Identifying a preinstruction to postinstruction factor model for the Force Concept Inventory within a multitrait item response theory framework

Philip Eaton^{*} and Shannon Willoughby[✉]

Department of Physics, Montana State University, 1325-1399 S 6th Avenue,
Bozeman, Montana 59715, USA

(Received 14 October 2019; published 28 January 2020)

As targeted, single-conception curriculum research becomes more prevalent in physics education research (PER), the need for a more sophisticated statistical understanding of the conceptual surveys used becomes apparent. Previously, the factor structure of the Force Concept Inventory (FCI) was examined using exploratory factor analysis (EFA) and exploratory multitrait item response theory (EMIRT). The resulting exploratory driven factor model and two expert-proposed models were tested using confirmatory factor analysis (CFA) and were found to be adequate representations of the FCI's factor structure for postinstruction student responses. This study compliments the literature by using confirmatory multitrait item response theory (CMIRT) on matched preinstruction to postinstruction student responses ($N = 19745$). It was found that the proposed models did not fit the data within a CMIRT framework. To determine why these models did not fit the data, an EMIRT was performed to find suggested modifications to the proposed models. This resulted in a bi-factor structure for the FCI. After calculating the student ability scores for each trait in the modified factor models, only two traits in each of the models were found to exhibit student gains independent of the general, “Newtonian-ness” trait. These traits measure students’ understanding of Newton’s third law and their ability to identify forces acting on objects.

DOI: [10.1103/PhysRevPhysEducRes.16.010106](https://doi.org/10.1103/PhysRevPhysEducRes.16.010106)

I. INTRODUCTION

Concept inventories have become some of the most widely used tools in physics education research (PER) since their introduction. One of the first successful conceptual inventories created was the Force Concept Inventory (FCI), which was constructed in 1992 and updated in 1995 [1]. Because of the success of this instrument many similar assessments have been created to probe conceptual topics in physics, such as the Force and Motion Conceptual Evaluation [2], the Energy and Momentum Conceptual Survey [3], the Conceptual Survey of Electricity and Magnetism [4], the Brief Electricity and Magnetism Assessment [5], and the Thermal Concept Evaluation [6], among others. Many studies have tested these assessments in an attempt to understand the psychometric properties of these tools, see Refs. [2–11].

Within PER, conceptual assessments, such as the FCI, are often used to identify effective research-based teaching strategies and curricula. For example, a significant portion of the evidence used to support the utilization of active learning techniques comes from student preinstruction to postinstruction gains on validated conceptual assessments [12–14]. The assessments used in these types of studies must be well understood to lend validity and strength to the conclusions derived from their statistics.

The psychometric analyses performed thus far on the FCI can be effectively broken up into two types: unidimensional analysis and multidimensional analysis. The unidimensional analyses performed on the FCI include, but are not limited to, item response curves [15,16], (unidimensional) item response theory (IRT) [17–19], and classical test theory (CTT) [12–14,19–21]; multidimensional studies include, but are not limited to, exploratory factor analysis (EFA) [22–24], confirmatory factor analysis (CFA) [25], multitrait (multidimensional) item response theory (MIRT) [26,27], cluster analysis [28,29], and network analysis [30,31]. These studies have been summarized in Table I.

Previously, we used CFA to test the factor structure of the FCI [25]. This analysis was driven in part by the “H&H debate” [32–34], which discussed the validity and stability of the FCI’s factor structure. This study found that the

^{*}philip.eaton@montana.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

TABLE I. Analyses that have been performed on the FCI. This list is by no means complete.

Unidimensional analyses	
Item response curves	[15,16]
Item response theory	[17–19]
Classical test theory	[12–14,19–21]
Multidimensional analyses	
Exploratory factor analysis	[22–24]
Confirmatory factor analysis	[25]
Multitrait item response theory	[26,27]
Cluster analysis	[28,29]
Network analysis	[30,31]

original factor structure proposed by the creators of the FCI did not fit student response data from the assessment. However, the creator’s model was found to fit the data adequately after adjustments, obtained through CFA modification indexes, were made to the creator-proposed model [25]. The CFA study also investigated two other factor models, one expert proposed and another that was driven by the EFA of the FCI presented by Scott and Schumayer in Ref. [22], and found them both to adequately explain the student response data for the FCI.

This study intends to extend the claims of our previous CFA work by using MIRT. It should be noted that MIRT has been used to study the FCI previously, see Refs. [26,27]. Scott and Schumayer [26] used MIRT to verify an exploratory factor model they found using exploratory factor analysis in Ref. [22]. This model is used in this study as one of the models being tested. Stewart *et al.* [27] presented a theoretically driven MIRT model that attempted to model the fragmented thought process of students while taking the FCI. However, only parsimonious fit statistics were reported for this model. To help identify if the model had good fit with the data a standardized root mean square of the residuals (SRMSR) should have, at minimum, been presented. We used Stewart *et al.*’s model and fit it to the data gathered for this study and found that there was an inadequate goodness of fit with the data, with a $SRMSR > 0.1$. Although the model presented by Stewart *et al.* has a strong theoretical basis, it does not adequately represent data from the FCI used in this study.

The purpose of this study is twofold: (i) to strengthen the CFA results of Ref. [25] using MIRT to supply a deeper understanding of the complex factor structure of the FCI, and (ii) to identify the concepts the FCI can probe independent of other factors in the factor model using a MIRT framework. This will be done by answering the following research questions:

RQ1: Do the expert-proposed models tested in Ref. [25] adequately explain the data in a confirmatory multitrait item response theory framework?

RQ2: Using exploratory multitrait item response theory as a guide, can the expert-proposed models from Ref. [25]

be modified to better fit the data, and what do these modifications suggest about the FCI?

RQ3: Provided an adequate confirmatory model is found, which of the models’ factors demonstrate measurable student gain from preinstruction to postinstruction independent of the other factors?

The article is organized as follows: Sec. II discusses the characteristics of the data used in this study as well as how the data was prepared for analysis. Next, a brief explanation to motivate the need for a multitrait model is presented in Sec. III, followed by the methodology in Sec. IV. The results of the methodology and a subsequent discussion can be found in Secs. V and VI. Lastly, Secs. VII, VIII, and IX address the limitations of, future works related to, and the conclusions of this study, respectively.

II. DATA

The data used in this study were received from PhysPort [35]. PhysPort is a support service that synthesizes physics education research and details effective curricula for instructors. Further, PhysPort collects and manages a plethora of PER assessments and maintains a database of student responses for many of these assessments.

