity and acceleration (DF1, DF2), with Newton's 1st law (L1) and the addition of forces (DF3), and with Newton's 2nd law (L2) and the law of gravitation (L4). Additional instruction may be required to allow students to fully integrate these concepts. MIRT, then, may also represent a tool which can be used to probe the structure of knowledge and to quantitatively characterize expert and novice differences and to localize where additional integration of knowledge is needed.

B. Comparison with the original FCI model

The FCI authors suggested a detailed structure for the FCI dividing the test into 6 general categories and

23 fine-grained principles (see Table I in Ref. [1]). The fine-grained principles play the same role as the principles in the theoretical model in Table III. The FCI was revised in 1995; the revised test included 3 new problems which were not categorized. These items, revised FCI items 5, 18, and 30, will not be included in this analysis.

Fitting a model implementing the structure suggested in the original FCI paper on the set of items 1, 2, 3, 4, 7, 8, 12, 13, 14, 15, 17, 19, 20, 21, 25, 28, and 29 from the revised FCI produced a model with AIC = 75260 and BIC = 75453. Using the constrained MIRT model of the previous section on the more restricted problem set produced a model with substantially better model fit [AIC = 74812; BIC = 75277], a very strong change in AIC and BIC. A likelihood ratio test showed that the constrained MIRT model had significantly better model fit [$\chi^2(13) = 474$, $p < 0.001$]. As such, while the model proposed by the FCI authors captured their motivation as the creators of the instrument, model 3 produced a better fit for this student population.

V. DISCUSSION

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

RQ1: What factor structure is extracted for the FCI by MIRT? Is this structure consistent with the results of other factor analysis? MIRT identified a 9-factor solutions as optimal for the full 30-item FCI. Other studies have identified 5-factor [8] and 6-factor [9] post-test models as optimal. It is possible that the larger sample used in the present study combined with the strong incentives given for correctly answering the items allowed this study to resolve more detailed structure in the FCI. The 9-factor model, while the best statistically based on likelihood ratio tests, was not the best model on all fit statistics (Table II). The fit statistics could also support the identification of either a 5-factor or 6-factor model. All three of these studies identified more factors than Huffman and Heller [7]; however, this may have resulted from the differing size and quality of the samples as well as the different criteria used to select the optimal number of factors.

The factors extracted can also be compared. If the 9-factor solution found in this study resulted because of superior resolution of the factors, we would expect some of the factors in the previously reported models to split to form the additional factors in this study. Some commonality can be found between the 5-factor [8], 6-factor [9], and our 9-factor model. The groups of physically similar items $\{5, 18\}$, $\{6, 7\}$, $\{17, 25\}$, and $\{4, 15, 16, 28\}$ do factor together in all models, except that item 16 often does not factor with the Newton's 3rd law group. The 5-factor model shows the same tendency of blocked items to factor together that we saw in the 9-factor model; this effect was less pronounced in the 6-factor model. All the factor models are difficult to support in terms of the actual structure of the physical principles needed to solve the problems shown in Table III. As such, it is difficult to support the proposition that EFA is providing fundamental insights into the knowledge structure of physics students as measured by the FCI.

RQ2: Can parts of this factor structure be explained by factors other than the structure of student knowledge of Newtonian mechanics?

Correlation analysis identified two nonphysical sources of relations between FCI items which could affect the factor structures: correlations through the blocking of items into groups and correlations through total score. The effect of blocking was clear in Table I with most blocked questions sharing the same factor with the exception of items 5–6. The strong correlation of many blocked items can also be seen in the overall correlation matrix (Fig. 1). Further analysis retained only the first item in each group; the nonphysical correlations created by blocking could not be corrected statistically. While the possible correlation of blocked items seems relatively uncontroversial, we know of no previous research that identifies it as a possible source of a nonphysical perturbation on the factor structure or other analysis. The possible correlation between total score could be deduced through the studies showing the FCI as a very internally consistent instrument [16,21] as well as Huffman and Heller's identification of the FCI as a single-factor instrument [7]. This internal consistency is clearly demonstrated in Fig. 1 showing all correlations are positive. The possibility of the difficulty of an item impacting the factor structure was discussed briefly by Scott, Schumayer, and Gray [8].

The correlations through overall test score were removed by calculating a partial correlation matrix (Fig. 2) which continued to show the effect of problem blocking and revealed a third source of correlation. There were four groups of items in the FCI which are answerable using very similar physical principles. One group, items requiring Newton's 3rd law for their solution, was expected. This group forms one of the factors in each published analysis [8,9,24] except Huffman and Heller [7]. The other three groups do not seem to represent special combinations of reasoning particularly important to understanding mechanics and the repetition of these principles seems likely to be accidental. These groups $\{5, 18\}$, $\{6, 7\}$, and $\{17, 25\}$ had large factor loadings in the same factor in all published models. It seems likely that the repetition of these blocks artificially influenced the factor structure; many other equally important combinations of physical reasoning could have been repeated.

