Evidence of measurement invariance across gender for the Force Concept Inventory

Philip Eaton*

School of Natural Sciences and Mathematics, Stockton University, Galloway, New Jersey 08205, USA

(Received 24 November 2020; accepted 19 March 2021; published 26 April 2021)

A performance gap between the genders has been observed on the Force Concept Inventory (FCI) almost since its introduction. Many studies have sought to characterize this gender gap, however, few have tested the consistency of the factor structure across the genders. This study fills in this gap by offering the first piece of evidence that the genders (male and female due to data constraints) are interacting with the FCI in similar manners. Using multigroup measurement invariance techniques, a preinstruction sample of 6238 males and 2874 females, and a postinstruction sample of 6338 males and 2955 females, the latent variable structure of the FCI was tested for consistency between the genders. As this technique is not often used by the physics education research community, significant time was spent explaining the methodology. It was found that the Eaton and Willoughby five-factor modified (EW5M) model of the FCI factor structure exhibited strict invariance between the genders. Additionally, a single-factor model showed strong invariance and partial strict invariance for the FCI. Latent means from the EW5M model revealed that females underperform compared to males across all of the factors of the FCI, but these gaps in performance decreased over the course of instruction, however, some differences are still large. The results of this study suggest observed performance differences on the FCI between the genders may not be due to gender specific factor structure differences. However, this result is sample dependent and should be verified by other studies using different, independent samples.

DOI: 10.1103/PhysRevPhysEducRes.17.010130

I. INTRODUCTION

The Force Concept Inventory (FCI) is one of the most popular conceptual assessments currently used in physics education and physics education research [1]. This assessment was originally designed in 1992, was updated in 1995, and presently contains 30 multiple-choice items [1]. The FCI is commonly used to investigate student understanding of introductory Newtonian mechanics, such as Newton's three laws, kinematics, force identification, etc., at the high school and introductory college or university level. The introduction of the assessment marked a turning point in how physics education research was performed. Now, well-understood research-based assessments are being increasingly used in physics education studies. These assessments can be used by all instructors and researchers and are regularly tested for reliability and, subsequently, validity.

The evidence for the validity of an assessment is built up over time as the collection of studies analyzing characteristics of the assessment grows. The FCI has been subjected to numerous statistical analysis since its introduction. The following is a list of the kinds of statistical studies, with some citations, that have been performed on the FCI: item response curves [2–4], (unidimensional) item response theory [5–7], classical test theory or classical item analysis [7–11], local dependence analysis [12], exploratory factor analysis [13–15], confirmatory factor analysis [16], multitrait (multidimensional) item response theory [17–19], cluster analysis [20,21], and network analysis [22,23]. This list is not complete, but gives a sense of the amount of research performed in an attempt to better understand this assessment. The consensus of these studies is that the FCI is a well-functioning instrument and appears to possess favorable statistical properties overall. This does not mean the FCI *is* a valid assessment, simply that there is considerable evidence that *suggests* the FCI is a valid assessment.

With this validity evidence in place, it may be surprising to some instructors and researchers that males frequently outscore females on the FCI [7,24–27]; this is not a complete list of the studies that have made this observation. Though performance differences between the genders on the FCI would not make the assessment invalid, it does raise serious questions about the source of these differences. For example, are these differences due to males and females interpreting the items of the FCI differently? Are any of the items more aligned with societal norms for males versus females (i.e., bias)?

Characterizing and quantifying the gender performance gap on the FCI is imperative. On one hand, if these

*philip.eaton@stockton.edu

differences are identified as being sourced by the assessment (e.g., item bias) then the assessment should be modified to fix the issue(s) and no new, specific pedagogical tools need to be developed. On the other hand, if it is found that all or part of the gender gap is a result of true group differences, then developing pedagogical tools to counter these differences becomes more important, and possible.

When an assessment is expected to perform identically between two groups, the invariance of the assessment's factors should be assessed; this is done through multigroup measurement invariance strategies [28–31]. For example, if an assessment was originally written in English and was later translated to Japanese, then the invariance of the factor model should be studied. Since the assessment would be expected to function identically in both languages, it would be hypothesized that measurement invariance is present between the versions of the assessment, for example, see Ref. [32].

