Confirmatory factor analysis applied to the Force Concept Inventory

Philip Eaton[*] and Shannon D. Willoughby[†]
Montana State University, Bozeman, Montana 59717, USA

(Received 19 January 2018; published 19 April 2018)

In 1995, Huffman and Heller used exploratory factor analysis to draw into question the factors of the Force Concept Inventory (FCI). Since then several papers have been published examining the factors of the FCI on larger sets of student responses and understandable factors were extracted as a result. However, none of these proposed factor models have been verified to not be unique to their original sample through the use of independent sets of data. This paper seeks to confirm the factor models proposed by Scott *et al.* in 2012, and Hestenes *et al.* in 1992, as well as another expert model proposed within this study through the use of confirmatory factor analysis (CFA) and a sample of 20 822 postinstruction student responses to the FCI. Upon application of CFA using the full sample, all three models were found to fit the data with acceptable global fit statistics. However, when CFA was performed using these models on smaller sample sizes the models proposed by Scott *et al.* and Eaton and Willoughby were found to be far more stable than the model proposed by Hestenes *et al.* The goodness of fit of these models to the data suggests that the FCI can be scored on factors that are not unique to a single class. These scores could then be used to comment on how instruction methods effect the performance of students along a single factor and more in-depth analyses of curriculum changes may be possible as a result.

DOI: 10.1103/PhysRevPhysEducRes.14.010124

I. INTRODUCTION

Expertly constructed assessments are used in classrooms to measure conceptual changes of the students or class compared to other students or classes compared to other students and classes. As a result, it is of great importance to understand what the assessments are actually measuring. For instance, from an expert's point of view, the Force Concept Inventory [1] is a test of Newton's laws and the kinematics related to those physical laws. Issues related to what the FCI actually measures have been raised and discussed previously [2–4].

Recent thrusts of research have investigated this topic using techniques such as exploratory factor analysis (EFA) [5–7] and multitrait item response theory [8]. However, none of these papers have confirmed that either the creator's factors nor the factors that they found in their data exists in *other students' responses.* In a series of papers, Huffman and Heller [2,4] claimed that the factors as laid out by the creators in Ref. [1] do not exist in student responses and that the FCI does not measure what has been claimed. In contrast, this paper uses confirmatory factor analysis (opposed to exploratory) to supply evidence that the measurement model proposed in Ref. [1] fits student data in a satisfactory way.

A measurement model, otherwise known as a factor structure, is a description of how items (e.g., questions of an assessment) load onto an associated factor. Exploratory factor analysis is a tool that attempts to extract the best model for the data that groups the correlations among the student responses. EFA does not attempt to confirm or deny the presence of prespecified models. Since most students think about Newtonian mechanics in a mixture of novicelike and expertlike ways, it is not surprising that EFA does not return a completely expertlike measurement model. Because of this, confirmatory factor analysis (CFA) should be used—and not EFA—to test if the models as suggested in Refs. [1] or [5] actually describe the correlations amongst the questions on the FCI.

Using CFA on a sample of 20 822 student responses, this paper offers evidence that the expertlike model proposed by the creators of the FCI actually models the responses of students in a satisfactory way. Further, the EFA driven model found by Scott *et al.* [5] was also [be] tested against this set of data. The data used in Ref. [5] have been further analyzed in Ref. [6] for primary misconceptions students had postinstruction. The potential fit of Scott *et al.*'s model onto this alternate data set could lend a suggestion for the primary misconceptions held by postinstruction students in general. Lastly, another expertlike model (created by Eaton and Willoughby) was to be compared to the model proposed in Ref. [1] to see which one better describes student responses. The research questions this paper seeks

[*]philip.eaton@montana.edu
[†]shannon.willoughby@montana.edu

to answer are how well do the models tested in this study fit a larger sample, and smaller subsamples, and through the use of CFA, what can be concluded about the factor structure of the FCI and the misconceptions of the students as a result?

In Sec. II a brief explanation of how the data were obtained will be detailed, followed by an explanation of confirmatory factor analysis, the meaning of model-fit statistics, and a description of how the stability of the models were tested in this paper. The model specifications can be found in Sec. III. The results of the application of the CFA and the stability analysis are presented and discussed in Sec. IV. The conclusion of the paper follows in Sec. V.

II. METHODOLOGY

A. Data collection

The data used in this study came from the PhysPort database [9]. With IRB approval, 22 028 postinstruction student responses for the 1995 FCI were obtained. The students that make up this data set come from both algebra- and calculus-based classes, and from potentially all levels of active learning classes, from traditional lecture to flipped classes. The data submitted to PhysPort is self-submitted but has many checks that it must go through before it is admitted into the actual database; details on this process can be found in Ref. [9]. After getting the data, it was cleaned by removing all surveys with any blank responses and further by removing any submissions that were all A's, B's, …, E's. After this cleaning process the final data set contained 20 822 student responses.

B. Explanation of CFA

Confirmatory factor analysis is a subdiscipline of a larger latent variable analysis theory known as structural equation modeling. The purpose of CFA is to confirm that a model proposed by a researcher fits, or sufficiently describes the correlational groupings of items within a given data sample. This is fundamentally different than exploratory factor analysis, the purpose of which is identifying the measurement model that best describes a specific set of data. As an example, suppose two different sections of an introductory physics course with the same lecturer, teaching style, semester, etc., were given the FCI. When EFA is applied to the response data for these sections, slightly different factors could be found. The differences in the factor structures between these hypothetical sections would be expected to be small; however, significant differences are not impossible. This can happen *because* EFA is entirely data driven in a way that CFA is not. CFA, however, could be used to compare the fit of the factor model generated by one of the classes onto the other class in an effort to confirm that the correlation structure of the classes is similar. This kind of comparison is not possible with EFA. So, when attempting to verify whether or not a factor structure is present within a set of data EFA is not capable of supplying an answer whereas CFA can provide verification.