RQ3: If blocked items and repeated reasoning groups are removed, is the resulting factor structure consistent with Newtonian mechanics? An EFA was also presented for a reduced set of FCI items which removed all but the first item in each problem block and removed the second item of the $\{5, 18\}$ and $\{17, 25\}$ groups and the first item of the $\{6, 7\}$ group. This EFA found a 6-factor solution

(Table IV); however, the factors make little physical sense. Factor 1 mixed a Newton's 1st law problem involving an elevator with the analysis of two plots with zero acceleration. Factor 2 contains a mixture of items including Newton's 3rd law, one-dimensional constant acceleration, and a position vs time plot involving objects of constant velocity and acceleration. Factor 3 includes three Newton's 3rd law items but also two-dimensional zero acceleration motion and one-dimensional motion under gravity. As such, factor analysis, once nonphysical or accidental correlations are removed, does not extract a factor structure consistent with Newtonian mechanics. As the designers intended, the FCI is a single-factor instrument [25]. The reason for the coherence can be seen in Fig. 3 where many items test multiple general domains.

RQ4: Can theoretically constrained MIRT produce a model of the physical constructs measured by the FCI? If so, what is the optimal model of the FCI for this student population?

Constrained MIRT allowed a confirmatory exploration of a set of related models grounded in the traditional theoretical framework of Newtonian mechanics. This exploration showed, while expert solutions to the FCI were cast in a number of lemmas which converted the mathematical framework of mechanics to language-based principles, that these were not needed to understand the structure of student understanding. This implies student thinking can be productively understood by a set of core principles grounded in the model of Table VI.

The optimal model 3 supported the differentiation in student thought between Newton's 1st law and Newton's 2nd law as well as the difference between one-dimensional and three-dimensional constant acceleration kinematics. Facility with the vector addition of forces was also shown to be distinct from facility with Newton's 2nd law.

Table VII shows the optimal MIRT model 3. The number in parenthesis next to the principle label is the discrimination for the principle. Because an overall factor loading on all items was included, a_0, the discrimination, $a_{j>0}$, of the individual principles represents the additional effect of the specific ability over the student's general ability with Newtonian mechanics. Some of the discrimination parameters are very small indicating that the item does not require additional facility with the principle over the student's general ability to answer FCI questions correctly. Some discriminations are negative which may be a sign of a problematic item. Items with only one strongly discriminating principle might be claimed to be good marker items for the skill represented by the principle. Items 1, 2, 12, 14, and 21 require multiple principles but discriminate on one principle more strongly than the others. These questions might be used to characterize students' knowledge on the high discrimination principle. Items 4, 15, 19, and 28 require only one principle, and therefore could be used as a measure ability to perform this principle; however, three of the four represent Newton's 3rd law. Items 5, 17, 18, and 25 require multiple principles with commensurate and large discriminations. These items measure multiple abilities at the same time, but do not differentiate between the abilities. Finally, a number of items have small discrimination values for all principles: items 3, 7, 13, 29, and 30. These items do not contribute additional information about specific abilities. Item 8 had negative discrimination; this may indicate the item is not functioning correctly.

MIRT provides a new lens for examining physics evaluations. If this lens proves valuable, it will suggest certain desirable properties in future evaluations. First, the structure and number of items should allow noncompensatory MIRT models to be fit to extract item or principle-level difficulty parameters. Second, each item should provide additional information about some ability. Third, the instrument should be invertible so that a linear combination of the scores on a subset of items provides an estimate of the ability for a each principle, thus giving practitioners a detailed characterization of their learning outcomes.

RQ5: Does the structure proposed by the FCI's authors provide a superior description of the instrument to the optimal model identified by MIRT? The structure suggested by the authors of the FCI [1] was also fit to the data set and the result compared to the optimal model 3 identified by MIRT. Model 3 outperformed the model suggested with the publication of the FCI. As such, part of the reason the published structure has not been recovered may be that other models fit the FCI better. This seems unlikely to be the primary reason for the mismatch between the proposed model and model 3. Table VII and Fig. 3 as well as Hestenes and Halloun's insistence that the FCI measures a single Newtonian force concept [25] show that the instrument simply was not constructed to factor well. There are very few items that use a single principle and only Newton's 3rd law, not Newton's 1st or 2nd law, is used independently and is repeated multiple times in the unblocked model (Table VII). Most FCI items measure multiple physical principles at once.