Specifically, "measurement invariance" is a property of an assessment that signifies the same latent variable constructs are being measure across the groups being examined. This kind of invariance can be separated into different levels like configural, metric, strong, and strict. Configural invariance occurs when the assignment of items to factors are the same across the groups, and metric invariance goes one step farther by forcing items and their factors to be related in the same manner up to a constant across the groups. Strong invariance takes the metric invariance and additionally restricts the constant so that items and their factors are identically related across groups. Lastly, taking strong invariance and further constraining the errors of the items to be the same across the groups is called strict invariance. It should be noted that as this analysis is for the factor structure of the assessment, the results recovered will be sample dependent. That is, multiple studies using the same analysis, but different samples should be performed to corroborate or refute the result presented in this article.

As research currently stands, the measurement invariance of factors of the FCI, like those proposed in Refs. [1,13,16], has not been tested outside of a longitudinal analysis [33]. This longitudinal study tested the assumption of measurement invariance for the factor structure of the FCI from pre- to post-instruction in a calculus-based freshman physics course covering Newtonian mechanics (typically called Physics I). It was shown that the factor structure of the FCI had strong invariance and partial strict invariance under both the Scott, Schumayer, and Gray five-factor modified model and the Eaton and Willoughby five-factor modified model [13,16,33].

This study will supply more evidence for or against the measurement invariance of the FCI's factor structure by examining the following research question:

To what extent does measurement invariance hold between the genders on the FCI under the Eaton and Willoughby five-factor modified model and a single-factor model?

It should be noted here that the analysis presented in this study is sample dependent. That is, it should not be assumed as a generally applicable result until more similar analyses have been performed on separate samples from varying types of introductory physics courses (algebra-based, calculus-based, high school, etc.).

II. METHODOLOGY

A. Overview of measurement invariance

Confirmatory factor analysis is a statistical theory that models the latent variable structure (or factors) of an assessment [34]. Specifically, factors of assessments are tested to supply evidence they are consist with an expert's interpretation of how the items (i.e., questions) of an assessment fit together.

Factors are groups of items which are theoretically assumed or have been shown from exploratory analysis to be related through a latent variable, which is often interpreted as a concept like Newton's third law or force identification, for example. It should be noted that these conceptualizations are an expert's, or a group of experts', interpretation of the factors, and not all interpretations will necessarily be the same.

When a factor model of an assessment is shown to have good fit with a sample of data, it can be inferred that the sample is likely relating the items in a manner consistent with the proposed factors [34]. In turn, score interpretations for the assessment can be made with greater confidence, and more evidence for the validity of the assessment is gained in the process.

Measurement invariance is a property of an assessment and indicates that the same latent variable structure is being measured between different groups (i.e., across race, ethnicity, gender, semester, etc.) [28–31]. If an assessment is found to have measurement invariance between two groups, then it can be assumed that the latent variables of the assessment are being interacted with in a similar manner. In the event measurement invariance is not established, then it can be assumed that the assessment is likely not being conceptualized in a similar manner between the groups. This differential functioning could be cultural, due to language differences, due to true differences between the groups being compared, etc.

For example, one study in 2017 found that Japanese students scored better overall on the Force and Motion Conceptual Evaluation (FMCE) compared to American students. The Japanese students took the Japanese translation of the FMCE and the American students took the English version [4]. Another study in 2012, analyzed a different sample of Japanese and American students and

found that the most incorrect responses to items on the FCI were the same between the two groups in the sample [35]. These performance similarities and differences *could* imply Japanese students have a better understanding of Newtonian mechanics compared to American students. However, since measurement invariance between the different language versions of the FMCE and FCI has not been demonstrated, it is hard to properly interpret these results. That is, it is unclear if these results are due to structural differences in the versions of the assessment or due to true differences in the samples [4].