EFA and CFA are commonly used tools in psychology research, as well as in economy, sociology, etc. [10]. When developing an assessment, EFA can be used to get a structure for the assessment when no guesses about the factor structure can be made based on expert opinion. This structure found with EFA is used as a model (perhaps with minor modifications) and is tested for validation using CFA on a *separate* set of data. EFA does not need to be used if the structure of the instrument can be inferred by an expert or is built in, as was attempted by the creators of the FCI.

CFA can be broken up into a four-stage process: (i) model development, (ii) estimation of the model's free parameters, (iii) calculation of the model-fit statistics, and (iv) model refinement. Because of the current absence of papers dealing with the application of CFA on conceptual assessments, the steps of CFA will be discussed in detail here; an in-depth discussion can be found in Ref. [10] and many other textbooks on the subject. If the reader is aware of CFA and how it is done, they can move beyond the remainder of this section as well [as] the description of the fit statistics in Sec. II C.

The specifications of the measurement model can come from a number of sources. The two most common sources of model specification are a previous application of EFA on another set of data, or theoretical or expert motivation. In both cases the model is a prescription of the number of latent variables (otherwise known as factors), a specification of how latent variables are measured by the items that make up the assessment (item-factor loadings), correlations amongst errors in the residuals, and other considerations. The residual correlation matrix, often called the residuals, is the difference between the true sample correlation matrix and the model generated correlation matrix. For a model with no correlated errors identified, the only parameters that are freely estimated are the loadings of the questions onto their respective factors and the covariance matrix between the latent variables. All other parameters are set to zero in an effort to make the model as parsimonious as possible.

Once the model has been specified, meaning all of the parameters that need to be estimated have been identified, parameter estimation techniques can be used to estimate or calculate the following values: the loading values of the items onto a factor, the reproduced covariance (or correlation) matrices for the items and latent variables, and the residuals. There are a number of standard parameter estimation techniques that are used in practice, such as maximum likelihood, weighted least squares, and Bayesian estimation techniques.

Model fit statistics can be calculated once the model parameters have been estimated. These model fit statistics allow for one to judge the goodness-of-model fit, and they also allow for a comparison of how different models fit the data. A discussion of the specific fit statistics used in this study can be found after this overview of CFA in Sec. II C.

Models can be refined to improve fit onto a set of data through the use of modification indices and the residuals. Modification indices are a measure of approximately how much the χ^2 of the model's fit will decrease if the parameter identified were allowed to be freely estimated within the model, rather than set to zero. Modification indices can suggest two kinds of changes to models being fit: (i) include correlations between the items, or more specifically between the sources of error for the two items, and (ii) change the loading of questions onto different factors.

The first suggestion of the modification indices can arise when two questions are very similar in content and could cause a student to get both questions wrong for related reasons. This correlation can be encapsulated within the model as an added residual correlation parameter. The second suggestion emerges when questions also correlate well with another group of questions other than the factor to which it is currently assigned (i.e., a suggested cross-loading). Within this study, these kinds of suggestions were not heeded in an effort to keep the factor structures unchanged. This is consistent with the goals of this study, in which we seek to validate specific factor structures; altering the loadings of questions onto factors would change the structures and thus could invalidate the conclusions being made.

Further use of the residuals can guide an expert in making modifications to models by identifying correlations that are poorly estimated and include inserting those correlation parameters directly into the model. A detailed description of these statistics, and their uses, can be found in Ref. [10].

With guidance from the modification indices and an expert's decision, modifications can be made to models. Following these modifications, the model parameters can be estimated, fit statistics calculated, and modification indices and residuals analyzed again in an effort to produce a better fitting model. This iterative process can be repeated until a desired model fit has been achieved. In this study, since the sample was made up of post-test responses, novicelike suggestions from the modification indices occasionally were the most favored change to the models. These kinds of modifications were avoided in favor for expertlike suggestions to get a better fit to the data while retaining the expertlike nature of a model. For example, when questions 6 and 7 on the FCI are considered, it can be seen that they ask about related concepts. In both questions the students are asked to identify the path a ball will take after losing either a normal force or a tension force that was causing the ball to travel along a circular path. From an expert's perspective it should be expected that if a student gets one of these questions right then they will likely get the other one correct as well, and vice versa. Therefore, we added a residual correlation between questions 6 and 7 in each of the three models.

For this study the open-source R software "lavaan" [11] was used to help in the estimation of the model parameters using maximum likelihood estimation, calculation of the standard errors of approximation, modification indices, model-fit statistics, and many other useful statistics. In some software packages the covariance between the latent variables is not freely estimated, leaving only the variances to be estimated; lavaan's default in this regard is to freely estimate the covariance between all of the latent variables unless otherwise specified.

C. Model fit statistics

Fit statistics used in CFA can be broken up into three categories: absolute fit, parsimony correction, and comparative fit [10]. The statistics that are generally calculated in CFA do not necessarily uniquely fit into one of these categories, but each category can be better described by one statistic over another. It should be noted that there is debate on what is considered a good model fit for all of the fit statistics presented here; for some papers that have helped guide the debate, see Refs. [12–14]. The categories of the fit statistics and the statistics themselves are briefly described below.