The original PhysPort data used in this study consisted of 22 029 matched preinstruction-to-postinstruction student responses for the 1995 version of the FCI. The sample was made up of student responses from a mixture of algebra- and calculus-based introductory physics I courses and a mixture of teaching strategies (e.g., traditional lecturing, flipped classroom, etc.). Statistical results from a dataset of this kind will thus pertain to the more general population of “students in introductory physics I.” This means the effects of variables such as “mathematical sophistication of the course” and “teaching strategies” cannot be assessed using this data (among others, like student’s gender, student’s academic year taking the course, the semester or quarter the course was taken, etc.).

Any student responses that contained a blank response were removed from the sample in a listwise deletion, which left 19 745 student responses in the sample. These 19 745 student responses were then split in into two halves using their even and odd identification numbers given in the original PhysPort data. The odd half of the data, 9841 student responses, was used to generate the exploratory models, which were to be used to suggest possible modifications to the expert models. The even half of the data, 9904 student responses, was used to test the fit of the exploratory model to a different, independent sample to supply evidence for the possible generalizability of the results. Further, both the odd and even halves of the data were used to assess the goodness of fit of the original and modified expert-proposed models. It should be noted that the data used in this study and in Ref. [25] are the same, with the exception of the removal of students who did not complete both the preinstruction and postinstruction FCI.

III. MOTIVATION FOR A MULTITRAIT ANALYSIS

Unidimensional models inherently assume that the assessment being modeled only measures a single trait, or student ability. Models of this nature enable researchers and instructors to readily compare students' performances to one another and to compare a single student's performance on an assessment from preinstruction to postinstruction through the use of one statistic (i.e., aggregate total score or a unidimensional IRT student ability score).

Assuming an assessment is measuring a single trait may oversimplify, and potentially misrepresent, what students specifically learned from preinstruction to postinstruction. A pre-to-post change in a unidimensional total score will indicate that the students learned something relevant to the assessment, but does not specify which concepts may have been learned. For instance, Fig. 1 shows the FCI scores for 9841 matched preinstruction to postinstruction student responses using the aggregate total score (total number of questions correct) and the student ability scores from a one-dimensional two-parameter logistic model from IRT. These results indicate that student abilities from IRT often do not add more information than the aggregate total score. A linear regression between estimated student abilities and total scores for the FCI found a regression correlation of $r^2 = 0.994$ [18]. This suggests that the total score and student ability on the FCI give effectively the same information. It can be seen that both methods of scoring the assessment indicate that the students did improve over the course of instruction. However, these differences are blind as to which conceptions the students' specifically learned on the assessment.

Conclusions based only on these models could be misleading depending on the research questions being asked. For example, an assessment could be built to probe Newton's first, second, and third laws independently from one another; an assessment like this would be said to have scales representing the three laws of Newton. Suppose this assessment was scored using a unidimensional model for preinstruction and postinstruction results to investigate the effectiveness of new curricula which specifically targets Newton's first law. Further, suppose that the students did not effectively learn Newton's first law, but did have "strong" gains along the Newton's second and third law scales. The results from the unidimensional scoring would be blind to this specific situation, and would conclude that the new curricula did effectively teach Newton's first law, even though it actually did not. Although this example is extreme, it succinctly demonstrates the crucial need for physics education researchers to understand the factors present on the assessments being used in practice.

To properly measure student growth on an assessment that claims to measure multiple conceptions, a multidimensional (or multitrait) model should be used. There are many ways to identify and generate multidimensional

FIG. 1. Histograms of student scores preinstruction (red) and postinstruction (blue). The top plot used the traditional total score for students found by adding up their total number of correct responses, and the bottom plot used the estimated student abilities from a one-dimensional two-parameter logistic model in IRT. Both methods show a flattening to the distribution of scores, and a definite shift to higher scores, which indicates the students performed better on the assessment postinstruction.

models for assessments, like factor analysis and multitrait item response theory (MIRT). We chose to use MIRT since it is a model driven analysis that readily assigns scores to students in the form of an estimated student ability vector. A brief explanation of MIRT and how the models used in this study were created will be discussed in the following section.

IV. METHODOLOGY

The following discussions were kept short for brevity purposes; references cited within each subsection can be used to find more detailed information about the tools and techniques used in this study.

In short, the expert proposed models from Ref. [25] (see Tables II and III) were tested using confirmatory MIRT

TABLE II. The qualitative interpretations for each of the factors for the bi-factor expert-proposed models used in the CMIRT analysis.

Hestenes, Wells, and Swackhamer 6-factor model (HWS6)	
F1:	Newton's first law
F2:	Newton's second law
F3:	Kinematics
F4:	Forces
F5:	Newton's third law
F6:	Superposition of forces
Eaton and Willoughby 5-factor model (EW5)	
F1:	Newton's first law with kinematics
F2:	Newton's second law with kinematics
F3:	Identification of forces
F4:	Newton's third law
F5:	Superposition of forces and mixed methods
Scott and Schumayer 5-factor model (SS5)	
F1:	Newton's first law with zero netforce
F2:	Newton's second law with kinematics
F3:	Identification of forces
F4:	Newton's third law
F5:	Newton's first law with canceling forces

TABLE III. The original expert-proposed models tested using CFA in Ref. [25].

Hestenes, Wells, and Swackhamer 6-factor model (HWS6)						
F1	F2	F3	F4	F5	F6	
6	9	12	1	13	4	17
7	22	14	2	18	15	25
8	26	19	3	25	16	26
10	27	20	5	30	28	
23		21	11			
24						
Eaton and Willoughby 5-factor model (EW5)						
F1	F2	F3	F4	F5		
6	20	9	21	5	4	17
7	23	12	22	11	15	25
8	24	14	27	13	16	26
10		19		18	28	
				30		
Scott and Schumayer 5-factor model (SS5)						
F1	F2	F3	F4	F5		
6	12	19	22	5	4	16
7	16	20	23	11	15	17
8	24	21	27	13	28	25
10	29			18		
				30		

(CMIRT). Following this, exploratory MIRT (EMIRT) was performed on the odd half of the data to find suggestions for how the expert models could be improved if necessary. Subsequently, CMIRT was used again to reexamine the modified expert models to verify good model fit with both the odd and even halves of the data. Once good model fit was verified, the students' abilities were estimated and plotted on radar plots, discussed in Sec. IV E.