When measurement invariance is not present between two groups, scores from the assessment will depend on not only a respondent's latent ability, but also on their group membership. This means individuals with identical latent abilities may not be expected to receive similar scores on the assessment due to potential group membership differences. For example, on the five factor model for personality traits (i.e., the "big five") it has been shown that adult women score higher on neuroticism and agreeableness compared to adult men [36]. However, it was later found in a different sample that strict measurement invariance did not exist for the five factor model between men and women [37]. This resulted in structural induced differences in scores on the personality traits that were not necessarily native to the groups themselves. After controlling for these differences, it was found that the statistical noninvariance had not affected the overall qualitative results made previously [37]. It is important to note that, although the affect of noninvariance in this study was small, this example does not guarantee a small impact for other assessments and/or samples.

In the case of the FCI, measurement invariance between binary genders (self-identified male and female indicators for this study due to data limitations) has not been assessed. As a result, it is hard to directly compare the performance of these groups on the FCI without being able to estimate the impact of potential structural differences. Without this estimation, it is impossible to distinguish between differences in total scores due to structural differences and differences native to the groups.

Testing for measurement invariance of a factor model has generally accepted procedures for continuous and discrete-ordered independent variables [28–31,38]. As the FCI's items are traditionally graded dichotomously, a measurement invariance analysis should be carried out using the discrete-ordered procedures. The procedure will be briefly discussed here for comparing two samples, but can be extended to compare more than two samples at once [28–31,38].

To begin, the proposed factor model must be shown to adequately fit the samples being examined. If the factor model does not fit one of the samples, then there is evidence to suggest that measurement invariance is not present between the two groups. It is generally accepted that the confirmatory fit index (CFI), the Tucker-Lewis fit index (TLI), and the root mean square error of approximation (RMSEA) be used to assess the goodness of fit of models to samples [16,34,39]. A model is said to have adequate fit to the sample if fit values meet the following criteria: CFI ≥ 0.95, TLI ≥ 0.95. RMSEA ≤ 0.08 [39]. If there is adequate fit between the model and both groups, then configural invariance can be tested.

Configural invariance tests that both samples have the same number of factors, with the same questions on each factor [28,31,38,40]. This is done by freely estimating the factor loadings and thresholds of the items in the factor model, but constraining the residual variances to 1 and latent means to 0 for both groups [28,31,38]. These constraints make this fitting different from simply fitting the model to each group independently, as is done immediately before this step. Factor loadings are related to the correlation between the item and factor, and thresholds are the level of latent ability a student needs to transition from answering the question incorrectly to correctly. Residual variance is the error in the estimated variances of the items and factors, and the latent means represent the expected value of the group's latent variables when the predictor is set to zero. That is, the latent mean gives an estimation of where the "center," or intercept, of the latent ability scale is for each factor.

Since factor loading values and thresholds must be varied together for ordered-categorical variables, the process of checking for metric invariance is skipped for this procedure [28,31,38]. Metric invariance is a check to verify if the factor loadings are the same across the groups while leaving the thresholds free to vary. This is not possible for ordered-dichotomous data since the factor loadings and thresholds must be estimated simultaneously.

Thus, the next step is to check for strong invariance between the groups. This is done by fixing the factor loadings and thresholds for the items to be the same across the groups. The residual variance and the latent means are fixed to 1 and 0, respectively, for the reference group and are freely estimated for the focus group [28,31,38]. The reference group is the group whose performance the focus group is compared against. For this study the reference and focus groups were made up of the males and females in the data, respectively.

At this level, since the factor loadings and thresholds are fixed, the estimated latent means will give a measure of how the focus group performed compared to the reference group while controlling for factor structure differences. If the latent means of the focus group are estimated to be near zero, that means the scales of the factors are being anchored to nearly the same value for both groups [28,31,38,40].

The final check is for strict invariance. To test for strict invariance, take the strong invariance model and further restrict the residual variances of the focus group to be equal to 1. This imposes an identical factor structure on both groups (strong invariance) with the addition of both groups

having the same errors of estimation for the item variances. It should be noted that this level of invariance is generally not observed in practice as it is a highly constrained model [40].

Incidentally, strict invariance is not mandatory for comparing the latent variable means of two groups. Because the residual errors contain both random errors and item-specific errors, there is little reason to assume these will be the same for two random samples from the same population, let alone for 2 samples from different populations.