1. Absolute fit

Absolute fit indices describe how well the model fits the data in an overall sense. This fit does not take into account the fit of the chosen model compared to another model, but is based only on how well the chosen model is able to recreate the correlation matrix of the data. For example, the χ^2 statistic is the difference of the natural logarithms of the determinants of the observed and model generated variance-covariance matrices multiplied by the number of responses minus 1 $[\chi^2 = (\ln|\mathbf{S}| - \ln|\mathbf{\Sigma}|)(N-1)]$. This is a measure of absolute fit of a single model since it makes no reference to another model when it is calculated. The standardized root mean square residual (SRMR) is another fit statistics that describes the absolute fit of the model to the data. The SRMR can be calculated using

$$\mathrm{SRMR} = \sqrt{\frac{1}{a}\sum_{\substack{i=1\\ j<i}}^{n} r_{ij}^2},$$

where a is the number of elements on and below the diagonal of the correlation matrix, r_{ij} are the elements of the residual correlation matrix, n is the number of items, and the summation is on and below the diagonal of the residual matrix. The SRMR statistic has values between 0.0 and 1.0, with 0.0 being a perfect model fit and farther away from 0.0 indicating a poorer model fit. For the SRMR, values close to 0.08 and below are considered in line with good model fit.

2. Parsimony correction

Statistics that fall into this category are different from others in the sense that they introduce penalties for a model having poor parsimony. A model has poor parsimony, or is not parsimonious, if it contains more freely estimated parameters than needed to achieve good model

fit. For example, two models could fit a set of data with the same absolute fit statistics, but one model may be more parsimonious than the other. So, there needs be a way to differentiate between these two models so that the better of the two models can be identified based on fit statistics alone. The parsimony correction index can be used to select the preferred model from the ones with the same, or similar, absolute fit statistics, thereby meeting the goal of using factor analysis to find the most parsimonious model which fits the data. Some very common indices that are used for this category are the root mean square error of approximation (RMSEA), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). Each of these three statistics uniquely introduces a penalty for a model being nonparsimonious. For example, to calculate the RMSEA the following can be used

$$\mathrm{RMSEA} = \sqrt{\frac{\chi_T^2 - df_T}{N df_T}},$$

where df_T is the degree of freedom, χ_T^2 is the chi-squared statistic of the model being tested, and N is the number of observations or students. The values for the RMSEA begins at 0 and is unbounded above. An acceptable value for this statistic are values close to 0.06 and below, where a value of 0.0 is said to be a perfect model fit.

The other two statistics that take model parsimony into account are the AIC and the BIC, which can be calculated with

$$\mathrm{AIC} = 2b_T - 2\ln(L_T),$$

$$\mathrm{BIC} = b_T \ln(N) - 2\ln(L_T),$$

where b_T is the number of freely estimated parameters, N is the number of observations in the sample, and L_T is the likelihood of the model being tested. As can be seen from their equations, the AIC and the BIC are very similar statistics in that they both reward goodness of fit through the likelihood and penalize for increasing the number of estimated parameters. Thus, smaller AIC and BIC values are indicative of the preferred model in terms of comparison to fit and parsimony. This enables the comparison of models which have differing numbers of factors and items, as is the case in this study. When comparing multiple models to each other, the one with the smallest AIC and BIC values is taken to be the preferred model.

3. Comparative fit

The last category of fit indices are ones that compare the fit of the model being tested to a baseline model. The baseline model takes the covariance between all of the items to be zero, and the variances are freely estimated. As one would expect, since the model being tested is being compared to one that makes no assumption about the relationship between the items, the comparative fit indices often look far more favorable than other fit indices presented. However, in comparative studies some of these indices are found to be some of the best behaved amongst all of the indices presented [10]. The two best behaved statistics are the comparative fit index (CFI) and the Tucker-Lewis index (TLI). These can be found using

$$\mathrm{CFI} = 1 - \frac{\max[\chi_T^2 - df_T, 0]}{\max[\chi_T^2 - df_T, \chi_B^2 - df_B, 0]},$$

$$\mathrm{TLI} = \frac{\chi_B^2 - (\chi_T^2/df_T)df_B}{\chi_B^2 - df_B},$$

where χ_T^2 and χ_B^2 is the model being tested and the baseline model's χ^2 value, respectively, and df_T and df_B are the degrees of freedom of the test model and the baseline model, respectively. The CFI values range from 0 to 1, with 0 indicating no fit, and 1 indicating a good fit for the model compared to the baseline. The TLI calculation can yield values outside of the 0 to 1 range; values less than 0 are rounded up to 0 and values greater than 1 are generally rounded down to 1 [10]. This means the TLI can be interpreted in the same way as the CFI. Accepted fit values for both of these statistics are around 0.90 up to the maximum values of 1 for each statistic. Some sources state that roughly 0.95 and above is indicative of a good model fit, however, there is still debate over what dictates a good model fit, and there is no strict agreement as of yet [10].

4. Local strain

The statistics discussed above describe the model's fit to the data in a global sense, meaning they do not look at the residuals individually but all at the same time. Sometimes all of the residuals but one will be within acceptable bounds. These locations of misfit are referred to as a local strain within the model-data fit. Local strain can be found by visual inspection of the modification indices and the residuals, and can be reduced by including residual correlation parameters within the model specifications. It is important to note that the size of the residuals depends on the sample size of the data since the size of the standard error decreases with larger sample sizes. Some methodologists recommend using larger cutoff values for the max allowed residual error as a result [10]. Since the total sample size used for part of this study contained more than 20 000 student responses, the typical bounds for acceptable residuals may be too constraining. When the conventional cutoffs for local strain were enforced for these models, it resulted in models with perfect model fit, and an abundance of residual correlations being added to the models (>20 in total for all three models). As a result of this, and in an effort to keep the models as parsimonious as possible, the demand for no local strain was not enforced.