This work identified the blocking of items in the FCI as a source of correlations not related to the student's ability to answer conceptual physics questions. To eliminate these correlations, only the first item in a problem block should be retained; as such, items 6, 9, 10, 11, 16, 22, 23, 24, 26, and 27 were removed from the FCI producing a 20-item version of the FCI. The model in Table VI can be used to understand the effect of this reduction. The blocked items to be removed are shown in bold in both Table VI and Fig. 3. Removing these items eliminated principles F7 and C5 while reducing coverage of R2 and C4. In general, these reductions still leave the coverage of the FCI intact although the elimination of an explicit use of friction is a loss.

Traxler *et al.* [16] also suggested a reduced 19-item instrument (including FCI questions: 1, 2, 3, 4, 5, 7, 8, 10, 11, 13, 16, 17, 18, 19, 20, 25, 26, 28, and 30) to remove

items with reliability problems and to remove items unfair to either men or women. The items removed to produce the 19-item instrument are underlined in Table VI and Fig. 3. While this reduction removes seven items from both kinematics and dynamics in Table VI, the coverage of kinematics required more principles than that of dynamics. The removal of unfair items from R1, R2, and C4 may substantially change the coverage of the instrument. Removing both blocked and unfair items further reduces the coverage.

To produce a fair instrument while maintaining coverage, it may be necessary to retain some blocked and unfair items but to balance the degree and number of unfair items for both men and women. Traxler *et al.* [16] reported that two of the removed items were unfair to men, items 9 and 15. If these two items are retained as well as items 14 and 27, which were unfair to women with similar differential item functioning statistics, the overall score should be gender fair. Blocked items 11 and 26 could also be retained to maintain coverage. Retaining these items would increase coverage of some kinematic principles while providing coverage of F7 and C5. This would leave a reduced 21-item FCI instrument containing items 1, 2, 3, 4, 5, 7, 8, 9, 11, 13, 14, 15, 17, 18, 19, 20, 25, 26, 27, 28, and 30. Blocked items 10 and 16 were removed because there was sufficient coverage of the principles required for their solution.

VI. LIMITATIONS

This work was performed with a single sample drawn from a single institution. Additional studies are necessary to determine if the conclusions are general. The sample was analyzed in aggregate; additional analysis is needed to determine if the results apply to all student subgroups. The analysis did not consider the role of misconceptions; an extended theoretical model including misconceptions should also be tested.

This work began with a model constructed from a sample of expert solutions at a single institution. Alternate models certainly can and should be constructed; MIRT provides the tool needed to determine which model best fits student thinking. The model presented in this work should not be considered the end point, but the beginning of a more detailed exploration of conceptual Newtonian mechanics that will take many years to complete. For researchers wishing to test alternate models or compare models between institutions, contact the corresponding author to request the data.

VII. IMPLICATIONS

This worked showed a theoretical model of introductory mechanics could be useful in understanding the results of conceptual inventories. Such models can be constructed for other conceptual areas of physics and could form a basic tool for understanding the detailed results of PER instructional innovations. The constrained MIRT analysis technique allowed the fine-grained exploration of the constructs measured by the FCI and may be a powerful tool for improving our understanding of student knowledge. EFA did not produce a factor structure that was useful in understanding the FCI and it is likely that purely exploratory tools may not yield generalizable results. Part of the reason for the failure of EFA was correlations produced by the blocks of questions in the FCI. The practice of using blocks of questions with the same stem may make PER instruments difficult to interpret statistically and should be discontinued.

This work showed that if all blocked items identified as problematic because of correlations produced by blocking and all items identified as unfair or unreliable by Traxler *et al.* [16] are removed that the coverage of kinematics of the modified FCI is reduced. This work proposed a 21-item reduced FCI to maintain coverage while balancing unfair items; to have to decide between coverage and fairness is unacceptable. While this 21-item instrument could be used for the near future, the identification of unfair items and blocked items as problematic in addition to the lack of coherent sub-scales suggest that it is time to revisit the construction of the FCI and to modernize it to remove some of the difficulties identified in recent research.

VIII. FUTURE WORK

This work will be extended to analyze other conceptual instruments popular in PER including the FMCE [3] and the CSEM [4]. The work will also be extended to determine if the results are consistent between men and women and to determine if this method can help in understanding the differences observed in men's and women's performance on conceptual evaluations.