Each type of measurement invariance is compared to the previously tested type in a sequential process. That is, the fit values for strong invariance are compared to those for configural invariance. Similarly, strict invariance is compared to strong invariance. A difference in the CFI between consecutive models larger than 0.01 in magnitude indicates an unacceptable change in the fit of the model [41]. This is also true for changes in the RMSEA larger than 0.015 in magnitude [41].

B. Assessment and factor model specification

The FCI was originally designed to probe 6 domains of Newtonian mechanics, but research has suggested that it is actually only probing 5 different domains [1,13,16,19]. One of the factor models suggested for the FCI is the Eaton and Willoughby five-factor model. This model was designed as a hybrid of the original creator's model and one found through exploratory analysis [1,13,17]. This model does not include 4 items from the FCI (1, 2, 3, and 29), however, it has been shown these could be reinserted to achieve acceptable factor reliability [33]. This updated model is called the Eaton and Willoughby five-factor modified model (EW5M) [33]. The item-factor assignments can be found in Table I, along with conceptual identifiers for each factor.

It has been shown that the FCI may be modeled using a bifactor structure, which suggests a single-factor model may adequately represent student responses to the FCI [19]. However, the FCI has been shown to have local dependence between some of items and should technically be modeled using multiple factors [12].

Incidentally, some instructors and researchers may be unaware of the need, or how, to score the FCI in a multidimensional manner. As a result, many instructors and researchers may still score the FCI along a single factor (e.g., the total aggregate score). Because of this, the single-factor model will be analyzed in this study to investigate if these scores are being significantly impacted by potential observed structural differences between the groups. Thus another model analyzed in this study is a strict single-factor model.

TABLE I. The item-factor relations for the Eaton and Willoughby five-factor modified model. The items added to the original Eaton and Willoughby model are indicated in bold.

	Factor name	Items
F1	Newton's 1st law + kinematics	6, 7, 8, 10, 20, 23, 24
F2	Newton's 2nd law + kinematics	**1**, **2**, **3**, 9, 12, 14, 19, 21,22, 27
F3	Newton's 3rd law	4, 15, 16, 28
F4	Identification of forces	5, 11, 13, 18, **29**, 30
F5	Superposition	17, 25, 26

C. Data

The data used in this study were received from PhysPort and contained a mixture of pre- and post-test student responses to the 1995 version of the FCI [42]. The data were a collection of algebra- and calculus-based introductory physics courses from a myriad of American post-secondary education institutes. Specifically, these data were voluntarily given to PhysPort by the instructors of the courses which made up the sample.

The data were separated into two groups via test administration (pre- and post-test) and all students without gender identifiers were removed. After inspecting the resulting data, it was revealed that only the gender identifies for males and females had large enough sample sizes to be used in factor analysis. At this point only students whose gender was listed as male or female were left in the sample. Thus, subsequent analyses were only performed for these genders.

Listwise deletion was then applied to remove incomplete student response vectors. Listwise deletion is a method for handling missing data in which students who are missing even a single response to the assessment are removed from the sample. This resulted in less than 10% of the data being removed from each of the groups (pre or post-test and male or female).

Finally, the data were graded dichotomously, meaning questions were scored as being either correct or incorrect represented as "1" and "0" in the data, respectively. The resulting pretest sample contained 6238 males and 2874 females, and the post-test sample contained 6338 males and 2955 females; test statistics for each of these samples can be found in Table III.

D. Analysis

All of the analyses in this study were performed using the statistical software R [43], specifically the lavaan package [44]. It should be noted that lavaan's default setting is the delta parametrization for structural models, however, the theta parametrization should be used when performing the analysis discussed previously. This parametrization allows for the residual variances of the factors to the treated as parameters, which can be set to 1 or freed depending on the invariance being tested [31]. It should be noted that the delta and theta parametrization will produce identical results, they are simply two different ways of parametrizing a confirmatory model [31].

III. RESULTS

All of the fit statistics discussed in the following sections can be found in Table II.

It was found that the male and female groups fit both the EW5M and the single-factor models with adequate fit indexes. Similar to the results of Ref. [33], it was found that the models fit the pretest responses to the FCI better than the post-test responses; however, these differences are not considered significant for the EW5M model (ΔCFI < 0.01 for both groups).