Additionally, two of the models being examined are expertlike models, making the complete removal of local

strains difficult. Since the sample of students in this data set is not totally expertlike, the expectation that they will fit these models with absolutely no local strain is unreasonable. This was tested for each of the three models by introducing residual correlations into the models until a fit on half of the total student sample (≈10 000 students) that contained no local strain was found. These "no strain" models, which actually had perfect fits to the data according to most of the fit statistics, were then fitted to the other half of the student data. After inspecting the resulting residuals, it was found that local strain reappeared. Thus, the local strain was found to depend on the composition of the class being considered. Upon investigation of the local strains that developed in the new fit, they were all found to be linked to non-expert-like correlations. As a result, some of the local strains within these models were a result of the unique misfits that manifested from the novicelike nature of the students under investigation. A future study that attempts to construct a model that alleviates all local strain or one that looks at the information the local strains about the students is recommended. This local strain issue does not detract from the conclusions made in this paper about the goodness of fit for the models tested since the four fit statistics used were found to all be within acceptable bounds.

D. Random class generation

For the first part of this study the models were fit to the entire 20 822 student sample. The fit statistics for the models are presented in the results section in Table IV. Another goal of this study was to test whether these models could consistently fit smaller sample sizes. The fit, or misfit, of the models when the number of the students in the classes were made smaller will be referred to as the stability of the model within this study.

A model may fit a large set of data, but as the number of students in the sample decreases the individual misconceptions of each student become more prominent within the correlation structure. Because of the desire to retain the expertlike nature of the models being examined the fit of smaller sample sizes is not guaranteed. If an instructor wanted to investigate how well their class matched an expertlike model they may be unable to use one, or both, of the expert models presented due to the potential instability of the models.

To test the stability of the models, sample classes comprised of 4000, 2000, and 1000 students were uniformly drawn from the ranked 20 822 sample population. This results in smaller samples that have similar means and standard deviations as the sample population they were pulled from. For each of these class sizes 2000 classes were drawn with no duplications in classes, meaning no two classes had the exact same students. The global fit statistics were calculated for each of these classes and the means and standard deviations were calculated for each of the fit statistics. Using the rate of misfit, the stability of each of the models can be determined, with a larger misfit rate indicating a less stable model.

III. MEASUREMENT MODEL SPECIFICATIONS

This study focused on testing three models. The development for two of the models is left to the papers in which they were created. One of these models was found by Scott *et al.* [5] through the use of EFA on a sample of 2109 post-instruction student responses. This model is called SSG5 (Scott-Schumayer-Gray, 5 factors). The measurement model for SSG5 can be found in Table I. Within Tables I, II, and III the numbers in each column represent the question numbers from the 1995 FCI, the columns are the factors that

TABLE I. The factor model found by Scott *et al.* [5] for 2109 students postinstruction. The added residual correlations were due to suggested modification indices and from expert consideration of the questions themselves. The numbers in the table indicate the question numbers from the 1995 FCI, and the double headed arrow and the ∼ symbol represent an estimated correlation for those two questions within the model.

Identification of Forces	1st Law w/ 0 force	2nd Law w Kin.	1st Law with canceling forces	3rd Law
5	6	19	16	4
11	7	20	17	15
13	8	21	25	28
18	10	22		
30	12	23		
	16	27		
	24			
	29			
Added residual correlations:				
5∼18	6∼7	19∼20	8∼23	4∼15
29∼30	10∼24	21∼22	23∼24	

TABLE II. The factor model suggested by Hestenes *et al.* [1] with modifications made due to questions not fitting with the data. The added residual correlations were due to suggested modification indices and from expert consideration of the questions themselves. Questions from the FCI are represented as a number in the table, and the double headed arrow and the ∼ symbol have the same meaning as in Table I.

Kin.	2nd Law	1st Law	3rd Law	Forces	Superpos.
12	9	6	4	1	17
14	22	7	15	2	25
19	26	8	16	3	26
20	27	10	28	5	
21		23		11	
		24		13	
				18	
				25	
				30	
Added residual correlations:					
19∼20	8∼9	6∼7	15∼16	1∼2	
21∼22	8∼23	10∼24	23∼24	5∼18	

TABLE III. The factor model developed by Eaton and Willoughby. The added residual correlations were due to suggested modification indices and from expert consideration of the questions themselves. The numbers in the table indicate the question numbers from the 1995 FCI, and the double headed arrow and the ∼ symbol represent an estimated correlation for those two questions within the model.

1st Law+Kin.	2nd Law+Kin.	3rd Law	Force Ident.	Mixed
6	9	4	5	17
7	12	15	11	25
8	14	16	13	26
10	19	28	18	
20	21		30	
23	22			
24	27			
Added residual correlations:				
6∼7	10∼24	21∼22	5∼18	17∼25
8∼23	23∼24	19∼20		

the models used, and the double-headed arrows and the ∼ indicate correlations that were included in the model.

The other two models considered in this study were developed through expert considerations of the questions on the FCI. In Ref. [1] the creators of the FCI proposed a measurement model for the questions on the assessment. This model left alone was found to have many cross loadings and was reduced through the use of model fit statistics and modification indices to make the model more parsimonious. This process resulted in the removal of question 29 from the model due to poor performance. Table II shows the measurement model that came as a result of reducing the original model in Ref. [1]. This model is called HWS6 (Hestenes-Wells-Swackhamer, 6 factors).