A. Multitrait item response theory

Item response theory postulates that the manner in which a student responds to an item is a consequence of the interaction between the student and item characteristics. MIRT further postulates that a single assessment can probe multiple traits (i.e., latent abilities). As a result, students and items will have different characteristics for each trait in the multitrait model of the assessment. For example, each item will have different characteristics and students will have different ability scores for each trait in the model. This enables researchers and instructors to determine which traits a student is excelling in and traits a student is in need of assistance.

The characteristics of an item are specified within a mathematical model called an item response function. An example of an item response function is the multitrait 2-parameter logistic (2PL) function [36], which can be written in the following manner:

$$P_{ij}(x_{ij} = 1 | \theta_i, \alpha_j, \delta_j) = \frac{1}{1 + e^{-D(\alpha_j \theta_i - \delta_j)}},$$

where P_{ij} is the probability that student i , with an ability vector θ_i , will answer item j correctly given the item's discrimination vector and difficulty, α_j and δ_j . The constant D is generally taken to be 1.702 to fix the scale of the logistic model such that $\theta = 1$ approximately corresponds to 1 standard deviation of the sample's student ability scores along a single trait. To fully describe the interaction of student i and item j along trait k one needs to know the student's trait ability θ_{ik} , the item's trait discrimination α_{jk} , and the item's difficulty δ_j . Since items are allowed to cross load onto multiple traits, each item may have multiple trait discriminations, one for each trait the item loads onto. The interpretation of the item parameters is not straightforward, but becomes easier when considering a single trait (i.e., setting all other trait abilities to zero).

For instance, the trait discrimination of item j along trait k , α_{jk} , describes the slope of the item response function at a probability of 50% along the axis of trait k . The larger α_{jk} is the more discriminating the trait is for that item, or the better the item is at differentiating between students with abilities above or below the difficulty of the item within that trait. So, students will generally need smaller gains along traits on which items have large discrimination values to significantly improve their odds of answering the

question correct, and vice versa. Typically, a single trait will contain a group of items that have large discrimination values compared to the other items on the assessment. These groups of items can be treated in a similar manner as the factors found through exploratory factor analysis [37].

Notice the item difficulty is the same regardless of the trait being considered. A trait difficulty for item j along trait k can be expressed as δ_j/α_{jk} , and can be shown to be the location along that trait's ability axis (all other trait abilities set to zero), where the probability of getting the question correct is 50%. So, the larger the discrimination an item has along a trait, the easier the item appears and vice versa. This is consistent with the idea that the better the item discrimination is along a trait, the less a student needs to improve that trait to significantly improve their odds of getting the item correct.

A more detailed discussion of MIRT can be found in the many resources dedicated to the subject. These details, as well as the parameter estimations techniques used to find the item characteristics and student ability scores, are out of the scope of this paper; more technical details can be found in the literature [26,27,37–41]. This study used maximum likelihood parameter estimation techniques using the open-source software package “mirt” [42] based in the programming language R [43].

B. Linking multitrait item response theory models

Because of the dependency on the data to set the metric in which the item parameters are estimated, one will need to link the metrics of models when trying to compare results from multiple sets of data. Within a unidimensional IRT model this linking appears in the form of a rescaling and a translation of the student ability metric. This becomes more complicated in a multitrait model where rescaling and translations need to be done along each of the traits, after rotating the ability axes to ensure the models are measuring the same traits. This can be done using a number of different methods, the details of which can be found in Refs. [44–48].

For this study, we decided to use a mean anchored (or mean anchoring) linking procedure. This was done by specifying the groups of questions within the MIRT model through prescribing the factor structure of the assessment, which is known as confirmatory multitrait item response theory. With the factor model in place, the preinstruction and postinstruction samples were used to find the item parameters for the specified model. The resulting item difficulty values were shifted from student centered to item centered by subtracting the mean item difficulty from all of the difficulty parameters for each sample. Finally, the pretest and post-test item parameters were averaged and these mean values were used to estimate the student abilities for both samples; thus the name: mean anchored. Goodness-of-fit statistics for the parameters generated

using this linking procedure were found to be acceptable and are reported in the Sec. V when discussing the results of each of the models used in this study.

FIG. 2. Preinstruction and postinstruction models' original estimated item parameters (y axis) plotted vs the combined mean anchored item parameters (x axis). The red, dashed line is the 1-to-1 line, a line of slope 1 and y-intercept 0. The points are, for the most part, well regressed onto the 1-to-1 line, meaning the item parameters are in good agreement with one another. For these plots a1–a6 represent the item discriminations for the 6 traits in the model and d stands for the item difficulty. Specifically, these plots are for the EW5 model, see Table VI.

After a scale linking is performed, it is important to check the quality of the linking. This can be done by plotting the item parameters (i.e., the trait discriminations and difficulties) of the original scale to the linked scale [48]. For example, Fig. 2 displays the relationship between the preinstruction, postinstruction, and the mean item parameters for the Eaton and Willoughby 5-factor model [25]. From these plots it can be seen that the item parameters are in good agreement with one another, with minor exceptions. Similar plots for the other models were generated, but did not differ much from these results. So, these plots were left out of the article for brevity. These plots, in conjunction with the acceptable goodness-of-fit statistics that result from models treated in this manner, indicates that the utilization of the mean anchoring procedure is acceptable.

C. Exploratory multitrait item response theory

One way to identify potential factor models that could be used in a CMIRT analysis is to explore the factor structure of the sample using an exploratory analysis. This was done by generating exploratory models ranging from 3 factors up to 7 factors. To ensure the resulting models would be expertlike in nature, postinstruction student responses were used to initially establish each model's factor structure.

The initial exploratory models allowed all of the questions to load onto all of the available factors. From here, questions were systematically removed from individual factors to create a more parsimonious factor structure. This was done by removing the smallest calculated item loading from the model one by one, while monitoring the model's goodness of fit. That is, the loading parameter that corresponded to the smallest loading value in the factor model was removed for a single iteration of this process.

Item loading values explain the amount of variance a factor explains for a particular item; a factor represents a trait in both factor analysis and MIRT. The closer the magnitude of an item loading is to 1, the more variance of the item is explained by the factor. As a result, a loading value is interpreted as being small if its magnitude is close to zero.

Beginning with item loading values with magnitudes below 0.05 for postinstruction data, items were removed from individual traits one at a time. After an item was removed, the new model was fit to the data and the goodness-of-fit statistics were calculated. The goodness-of-fit statistics were monitored to ensure that the fit of the model never became inadequate. If all of the items with loading values of magnitude less than 0.05 on the post-instruction data were removed, then the process was redone using the preinstruction data. This back and forth was continued, incrementing by 0.05 in the magnitude of the loading values, until the goodness-of-fit statistics for the model became unacceptable.