IX. CONCLUSIONS

The work examined the structure of the FCI with Multidimensional Item Response Theory; first as an exploratory method and then as a confirmatory method using constrained MIRT with a theoretical model of Newtonian mechanics. The exploratory analysis identified a 9-factor solution that showed some similarities to previously published solutions. Further analysis showed many of the factors in the 9-factor solution and the previously published solutions could have resulted from the use of multiple problem blocks and the repetition of physically similar items. Exploratory factor analysis was repeated, removing these correlated items; the resulting 6-factor solution could not be reconciled with the theoretical structure of Newtonian mechanics. Constrained MIRT was then employed to determine the optimal model of the FCI for the student population studied within the framework of a theoretical model. The optimal model contained only core principles of mechanics and did not contain subsidiary principles derived from these core principles. The optimal model differentiated between Newton's 1st and 2nd law;

between Newton's 2nd law and the principle of vector addition of forces; and between one-dimensional and three-dimensional kinematics. The optimal model identified by MIRT was substantially statistically superior to the original model proposed by the authors of the FCI.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, PHY-0108787.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[2] 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).
[3] 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).
[4] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Huevelen, Surveying students' conceptual knowledge of electricity and magnetism, Phys. Ed. Res., Am. J. Phys. **69**, S12 (2001).
[5] 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).
[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] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure? Phys. Teach. **33**, 138 (1995).
[8] 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).
[9] 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).
[10] E. Brewe, J. Bruun, and I. G. Bearden, Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **12**, 020131 (2016).
[11] R. P. Springuel, M. C. Wittmann, and J. R. Thompson, Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics, Phys. Rev. ST Phys. Educ. Res. **3**, 020107 (2007).
[12] J. Stewart, M. Miller, C. Audo, and G. Stewart, Using cluster analysis to identify patterns in students' responses to contextually different conceptual problems, Phys. Rev. ST Phys. Educ. Res. **8**, 020112 (2012).
[13] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[14] 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).
[15] 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).
[16] 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).
[17] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, Am. J. Phys. **74**, 449 (2006).
[18] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80**, 825 (2012).
[19] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).
[20] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).
[21] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).
[22] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).
[23] 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).
[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] 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).
[26] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[27] S. Ramlo, Validity and reliability of the force and motion conceptual evaluation, Am. J. Phys. **76**, 882 (2008).
[28] 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).
[29] 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).

[30] G. S. Gliner, College students' organization of mathematics word problems in relation to success in problem solving, School Sci. Math. **89**, 392 (1989).
[31] G. S. Gliner, College students' organization of mathematics word problems in terms of mathematical structure vs. surface structure, School Sci. Math. **91**, 105 (1991).
[32] B. S. Eylon and F. Reif, Effects of knowledge organization on task performance, Cognit. Instr. **1**, 5 (1984).
[33] 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).
[34] A. H. Schoenfeld and D. J. Herrmann, Problem perception and knowledge structure in expert and novice mathematical problem solvers., J. Exp. Psychol. Learn. Mem. Cogn. **8**, 484 (1982).
[35] F. Reif and J. I. Heller, Knowledge structure and problem solving in physics, Educ. Psychol. **17**, 102 (1982).
[36] I. D. Beatty and W. J. Gerace, Probing physics students' conceptual knowledge structures through term association, Am. J. Phys. **70**, 750 (2002).
[37] G. A Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information., Psychol. Rev. **63**, 81 (1956).
[38] J. Clement, Students' preconceptions in introductory mechanics, Am. J. Phys. **50**, 66 (1982).
[39] L. C. McDermott, Research on conceptual understanding in mechanics, Phys. Today **37**, 24 (1984).
[40] 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).
[41] E. Etkina, J. Mestre, and A. O'Donnell, *The cognitive revolution in educational psychology*, (Information Age Publishing, Greenwich, CT, 2005), pp. 119–164.
[42] National Research Council, *How People Learn: Brain, Mind, Experience, and School: Expanded edition* (The National Academies Press, Washington, DC, 2000).
[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] Stellan 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] 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).
[57] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).
[58] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).
[59] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
[60] Physport, https://www.physport.org. Accessed 8/8/2017.
[61] US News & World Report: Education, US News and World Report, Washington, DC, https://premium.usnews.com/best-colleges. Accessed 4/30/2017.
[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] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach* (Springer-Verlag, New York, NY, 2003).
[64] McElreath, *Statistical Rethinking: A Baysian Course with Examples in R and Stan* (CRC Press, Taylor & Francis Group, Boca Raton, FL, 2016).
[65] A. E. Raftery, Bayesian model selection in social research, Sociol. Methodol. **25**, 111 (1995).
[66] L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Struct. Eq. Modeling **6**, 1 (1999).
[67] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).
[68] B. T West, K. B. Welch, and A. T. Gatecki, *Linear Mixed Models: A Practical Guide to Using Statistical Software*, 2nd ed. (CRC Press, Francis & Taylor Group, Boca Raton, FL, 2015).
[69] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.14.010137 for traditional factor analysis, 3 and 5 factor MIRT factor models, and the constrained MIRT model with the factor loading on all items.
[70] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013).

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

[72] 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).

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

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

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

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