The (Eaton-Willoughby, 5 factors) EW5 model breaks the questions up into a mixture of the factors identified in Refs. [5,1] in an effort to create an expertlike model that is capable of fitting smaller sample sizes, which HWS6 had a hard time doing (as discussed in the results section). Instead of treating Newton's first and second laws and kinematics as completely different latent variables, they were combined, resulting in the following two factors: Newton's first law with kinematics and Newton's second law with kinematics. Thus, the EW5 model uses the following factors: each of Newton's three laws and their associated kinematics, force identification, and mixed concepts. This model can be found in Table III. Discussion of these factors follows.

The first two factors of the EW5 model are Newton's first and second laws plus the kinematics that result from these laws. In these factors kinematics pertains to path identification and describing how the speed of an object changes for systems that have zero and nonzero net forces, respectively. The questions that were placed into these factors can be found in Table III. When comparing the Newton's first law factors between the expert models they can be seen to be the same with the exception of one question, question 20. In HWS6 question 20 is put into the kinematics factor of the model, however since EW5 combines Newton's laws and their associated kinematics, question 20 appears in a different factor when the two models are compared. Similarly, questions 12, 14, 19, and 21 (all of the remaining question on HWS5's kinematics factor) moved to the factor in EW5 that combined Newton's second law and its associated kinematics.

The factors classified as Newton's third law in each of the models all have a set of core questions (4, 15, and 28). Question 16 does not appear in the SSG5 model because it did not load in the original EFA analysis done Scott *et al.* In fact, this question was found by Scott *et al.* to probe both Newton's first law and not the third law. In this question, a car pushes a truck while coasting at a constant speed, and from an expert's point of view is probing Newton's third law. This question was found by Scott *et al.* to challenge student understanding of Newton's third law [5].

The factor in EW5 identified as force identification is shared in SSG5, also called force identification. These questions all appear together in the HWS6 model in the forces factor. These questions have been found to create a strong grouping among student responses, and upon inspection all of these questions can be found to be about identifying the forces acting on objects that are stationary or moving at a constant velocity.

The last factor in the EW5 model, called mixed concepts, appears identically in the HWS6 model as the factor superposition principle. Instead of calling it the same name as HWS5, the name mixed concepts was chosen since these questions deal with multiple concepts simultaneously, and not just with superposition of forces. As an example, question 17 is about an elevator being pulled up at a constant speed by a cable. The question asks how the forces acting on the elevator compare to one another. This requires the students to understand how tension works, create a free-body diagram, and then apply Newton's first law to realize that the net force is equal to zero since the system is not accelerating. The other two questions on this factor, questions 25 and 26, are similar to 17 but require an understanding of kinetic friction at an introductory level.

There were some questions in the EW5 model that were left out: 1, 2, 3, and 29. These questions were left out on this expert model due to consideration of the Scott *et al.* model. They found that these questions did not fit into the EFA factors in a satisfactory way. Therefore, these questions were removed from the expertlike EW5 model in order to avoid poor fit using smaller student data samples.

The residual correlations applied to each of the models were found using modification indices from the fits of the models to the random subsamples. These correlations had the largest modification indices, were the most common, and were expertlike for the subsample fits. Correlations were added to the models until a nonexpertlike correlation was the largest suggested correction to the models. The procedure resulted in the addition of 9, 10, and 8 residual

correlations for the SSG5, HWS6, and EW5 models respectively. Other correlations could be added to improve the fits of the models, but this was not needed as the resulting fits were all within acceptable ranges.

Because of this model's consideration of the results in Scott *et. al.*'s resulting EFA factors, particularly the exclusion of questions 1, 2, 3, and 29, this model could be thought of as a hybrid model between expertlike and novicelike. The removal of those four questions was done in an effort to remove poorly performing questions from the model in an attempt to generate an expert model with better fits to the data. The reorganization of the questions from there was done through expert rationale with the intent of not recreating the HWS6 model. The resulting factors as prescribed by EW5 are reasonable from an expert's perspective, and as a result this model will continue to be referred to as an expertlike model.

IV. RESULTS

This section is broken up into two parts, the fit of the models onto the entire sample set, followed by the fit of the models on randomly drawn subsamples of the full sample. The first section shows that each of the models does a good job fitting the entire sample and the second section shows that some of the models have difficulty fitting smaller sample sizes, and are thus referred to as unstable.

A. Entire sample

The fit statistics for each model, with no added residual correlations, when fit to the entire data set can be found in Table IV. All of the models had acceptable fit statistics with no added residual correlations [CFI > 0.9, TLI > 0.9, SRMR < 0.08, RMSEA (Upper CI) < 0.06]. This suggests that the models adequately place questions onto factors in a manner that agrees with the data. Taking parsimony corrections into account, it can be seen that the HWS6 model performs poorly according to the AIC and BIC statistics.

Of all three models analyzed, SSG5 performed the best with the lowest AIC and BIC values of all the models. The goodness-of-fit for SSG5, as well as the other three models, indicates that classes come out of introductory physics with correlation structures that are similar to each other. Suppose the fit for SSG5 had been poor, that would mean that the model generated through EFA performed by Scott *et al.* was not the best way to represent the correlational groupings of the questions for this large sample. If this were the case then the stability of this model would be called into question, and it could be inferred that classes after instruction potentially have unique factor structures. This would imply that postinstruction, student responses in different classes would be correlated with different topics compared to students from another class. However, since SSG5 did have a good fit to the data, that appears to not be the case, and classes after instruction seem to have the same topical understanding of the questions, as measured by the FCI.