TABLE IV. The 7-factor exploratory model generated through the iterative item removal process described in Sec. IV C. The related goodness-of-fit statistics for this model can be found in Table IX for the exploratory and confirmatory samples of data.

F1	F2	F3	F4	F5	F6	F7
5	4	14	13	6	17	
11	15	21	29	8	25	
13	16	22	30	10	26	All questions
18	28	23		17		
30		27		23		
				24		

This procedure resulted in the most parsimonious models possible while still having acceptable goodness-of-fit indices. An example of a model generated from this procedure can be found in Table IV, whose details will be discussed in the results section (see Sec. V B).

The goodness-of-fit statistics used in this study can be loosely placed into three categories: (i) parsimonious fit indexes, (ii) absolute fit indexes, and (iii) comparative fit indexes. The following is a brief discussion of some of these statistics, and a more detailed explanation can be found in Ref. [25].

Generally, as more parameters are added to a model the fit of the model will become better. Parsimonious fit indexes add a penalty to the goodness of fit in a manner related to the number of parameters used in the model to correct for a model having more parameters and another. Some indexes of this type are the root mean square error of approximation (RMSEA) and the M2 statistic [49–51]. The RMSEA is bounded on bottom by 0 and it is unbounded above. If the top of the 95% confidence interval of the RMSEA is below 0.06, then the model is said to have an acceptable fit [52]. The M2 statistic could be used to indicate acceptable model fit, but was instead used for parsimonious model selection. If two models have the same absolute and comparative goodness-of-fit values, then the model with the smaller M2 statistic is the preferred model as it is the most parsimonious between the two.

Absolute fit indexes are fit statistics that do not need to reference another model to be calculated. For example, the standardized root mean square of the residual (SRMSR) is an absolute fit index since it uses only the residuals of the model to determine goodness of fit. The SRMSR is bounded between 0 and 1, where a value of 0 indicates that all of the residuals were zero, and thus good fit. A model is said to have an acceptable fit if its SRMSR is below 0.08 [52,53].

Lastly, comparative fit indexes are fit values calculated by comparing the fit of the proposed model to a baseline model. A baseline model is one which makes no assumptions about the structure of the assessment and takes all of the parameters to be uncorrelated. The comparative fit index (CFI) and the Tucker-Lewis index (TLI) are two

commonly used comparative fit indices. Both statistics are bounded on the bottom by 0 and on the top by 1. Values closer to 1 indicate good model fit, and values closer to 0 indicate that the proposed model and the baseline model fit the data with similar accuracies. Acceptable values for these statistics are above 0.90 or 0.95 depending on the desired model-fit strictness [52,53].

D. Confirmatory multitrait item response theory

Confirmatory multitrait item response theory (CMIRT) is similar to the exploratory analysis described previously, the primary difference is that the models investigated in CMIRT are expert-proposed factors rather than data derived factors. Expert-proposed models are generated through the careful consideration of the conceptual content of the questions on the assessment and may be partially, or wholly, guided by exploratory results. These expert-proposed models are assumed to encapsulate the conceptual content within the assessment in a better manner than exploratory driven models because they are generated by experts of the field.

The models tested in this study were previously studied by us in Ref. [25] using CFA; see Tables II and III for descriptions of the models. In total 3 factor models for the FCI were investigated, 2 expert-proposed models and 1 model which resulted from exploratory factor analysis, and were found to adequately represent the factor structure of the FCI [25]. The exploratory model (SS5) was generated by Scott and Schumayer in 2012 in which they used exploratory factor analysis on 2150 postinstruction student responses from an algebra-based introductory physics I course [22]. They then reconfirmed the existence of this structure by performing an EMIRT on the same set of data [26]. Another of the expert models analyzed was a slightly modified version of the factor model proposed by the creators of the FCI (HWS6) [1], and the last model was created using a combination of the creator's model and the EFA model from Scott and Schumayer (EW5) [25]. The factor structure and the qualitative meaning for each of these model's factors can be found in Tables II and III, respectively. A thorough discussion and justification of these models is left to Refs. [1,22,25].

To perform CMIRT, one begins by specifying the expert model and then fits it against data from students who took the assessment being investigated. Once the model has been fit to the data, goodness-of-fit statistics of the model can be generated. All of this was done using a maximum likelihood estimator from the “mirt” package in R [42].

When analyzing the expert-proposed models, the covariances between factors were set to zero to make the student ability scores easier to interpret and to make the models as parsimonious as possible. Although, one would expect the concepts measured by the FCI to be related to one another, this study seeks factors that are independent from one another. To model this, the factors were taken to be

uncorrelated within the exploratory and confirmatory analysis. This interpretation of factor independence is used to analyze the results of the modified expert models in Sec. VI.

E. Student ability radar plots

Once the modified expert-proposed models were confirmed to have a good fit with the preinstruction and postinstruction data, the linking procedure described in Sec. IV B was used and the student abilities were estimated. This enabled the student ability scores to be estimated on the same ability scale for preinstruction and postinstruction, which allowed for direct comparisons of the students' before and after instruction performances on the FCI.

The resulting student abilities were plotted in radar plots using the “fmsb” package in R [54]. A radar plot is a graphical technique that allows for multivariate results to be displayed in a two-dimensional plot. For the plots presented in this study, the outermost dotted line corresponds to a θ of 2 and the innermost dotted line to -2 . The student abilities of the sample for each trait are plotted along a spoke of the radar plot, with each spoke representing a single trait. An example of a preinstruction and postinstruction student abilities radar plot for the Eaton and Willoughby 5-factor model can be found in Fig. 3, where preinstruction results

FIG. 3. Preinstruction (red) and postinstruction (blue) student ability estimations for the EW5 model. The short and long dashed lines represent the preinstruction and postinstruction medians, respectively. The shaded regions represent the first and third quartiles of the students' postinstruction ability values for each of the traits. Similarly for the hashed regions and the preinstruction results. The outer- and innermost black dotted lines indicate $\theta = 2$ and $\theta = -2$, respectively. Notice only factor F3, F4, and F6 show growth from the pretest to post-test. A description of the factors for the EW5 model can be found in Tables II and VI.

TABLE V. Fit statistics for the expert-proposed factor models using the exploratory and confirmatory halves of the student response data. Since all three models fit both halves of the data equally well, the smallest M2 fit statistic (bold type) is used to identify the preferred model.

Exploratory half of the data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
EW5 Pre	9548.885	0.0446	0.04378	0.0453	0.042 06	0.9543	0.9512
EW5 Post	9225.783	0.043 76	0.0430	0.0445	0.0387	0.9596	0.9568
HWS6 Pre	10 390.98	0.0464	0.0457	0.0472	0.0443	0.9501	0.9467
HWS6 Post	11 346.57	0.0486	0.0478	0.0494	0.0391	0.9485	0.9449
SS5 Pre	111 870.03	0.0499	0.0491	0.0507	0.0427	0.9474	0.9438
SS5 Post	12 034.5891	0.0503	0.0495	0.0511	0.0427	0.9467	0.943
Confirmatory half of the data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
EW5 Pre	10 238.8	0.0460	0.0453	0.0468	0.0428	0.9509	0.9475
EW5 Post	8869.934	0.0427	0.0419	0.0435	0.0395	0.9602	0.9574
HWS6 Pre	10 390.98	0.0464	0.0457	0.0472	0.0443	0.9501	0.9467
HWS6 Post	11 346.57	0.0486	0.0478	0.0494	0.0391	0.9485	0.9449
SS5 Pre	13 035.23	0.0522	0.0515	0.0530	0.0498	0.9368	0.9325
SS5 Post	11 263.53	0.0484	0.0477	0.0492	0.0438	0.9489	0.9453

are displayed in red and postinstruction results are in blue. From the plots the traits on which the students experienced explanatory growth over the course of instruction become readily identifiable.

V. RESULTS

First the initial CMIRT results are discussed, followed by the subsequent EMIRT results. Lastly, the results for the CMIRT and the corresponding student ability radar plots are discussed.

A. Initial confirmatory multitrait item response theory results

The unmodified expert-proposed factor models generated unacceptable model-fit statistics, with TLI and CFI values below 0.9, within the CMIRT frame work. This suggests something is missing from these models, which is causing them to inadequately model the students' response vectors. Thus, an EMIRT was performed to identify potential modifications that could be made to the expert models to increase their fit to the data without unintentionally adding unmotivated or unjustified features.

B. Exploratory multitrait item response theory

The factor structures for each of the resulting 3- through 7-factor exploratory models can be found in Table VIII and their goodness-of-fit statistics can be found in Table IX, both located in the Appendix. From Table IX, it can be seen

TABLE VI. The bi-factor expert-proposed models used in the CMIRT analysis. The related goodness-of-fit statistics for these models can be found in Table V for the exploratory and confirmatory samples of data.

Hestenes, Wells, and Swackhamer 6-factor model (HWS6)							
F1	F2	F3	F4	F5	F6	F7	
6	9	12	1	13	4	17	
7	22	14	2	18	15	25	
8	26	19	3	25	16	26	All questions
10	27	20	5	30	28		
23		21	11				
24							
Eaton and Willoughby 5-factor model (EW5)							
F1	F2	F3	F4	F5	F6		
6	20	9	21	5	4	17	
7	23	12	22	11	15	25	
8	24	14	27	13	16	26	All questions
10		19		18	28		
				30			
Scott and Schumayer 5-factor model (SS5)							
F1	F2	F3	F4	F5	F6		
6	12	19	22	5	4	16	
7	16	20	23	11	15	17	
8	24	21	27	13	28	25	All questions
10	29			18			
				30			

that the models using 3–7 factors had acceptable model-fit values with the odd and even halves of the data. This suggests that the models that were generated from exploratory work are reliable and thus may be effectively modeling the population which the sample was drawn from. From the same table it can be seen that the models using 3–6 factor all have quite similar goodness-of-fit statistics. This is due to the manner in which the exploratory models were generated. Recall, the removal of items from traits was stopped only after the model-fit values became unacceptable. For these models, the SRMSR was the first fit value to become unacceptable.

The utilization of 7 factors resulted in a model that did not have any issues with unacceptable fit statistics. The reason the iterative removal of items from factors was stopped for this model was to avoid factors with less than 3 items, which is not recommended for factor models [52]. From Table VIII it can be seen that the 7-factor model is more simplistic compared to the other exploratory models. Since the reduction of this model was not stopped due to fit

concerns, the fit of this model passes even the strictest fit criteria for model goodness of fit.

The factor structure of the 7-factor exploratory model is actually a bi-factor model in that it contains a single, all-questions-included factor (sometimes called the general factor) with subfactor structure overlaid [55–57]. Coincidentally, the subfactor structure for the 7-factor model is almost identical to the proposed expert models from Ref. [25]. Comparing the composition of the factors for the 7-factor model and the expert models the following traits readily emerge:

- (i) F1—Identification of forces,
- (ii) F2—Newton’s third law,
- (iii) F3—Newton’s second law with kinematics,
- (iv) F5—Newton’s first law with kinematics, and
- (v) F6—Superposition of forces or mixed methods.

The initial poor fit of the expert-proposed models and their similarities to the subfactors of the 7-factor model suggests that a general trait is necessary to properly model the FCI within a MIRT framework.

TABLE VII. Preinstruction (red) and postinstruction (blue) student ability estimations for the expert proposed models. The short and long dashed lines represent the preinstruction and postinstruction medians, respectively. The shaded regions represent the first and third quartiles of the students’ postinstruction ability values for each of the traits. Similarly for the hashed regions and the preinstruction results. The outer- and innermost black dotted lines indicate $\theta = 2$ and $\theta = -2$, respectively. The factor compositions for these models can be found in Tables II and VI.

In short, a bi-factor model contains a general trait that is measured by all of the items on the assessment. Subtraits, or subfactors, are then specified on top of the general factor to account for residual variance shared by subsets of items on the assessment. A bi-factor model can appear when the assessment contains an overall scale (due to nonorthogonal factors) and orthogonal subscales [58,59]. Thus, it makes sense that the exploratory analysis of the FCI suggests that expert models should incorporate a general trait since the factors on the FCI have been found to be nonorthogonal in nature [22].

C. Confirmatory multitrait item response theory

Motivated by the results of the EMIRT analysis, the expert models were updated by adding a general trait, see Table VI. This adjustment resulted in acceptable goodness of fit for all three of the expert-proposed models with TLIs and CFIs all above 0.93 and RMSEAs and SRMSRs all below 0.06, see Table V.

VI. DISCUSSION

Research Question 1: Do the expert-proposed models tested in Ref. [25] adequately explain the data in a confirmatory multitrait item response theory framework?

The initial expert-proposed models used in Ref. [25] did not have acceptable goodness-of-fit statistics with the data in a CMIRT framework (TLIs and CFIs below 0.9).