The SSG5 model having the best fit of all the models is not surprising since it is a non-expert-like model that was derived from another student sample, so it inherently models some of the main misconceptions held by students postinstruction. Whereas, HWS6 and EW5 are expertlike models, and the differences in the fits between these expertlike models and SSG5 may be due to the presence of nonexpert thinking. Further it can be inferred [6] that the primary non-Newtonian world view that students have after instruction is probably the impetus world view.

Of the expert models tested, the AIC and BIC statistics suggest that EW5 fits the student data better than HWS6, that is the correlational structure of the EW5 model more accurately reflects student responses. This may be because HWS6 is too expertlike, and any presence of novicelike correlations causes the fit to be reduced more for this model than the other expertlike model. Ultimately, all of these models do a satisfactory job at describing the relationships between students' responses to the FCI. An obvious problem with this particular analysis is that classes are not generally in the 20 000-student range. Investigations into whether these models do a good job at fitting smaller samples was performed.

The results of the models with their residual correlations in place can be found in Table IV. As expected, the fit statistics for all of the models improved, either increasing or

TABLE IV. Fit statistics for the three models applied to the full sample, $N = 20\,822$, without and with residual correlations (Res. Cor. in the table).

	No. of Factors	No. of Res. Cor.	CFI	TLI	SRMR	RMSEA (Upper CI)	AIC	BIC
Without residual correlations								
SSG5	5	0	0.922	0.911	0.032	0.041 (0.042)	538 207	538 675
HWS6	6	0	0.911	0.900	0.037	0.040 (0.041)	654 026	654 622
EW5	5	0	0.915	0.904	0.038	0.042 (0.043)	585 590	586 082
With some residual correlations								
SSG5	5	9	0.973	0.968	0.021	0.025 (0.025)	532 708	533 248
HWS6	6	10	0.949	0.941	0.033	0.031 (0.032)	648 887	649 554
EW5	5	8	0.955	0.948	0.032	0.031 (0.032)	580 544	581 100

decreasing where appropriate. These results are supplied so that other models can be compared to the three analyzed in this study.

B. Subsamples of the entire sample set

The 2000 subsamples of 4000, 2000, and 1000 student classes, respectively, were generated by randomly drawing, without replacement, students from the entire 20 822 sample. The fit statistics were calculated for each of these classes for each model, and the results are presented in Tables V, VI, and VII for the SSG5, HWS6, and EW5 models, respectively. We will go through each model's performance one by one, after first describing the meaning of misfit in more detail.

A misfit results when the maximum likelihood algorithm converges to a nonacceptable solution. Occasionally errors such as variances for questions will come out greater than 1 or negative, correlation matrices for the latent variables may be nonpositive definite, correlation between questions may be greater than 1 or negative, etc. These kinds of errors indicate a model is ill specified for the data that it is being fit to, otherwise known as a misfit. When the models were being fit to the randomly sampled classes all instances that resulted in a misfit of the model to a class were counted. After all of the classes had been fit to the models the percentage of the classes that misfit the models were calculated. A cutoff percentage of 15% was used as an indication that a model had fundamental issues when trying to fit the smaller sample sizes. As a result, none of the fit statistics were presented for any model that had a misfit rate above 15% due to the instability of the model at that sample size.

The actual value used for the cutoff, 15%, was chosen in the spirit of the most lenient p-value commonly used in hypothesis testing of $p < 0.15$. The p value is the probability of finding the observed data when the null hypothesis is true. Generally, the null hypothesis is that the model being tested is stable, so the 15% cutoff for the misfit rate carries a similar meaning to testing the null hypothesis. Ultimately, this cutoff rate is arbitrary and could be increased or decreased to change the rigor of the stability testing.

Table V shows the performance for the SSG5 model. As can be seen there were no misfits for any of the sample sizes (which is not the case for the other models) and the fit statistics were satisfactory to claim good model-data fit. In fact, the fit statistics can be seen to change little from the 4000 student classes to the 1000. This indicates that this model is stable for the smaller sample sizes, and that using this model to fit individual classes may be possible. Since Scott *et al.* [6] found the most prominent misconceptions contained in their data, an instructor may want to check the fit of this model to their students to verify if they potentially possess similar misconceptions.

The HWS6 model had a relatively large number of misfits using the smaller sample sizes. In fact, for the samples of 2000 and 1000 students this model misfit the classes more that 15% of the time. However, when the model did not have a misfit it retained a good fit with the data. Because of the large misfit rate this model can be concluded to be unstable

TABLE V. Fit statistics' mean and standard deviations of 2000 samples of size 4000, 2000, and 1000 students for the SSG5 model. None of the smaller samples tested had any misfits, which is an indication that this model is considered stable within this study. As the sample sizes get smaller the fit statics get worse, but never get to the point of representing a poor model fit.