Research Question 2: Using exploratory multitrait item response theory as a guide, can the expert-proposed models from Ref. [25] be modified to better fit the data, and what do these modifications suggest about the FCI?

The initially poor model fits and the results of the EMIRT analysis, suggest that the FCI may be best represented by a bi-factor model, when working in a MIRT statistical framework. For the FCI, the unidimensional scale represents the students' overall "Newtonian-ness" and the subscales represent the specific concepts probed by the instrument (Newton's third law, identification of forces, etc.). In conclusion, the expert models perform well if a general trait is included in the models.

Research Question 3: Provided an adequate confirmatory model is found, which of the models' factors demonstrate measurable student gain from preinstruction to postinstruction independent of the other factors?

Over the course of instruction, each of the expert model's preinstruction to postinstruction student ability radar plots show an improved student Newtonian-ness. This can be seen in Table VII, where there is outward motion on the general factor for each of the models (the last factor in each model). This implies that some of the students' improvement on the items is explained by a general increase in their overall Newtonian-ness. It should be noted that while attempting to model the FCI using just a unidimensional MIRT model acceptable goodness-of-fit statistics could not

be obtained. This suggests the need for a multidimensional model to help explain some of the remaining variance and thus improve the model-fit statistics.

Improved student ability scores along the unidimensional trait is expected since the students' aggregate total scores over the course of instruction improved on average. However, these results do not include student growth with respect to any of the specific concepts probed. Student growth along any of the other factors, which represent independent concepts, is thus information gained independent of what a unidimensional model would explain. For example, in Table VII, the models all have the added general trait in the last factor of the radar plot. Thus, the growth along F3 and F4 in the EW5 model, F4 and F5 in the HWS6 model, and F3 and F4 in the SS5 model is information not captured by the unidimensional factors in each of the models. Traits with this kind of characteristic are defined as usable scales within the MIRT framework.

Regarding the subfactor structure, measurable gains were detected using the present data for students on the same two conceptual traits for all of the expert models, see Table VII in conjunction with Table II. These traits pertain to Newton's third law and identification of forces. This implies that student growth along these conceptual ideas could not be fully modeled using a unidimensional model. Although the other factors did explain some of the variance in the assessment, they did not help to explain what the students learned over the course of instruction independent from a general ability trait.

As a result of this analysis, the FCI can be thought of primarily as unidimensional assessment with subscales measuring concepts such as Newton's three laws, identification of forces, superposition of forces, and kinematics, depending on the expert model used. Of these subscales, this study identifies only 2 that are usable for identifying concepts students learned independently from the other traits on the FCI: Newton's third law and identification of forces. Recall this idea of independence comes from the fact that the traits were all taken to have a correlation of zero with respect to one another. This implies that the traits can be taken to be independent of each other within the model. So within a MIRT perspective, if a researcher wanted to independently comment on students' gains in Newton's first and second laws and kinematics the FCI could not be used in its current form.

It should be noted that a model was investigated which included only the general trait, and the factors pertaining to Newton's third law and identification of forces, leaving out the other factors which did not exhibit student ability growth. When this model was fit against the student response data using CMIRT, unacceptable fit indices were generated. So, the other factors in the models are necessary to adequately fit student data. This suggests that the FCI is a 5 or 6 factor assessment, but only two of these factors were found in this study to exhibit student ability changes independent of the general, Newtonian-ness trait.

Since the results of this analysis come from modeling the FCI, they can be thought of as describing the characteristics of the FCI independent of the students. This is a critical feature of item response theory that item characteristics are independent of the student characteristics. This suggests that after proper modifications are made to the FCI some of the issues detected may be fixed. For example, changes could be made to the questions measuring Newton's first and second laws to make them more independent in an effort to separate these concepts from one another. This could result in better resolution of these factors, and thus allow for detectable student ability growth along these factors independent of the general trait.

VII. LIMITATIONS

The results presented in this study were for a general population of introductory physics I students. Results based on more "pure" data may be slightly different and further research to check these results is suggested to strengthen the claims and utility of the results presented in this paper.

Since listwise deletion was used to clean the data used in this study, some bias may have been introduced into the results. Listwise deletion was used for this study for two reasons: (i) This study used the same original pool of data as Ref. [25], which used listwise deletion. To enable easier comparisons between the results of these studies listwise deletion was also used here. (ii) It has been found that regression results remain robust and relatively unbiased when using listwise deletion [60]. Since the 2PL model is similar to a logistic regression, the parameter estimations of MIRT should also be accurate. Further, standard errors found for regression models using listwise deleted data are found to still accurately estimate the true standard errors.

There could be other models not examined in this study that could produce a structure in which Newton's first and second laws and kinematics demonstrate student ability growth independent of the general unidimensional factor. However, considering the results of the EMIRT, these other models are unlikely to exist for the current form of the FCI since they would likely be significantly different from the discovered EMIRT models. This implies that these hypothetical models would not agree with the factors generated via student responses to the FCI and will thus not likely reflect an honest representation of the instrument.

VIII. FUTURE WORKS

Future work will be to extend this analysis to other well-used conceptual assessments in PER, like the Force and Motion Conceptual Evaluation [2], the Conceptual Survey of Electricity and Magnetism [4], and the Brief Electricity and Magnetism Assessment [5]. This would help supply evidence for the existence of stable scales for these assessments, which would benefit concept specific interventions within these physical topics (i.e., laboratories or tutorials).

IX. CONCLUSIONS

This study found that the original expert-proposed models for the FCI from Ref. [25] did not adequately fit the data within a MIRT framework. Suggestions for modifications to these models were found through the use of EMIRT, from which a bi-factor structure was discovered. Coincidentally, the 7-factor exploratory model had a striking resemblance to the proposed expert models analyzed in this study, with the notable inclusion of a general, unidimensional factor.

Since the FCI seeks to measure student Newtonian-ness by probing students' understanding of concepts like Newton's third law or Force identification, a bi-factor model is reasonable. The general factor can be assumed to measure the Newtonian-ness of students, while the subfactors of the model measure the individual concepts used to probe the general trait. The validity of such a model was tested by adding a general factor to the expert models. The new expert bi-factor models were found to have acceptable model-fits with the data preinstruction and postinstruction. This suggests that the FCI is likely best modeled as an assessment which measures the general Newtonian-ness of students and then measures subabilities covering the concepts Newton's first law with kinematics, Newton's second law with kinematics, Newton's third law, identification of forces, and superposition of forces.

From preinstruction and postinstruction student ability scores for these subfactors, it was found that only two concepts consistently displayed student growth from independent of the general factor. These concepts were Newton's third law and identification of forces. This implies that within a MIRT framework the FCI may not be sensitive enough to measure the fine-grain growths of all the concepts contained on the assessment.

The lack of independent student ability growth found for some of these concepts is important to understand with respect to future curricula research. For instance, when examining the effectiveness of kinematics curricula or tutorials, using the FCI is not advised since it appears to be insensitive towards measuring independent student growth in kinematics. Researchers should consider choosing a different assessment if they are attempting to measure differences in students' abilities within specific concepts of force and motion, such as kinematics or Newton's second law. This study found that the 1995 version of the FCI could be used to measure students' overall Newtonian-ness, and independently measure students' understanding of Newton's third law and/or identification of forces.

Currently, some PER curricula research is preformed with conceptual or "in-house" instruments whose psychometric properties are not yet well understood. To ensure the validity and reproducibility of such studies, we encourage researchers to only use instruments whose psychometric properties are well understood. To enable single-concept specific research in PER, currently used instruments need to be better vetted statistically and qualitatively. The authors suggest that the future direction of PER involve

the construction of single-conception assessments to avoid the inherent difficulties of creating a many-concept inventory which obeys a prespecified factor structure. Instruments of this nature would allow for more fine-grain curricula studies to be performed with more confidence, precision, and accuracy than may be possible with the current assessments available within the field.

ACKNOWLEDGMENTS

P. E. and S. W. thank Physport for allowing us to use their data. Further, P. E. and S. W. thank Dr. Keith Johnson and David Russell for their insightful discussions and thoughtful comments when reading over the manuscript. This project was funded by the Montana State University Physics Department.

APPENDIX: TABLES

The following are complementary tables for readers interested in the results of the exploratory MIRT analysis. The resulting exploratory factor structures for trials using 3 factors up to 7 factors can be found in Table VIII and the associated pre- and postinstruction fit statistics can be found in Table IX.

TABLE VIII. The 3- through 7-factor EMIRT models. Goodness-of-fit statistics of these models for the exploratory and confirmatory halves of the data can be found in Table IX.

7-factor exploratory model								
F1	F2	F3	F4	F5	F6	F7		
5	4	14	13	6	17	All questions		
11	15	21	29	8	25			
13	16	22	30	10	26			
18	28	23		17				
30		27		23				
				24				
6-factor exploratory model								
F1	F2	F3	F4	F5	F6			
2	2	1	3	1	17	1	11	21
5	4	2	9	2	19	3	12	22
11	15	8	14	3	20	4	13	23
13	16	10	19	8	22	5	14	24
17	28	23	21	9	25	6	15	26
18	29	24	22	10	26	7	16	27
22			23	13	28	8	18	28
25			26	14	30	9	19	30
26			27	16		10	20	
30								
5-factor exploratory model								
F1	F2	F3	F4	F5				
2	18	4	9	2	22	1	11	21
3	20	15	19	4	25	2	12	22
5	22	16	20	13	26	3	13	23
10	25	20		15	28	5	14	24
11	26	28		16	30	6	16	26
13	29	29		17		7	17	27
17	30					8	18	28
						9	19	30
						10	20	
4-factor exploratory model								
F1	F2	F3	F4					
2	18	4	2	22	1	11	20	

(Table continued)

TABLE VIII. (Continued)

4-factor exploratory model							
	F1	F2	F3		F4		
3	20	15	4	25	2	12	21
5	22	16	13	26	3	13	22
8	25	18	15	28	4	14	23
10	26	21	16	30	6	15	24
11	28	22	17		7	16	26
13	29	28			8	17	27
16	30	29			9	18	28
17					10	19	30

3-factor exploratory model							
	F1	F2		F3			
1	12	22	1	16	1	12	23
2	13	23	2	18	2	13	24
3	14	25	4	20	3	14	26
5	16	26	6	28	6	16	27
6	17	28	7	29	7	18	28
8	18	30	15		8	19	29
9	19				9	20	30
10	20				10	21	
11	21				11	22	

TABLE IX. Fit statistics for the exploratory factor models using the exploratory and confirmatory halves of the student response data.

Exploratory half of data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
7-factor exploratory model pre	7602.8436	0.0395	0.0387	0.0403	0.0338	0.9641	0.9616
7-factor exploratory model post	6821.7249	0.0373	0.0365	0.0381	0.0345	0.9707	0.9687
6-factor exploratory model pre	11 473.7269	0.0491	0.0483	0.0498	0.0748	0.9447	0.9408
6-factor exploratory model post	11 760.0204	0.0497	0.0489	0.0505	0.0719	0.9479	0.9443
5-factor exploratory model pre	12 182.2793	0.0506	0.0498	0.0514	0.0794	0.9411	0.937
5-factor exploratory model post	12 539.887	0.0514	0.0506	0.0521	0.076	0.9443	0.9405
4-factor exploratory model pre	10 701.1697	0.0473	0.0465	0.0481	0.0653	0.9485	0.945
4-factor exploratory model post	11 536.8648	0.0492	0.0484	0.05	0.0707	0.949	0.9454
3-factor exploratory model pre	12 728.7833	0.0518	0.051	0.0525	0.0753	0.9383	0.9341
3-factor exploratory model post	16 485.3091	0.0592	0.0584	0.0599	0.0739	0.9262	0.9211

Confirmatory half of data							
Name	M2	RMSEA	RMSEA_5	RMSEA_95	SRMRS	TLI	CFI
7-factor exploratory model pre	8149.188	0.0408	0.0401	0.0416	0.0345	0.9614	0.9587
7-factor exploratory model post	6603.938	0.0365	0.0357	0.0373	0.0355	0.9709	0.9689
6-factor exploratory model pre	11 812.43	0.0496	0.0489	0.0504	0.0739	0.9430	0.9391
6-factor exploratory model post	11 374.96	0.0487	0.0479	0.0494	0.0698	0.9483	0.9448
5-factor exploratory model pre	12 404.81	0.0509	0.0501	0.0517	0.0771	0.9400	0.9359
5-factor exploratory model post	11 978.75	0.0500	0.0492	0.0508	0.0740	0.9455	0.9417
4-factor exploratory model pre	10 921.81	0.0477	0.0469	0.0484	0.0634	0.9475	0.9438
4-factor exploratory model post	11 335.75	0.0486	0.0478	0.0494	0.0686	0.9485	0.9450
3-factor exploratory model pre	12 874.1	0.0519	0.0511	0.0527	0.0734	0.9377	0.9334
3-factor exploratory model post	15 830.35	0.0578	0.0570	0.0585	0.0718	0.9272	0.9222