Scott-Schumayer-Gray 5 factors—SSG5								
Without residual correlations								
	Mean	St. Dev.		Mean	St. Dev.		Mean	St. Dev
4000 students			2000 students			1000 students		
CFI	0.921	0.0043	CFI	0.920	0.0065	CFI	0.918	0.0097
TLI	0.910	0.0049	TLI	0.908	0.0074	TLI	0.906	0.0112
SRMR	0.034	0.0009	SRMR	0.036	0.0013	SRMS	0.039	0.0018
RMSEA	0.041	0.0011	RMSEA	0.042	0.0017	RMSEA	0.042	0.0026
RMSEA upper CI	0.043	0.0011	RMSEA upper CI	0.044	0.0017	RMSEA upper CI	0.046	0.0025
AIC	103 440	528	AIC	51 737	391	AIC	25 878	284
BIC	103 812	528	BIC	52 067	391	BIC	26 168	284
With residual correlations								
4000 students			2000 students			1000 students		
CFI	0.972	0.0024	CFI	0.971	0.0039	CFI	0.969	0.0064
TLI	0.967	0.0029	TLI	0.966	0.0046	TLI	0.963	0.0076
SRMR	0.023	0.0008	SRMR	0.026	0.0012	SRMS	0.031	0.0016
RMSEA	0.025	0.0011	RMSEA	0.025	0.0017	RMSEA	0.026	0.0028
RMSEA upper CI	0.027	0.0011	RMSEA upper CI	0.028	0.0017	RMSEA upper CI	0.026	0.0026
AIC	102 372	531	AIC	51 217	385	AIC	25 620	289
BIC	102 800	531	BIC	51 598	385	BIC	25 953	289

TABLE VI. Fit statistics' mean and standard deviations of 2000 samples of size 4000, 2000, and 1000 students for the HWS6 model. Many of the smaller samples tested misfit with the model, which is an indication that this model is not a good representation of smaller samples.

Hestenes-Wells-Swackhamer 6 factors—HWS6								
Without residual correlations								
	Mean	St. Dev.		Mean	St. Dev.		Mean	St. Dev
4000 students	Misfit rate = 5.05%		2000 students			1000 students		
CFI	0.910	0.0039	CFI			CFI		
TLI	0.899	0.0044	TLI			TLI		
SRMR	0.039	0.0010	SRMR			SRMS		
RMSEA	0.040	0.0009	RMSEA	Misfit rate >15%		RMSEA	Misfit rate >15%	
RMSEA upper CI	0.042	0.0009	RMSEA upper CI			RMSEA upper CI		
AIC	125664	585	AIC			AIC		
BIC	126136	585	BIC			BIC		
With residual correlations								
4000 students	Misfit rate = 3.10%		2000 students			1000 students		
CFI	0.950	0.0028	CFI			CFI		
TLI	0.942	0.0033	TLI			TLI		
SRMR	0.034	0.0011	SRMR			SRMS		
RMSEA	0.031	0.0009	RMSEA	Misfit rate >15%		RMSEA	Misfit rate >15%	
RMSEA upper CI	0.032	0.0009	RMSEA upper CI			RMSEA upper CI		
AIC	124655	580	AIC			AIC		
BIC	125190	580	BIC			BIC		

for small sample sizes. This instability at smaller sample sizes may be due to the fact that this model is too expert-like, as mentioned previously. As the sample sizes get smaller and smaller, misconceptions of an individual student become more prominent within the correlation matrix. As a result even subtle differences in response patterns from a handful of students could be enough to cause a misfit between this model and the data. This is conjecture, and more analysis should be done in a future study to address the specifics for why this model does so poorly at smaller sample sizes.

TABLE VII. Fit statistics' mean and standard deviations of 2000 samples of size 4000, 2000, and 1000 students for the EW5 model. Some of the smaller samples tested misfit with the model, but the rate of misfits was never above 3% for the model with correlations in place. This suggests that for samples smaller than 1000 students the model without correlation should be fitted to the data first and then correlation can be added after model-data fit has been established. As the sample sizes get smaller the fit statics get worse, but never get to the point of representing a poor model fit.

Eaton-Willoughby 5 factors—EW5								
Without residual correlations								
	Mean	St. Dev.		Mean	St. Dev.		Mean	St. Dev
4000 students			2000 students			1000 students		
CFI	0.914	0.0039	CFI	0.913	0.0061	CFI	0.911	0.0094
TLI	0.903	0.0044	TLI	0.902	0.0068	TLI	0.899	0.0106
SRMR	0.040	0.0011	SRMR	0.041	0.0015	SRMS	0.045	0.0022
RMSEA	0.042	0.0010	RMSEA	0.043	0.0015	RMSEA	0.043	0.0024
RMSEA upper CI	0.044	0.0010	RMSEA upper CI	0.045	0.0016	RMSEA upper CI	0.047	0.0023
AIC	112 521	516	AIC	56 271	395	AIC	28 170	284
BIC	112 912	516	BIC	56 618	395	BIC	28 474	284
With residual correlations								
4000 students			2000 students	Misfit rate = 0.30%		1000 students	Misfit rate = 2.85%	
CFI	0.954	0.0028	CFI	0.953	0.0046	CFI	0.951	0.0069
TLI	0.947	0.0032	TLI	0.946	0.0053	TLI	0.943	0.0080
SRMR	0.034	0.0011	SRMR	0.036	0.0017	SRMS	0.039	0.0022
RMSEA	0.031	0.0010	RMSEA	0.032	0.0016	RMSEA	0.033	0.0024
RMSEA upper CI	0.033	0.0010	RMSEA upper CI	0.034	0.0016	RMSEA upper CI	0.036	0.0023
AIC	111 564	539	AIC	55 788	409	AIC	27 925	297
BIC	112 005	539	BIC	56 180	409	BIC	28 269	297

The results of the other expert model, EW5, can be found in Table VII. This model appears to be a better representation of how the questions fit together as the misfit rate is drastically lower compared to the HWS6 model. The fit statistics are good for all of the categories and change very little as the class size decreases. Comparing the AIC and BIC for this model and HWS6 for 4000 students in a class, it can be seen that this model is better at describing the data with and without residual correlations in place. Between the expert models, EW5 appears to do a better job in general at fitting the data to latent variables.