- [1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
- [2] 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).
- [3] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, *Am. J. Phys.* **71**, 607 (2003).
- [4] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students? Conceptual knowledge of electricity and magnetism, *Am. J. Phys.* **69**, S12 (2001).
- [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] P. Wattanakasiwich, P. Taleab, M. D. Sharma, and I. D. Johnston, Construction and implementation of a conceptual survey in thermodynamics, *Int. J. Innov. Sci. Math. Educ. (formerly CAL-laborate International)* **21**, 29 (2013).
- [7] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010105 (2009).
- [8] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, *Phys. Rev. ST Phys. Educ. Res.* **4**, 020101 (2008).
- [9] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, *Am. J. Phys.* **76**, 882 (2008).
- [10] M. Ishimoto, G. Davenport, and M. C. Wittmann, Use of item response curves of the Force and Motion Conceptual Evaluation to compare Japanese and American students? Views on force and motion, *Phys. Rev. Phys. Educ. Res.* **13**, 020135 (2017).
- [11] P. Eaton, K. Johnson, B. Frank, and S. Willoughby, Classical test theory and item response theory comparison of the Brief Electricity and Magnetism Assessment and the Conceptual Survey of Electricity and Magnetism, *Phys. Rev. Phys. Educ. Res.* **15**, 010102 (2019).
- [12] M. Prince, Does active learning work? A review of the research, *J. Engin. Educ.* **93**, 223 (2004).
- [13] 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).
- [14] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410 (2014).
- [15] 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).
- [16] 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).
- [17] 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).
- [18] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, *Am. J. Phys.* **78**, 1064 (2010).
- [19] 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).
- [20] 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).
- [21] C. Henderson, Common concerns about the Force Concept Inventory, *Phys. Teach.* **40**, 542 (2002).
- [22] 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).
- [23] T. F. Scott and D. Schumayer, Conceptual coherence of non-newtonian worldviews in Force Concept Inventory data, *Phys. Rev. Phys. Educ. Res.* **13**, 010126 (2017).
- [24] 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).
- [25] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010124 (2018).
- [26] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, *Phys. Rev. ST Phys. Educ. Res.* **11**, 020134 (2015).
- [27] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multi-dimensional item response theory and the force concept inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010137 (2018).
- [28] 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).
- [29] 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).
- [30] T. F. Scott and D. Schumayer, Central distractors in Force Concept Inventory data, *Phys. Rev. Phys. Educ. Res.* **14**, 010106 (2018).
- [31] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler, Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis, *Phys. Rev. Phys. Educ. Res.* **15**, 020122 (2019).
- [32] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, *Phys. Teach.* **33**, 138 (1995).
- [33] 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).
- [34] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, *Phys. Teach.* **33**, 503 (1995).
- [35] S. B. McKagan, Physport (2011).

- [36] R. L. McKinley and M. D. Reckase, *An Extension of the Two-Parameter Logistic Model to the Multidimensional Latent Space* (ERIC, Iowa, IA, 1983).
- [37] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (Guilford Publications, New York, 2013).
- [38] S. Monroe and L. Cai, Estimation of a Ramsay-curve item response theory model by the Metropolis–Hastings Robbins–Monro algorithm, *Educ. Psychol. Meas.* **74**, 343 (2014).
- [39] R. J. Patz and B. W. Junker, A straightforward approach to Markov chain Monte Carlo methods for item response models, *J. Educ. Behav. Stat.* **24**, 146 (1999).
- [40] P. Osteen, An introduction to using multidimensional item response theory to assess latent factor structures, *J. Soc. Social Work Res.* **1**, 66 (2010).
- [41] F. B. Baker and S.-H. Kim, *Item response theory: Parameter estimation techniques* (CRC Press, Boca Raton, 2004).
- [42] R. P. Chalmers *et al.* mirt: A multidimensional item response theory package for the R environment, *J. Stat. Softw.* **48**, 1 (2012).
- [43] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2014).
- [44] Y. H. Li and R. W. Lissitz, An evaluation of the accuracy of multidimensional irt linking, *Appl. Psychol. Meas.* **24**, 115 (2000).
- [45] L. Yao and K. Boughton, Multidimensional linking for tests with mixed item types, *J. Educ. Measure.* **46**, 177 (2009).
- [46] W.-C. Lee and J.-C. Ban, A comparison of IRT linking procedures, *Appl. Meas. Educ.* **23**, 23 (2009).
- [47] W. Lin, Q. Jiahe, and L. Yi-Hsuan, Exploring alternative test form linking designs with modified equating sample size and anchor test length, ETS Res. Report Series **2013**, i (2014).
- [48] M. J. Kolen and R. L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices* (Springer Science & Business Media, New York, 2014).
- [49] A. Maydeu-Olivares and H. Joe, Limited information goodness-of-fit testing in multidimensional contingency tables, *Psychometrika* **71**, 713 (2006).
- [50] Y. Liu, W. Tian, and T. Xin, An application of M2 statistic to evaluate the fit of cognitive diagnostic models., *J. Educ. Behav. Stat.* **41**, 3 (2016).
- [51] F. Chen, Y. Liu, T. Xin, and Y. Cui, Applying the M2 statistic to evaluate the fit of diagnostic classification models in the presence of attribute hierarchies, *Front. Psychol.* **9**, 1875 (2018).
- [52] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. (Guilford publications, New York, 2015).
- [53] L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal* **6**, 1 (1999).
- [54] M. Nakazawa, fmsb: Functions for Medical Statistics Book with some Demographic Data, 2018. URL <https://CRAN.R-project.org/package=fmsb>, R package version 0.6.3.
- [55] L. R. Tucker, An inter-battery method of factor analysis, *Psychometrika* **23**, 111 (1958).
- [56] K. G. Jöreskog, A general approach to confirmatory maximum likelihood factor analysis, *Psychometrika* **34**, 183 (1969).
- [57] R. D. Gibbons and D. R. Hedeker, Full-information item bi-factor analysis, *Psychometrika* **57**, 423 (1992).
- [58] C. E. DeMars, A tutorial on interpreting bifactor model scores, *International Journal of Nondestructive Testing* **13**, 354 (2013).
- [59] S. P. Reise, T. M. Moore, and M. G. Haviland, Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores, *J. Personality Assess.* **92**, 544 (2010).
- [60] P. D. Allison, *Missing Data* (Sage Publications, Newbury Park, CA, 2001), Vol. 136.