For the EW5 model there were a few cases of misfit at the lower sample sizes with residual correlations in place. This makes sense since the residual correlations that were included were only expertlike. This means the model with the residual correlations in place can be considered to be more expertlike compared to the measurement model without correlations in place. As a result, classes with more novicelike correlations will potentially have a harder time fitting this model and thus the misfit rate increases. This can be seen to be the case in Table VII for the 2000 and 1000 sample sizes comparing the misfit rates with and without the residual correlations. As a result, if an instructor wanted to try to fit this model to their own class's results they should consider starting with the correlation free model to initially see how well their class fits the expertlike EW5 model.

V. CONCLUSIONS

Confirmatory factor analysis was applied to three models using a set of data with 20 822 postinstruction student responses to the Force Concept Inventory. Of the models, one was found through the use of exploratory factor analysis applied to 2109 students by Scott *et al.* [5]. The other two models were expert created models without the use of EFA, described in the model specification section. All of these models were found to fit the full sample with satisfactory fit values. This means that all three of these models could be used to describe the correlations between students' responses to the FCI after instruction.

Further analysis of the SSG5 data by Scott *et al.* [6] revealed that the impetus world view was the primary misconception held by the students whose response data generated the SSG5 model. Because of the good model fit of SSG5 with this larger data set it could suggest that the impetus world view is the chief issue students still have after instruction.

The fact that the HWS6 model has an acceptable fit with the full sample ($N = 20\,822$) suggests that this factor structure accurately represents postinstruction student responses on the FCI. Through the use of CFA (instead of EFA) this study finds that it agrees with Halloun and Hestenes in that the factor structure presented in Ref. [1] is an acceptable way to categorize the questions of the FCI.

This conclusion disagrees with the conclusions of Huffman and Heller presented and discussed in Refs. [2,4]. The disagreement between these studies could primarily be a result of the statistical tools chosen to answer the research question. EFA as [a] tool is not capable of confirming or denying the existence of a given factor structure within a set of data. This is not to say that the validity of a model cannot be inferred through the use of EFA on alternate sets of data, which is a common application. That no consistent model was found through the application of EFA on different sets of data may imply that the factors of the FCI are sensitive to the sample being analyzed. This study demonstrates that the FCI does measure what it was design to measure as described by Hestenes *et al.* from a factor perspective. Given the results seen herein, we hope to end the air of caution carried by researchers about the factor structure of the FCI and to end the debate that began in 1995.

Because these expert factor models have been confirmed for the FCI, instructors can now look at graded "chunks" of the FCI. For instance, using the EW5 model, the FCI can be thought of as testing the 5 factors indicated in the model specifications. Thinking specifically about Newton's third law, an instructor can inspect how a student (or the whole class) answered questions 4, 15, 16, and 28 and determine the extent to which Newton's third law is understood after instruction. This is a simple example of what can be done, but the power this gives instructors for targeting concepts their students are struggling with could help in assessing instruction methods for specific sections of the material.

After examining the fits of these models to smaller samples constructed from the larger data set, it was found that the SSG5 and the EW5 models did a good job fitting with little or no misfits. HWS6, however, had a hard time fitting these smaller samples. This seems to indicate that when students have conceptual issues with Newton's first or second laws they may also have difficulties with their associated kinematics. This is opposed to viewing Newton's laws and kinematics as being entirely separate factors and thus give the impressions that one could grasp one topic without fully understanding the other.

VI. FUTURE WORK

Further research into why the HWS6 model has a hard time fitting the smaller sample sizes is suggested. Also, investigating how these models fit data that is only from an algebra- or calculus-based class and a comparison of these results is currently being pursued. Other affects, like gender or different teaching styles, on the fit of these models is also being considered.

An investigation into what the suggested modification indices can reveal about classwide postinstruction misconceptions in an exploratory or confirmatory factor analysis style of analysis is being investigated as well.

ACKNOWLEDGMENTS

The authors would like to thank Keith Johnson and Barrett Frank for reading over the manuscript, offering insightful suggestions, and the many interesting discussions related to this project. This project was funded by the Montana State University Physics Department.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[2] D. Huffman and P. Heller, What does the force concept inventory actually measure?, Phys. Teach. **33**, 138 (1995).
[3] 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).
[4] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[5] 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).
[6] 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).
[7] M. R. Semak, R. D. Dietz, R. H. Pearson, and C. W. Willis, Examining evolving performance in the Force Concept Inventory using factor analysis, Phys. Rev. Phys. Educ. Res. **13**, 010103 (2017).
[8] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).
[9] PhysPort, Security FAQ for the Assessment Data Explorer (accessed November 20, 2017).
[10] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. (Guilford Press: A Division of Guilford Publications, Inc., New York, 2015), pp. 72–75.
[11] Y. Rosseel, lavaan: An R Package for Structural Equation Modeling, J. Stat. Softw. **48**, 1 (2012).
[12] L.-t. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Struct. Eq. Modeling: A Multidisciplinary J. **6**, 1 (1999).
[13] M. W. Browne and R. Cudeck, Alternative ways of assessing model fit, Socio. Methods Res. **21**, 230 (1992).
[14] R. C. MacCallum, M. W. Browne, and H. M. Sugawara, Power analysis and determination of sample size for covariance structure modeling, Psychol. Methods **1**, 130 (1996).