A. EW5M results

The results indicate good model fit for configural invariance pre- and post-test between the genders with all the reported fit indexes being within acceptable ranges (pre; post: CFI $= 0.9897$; 0.9847, TLI $= 0.9887$; 9832, RMSEA $= 0.0319$; 0.0362). This suggests that the EW5M factor structure is consistent between the genders. Next, strong invariance was found to have acceptable model fit (pre; post: CFI $= 0.9884$; 0.9827, TLI $= 0.9876$; 9815, RMSEA $= 0.0333$; 0.0380) and did not have a significant change in fit from configural invariance (pre; post: ΔCIF $= -0.0012$; -0.0020, ΔRMSEA $= 0.0015$; 0.0018). Lastly, strict invariance also had good fit with the sample (pre; post: CFI $= 0.9825$; 0.9784, TLI $= 0.9820$; 0.9777, RMSEA $= 0.0401$; 0.0417) and did not have a significant change in fit from strong invariance (pre; post: ΔCIF $= -0.0059$; -0.0044, ΔRMSEA $= 0.0068$; 0.0037). It can also be seen that the strict invariance fit did not significantly differ from the configural invariance fit (pre; post: ΔCIF $= -0.0071$; -0.0064, ΔRMSEA $= 0.0083$; 0.0055).

These results suggest that strict measurement invariance between the genders does exist for the FCI under the

TABLE II. Model fit and measure invariance results for the sample. A model is said to have adequate fit with the data if CFI $>\geq 0.95$, TLI ≥ 0.95, and RMSEA ≤ 0.080 [39]. When comparing nested models, model fit degeneration is deemed acceptable provided ΔCFI < 0.010 and ΔRMSEA < 0.015 [41]. Values that were found to be out of any of these ranges have been indicated in bold. The upper 90% C.I. column reports the upper 90% confidence interval for the RMSEA.

Pretest							
Model	CFI	TLI	RMSEA	(Upper 90% C.I.)	ΔCFI	ΔRMSEA	ΔRMSEA (Upper 90% C.I.)
EW5M							
Male	0.9909	0.9900	0.0330	(0.0341)			
Female	0.9833	0.9817	0.0292	(0.0309)			
Configural invariance	0.9897	0.9887	0.0319	(0.0328)			
Strong invariance	0.9884	0.9876	0.0333	(0.0342)	−0.0012	0.0015	(0.0014)
Strict invariance	0.9825	0.9820	0.0401	(0.0410)	−0.0059	0.0068	(0.0068)
Single-Factor Model							
Male	0.9770	0.9753	0.0520	(0.0530)			
Female	0.9652	0.9626	0.0417	(0.0433)			
Configural invariance	0.9750	0.9732	0.0490	(0.0498)			
Strong invariance	0.9731	0.9720	0.0500	(0.0509)	−0.0020	0.0010	(0.0010)
Strict invariance	0.9603	0.9602	0.0596	(0.0605)	**−0.0128**	0.0096	(0.0096)
Partial strict invariance[a]	0.9687	0.9677	0.0537	(0.0546)	−0.0053	0.0037	(0.0037)
Post-test							
Model	CFI	TLI	RMSEA	(Upper 90% C.I.)	ΔCFI	ΔRMSEA	ΔRMSEA (Upper 90% C.I.)
EW5M							
Male	0.9865	0.9852	0.0361	(0.0372)			
Female	0.9787	0.9762	0.0366	(0.0382)			
Configural invariance	0.9847	0.9832	0.0362	(0.0371)			
Strong invariance	0.9827	0.9815	0.0380	(0.0389)	−0.0020	0.0018	(0.0018)
Strict invariance	0.9784	0.9777	0.0417	(0.0425)	−0.0044	0.0037	(0.0036)
Single-Factor Model							
Male	0.9673	0.9648	0.0556	(0.0567)			
Female	0.9473	0.9433	0.0565	(0.0581)			
Configural invariance	0.9627	0.9600	0.0558	(0.0567)			
Strong invariance	0.9591	0.9575	0.0575	(0.0583)	−0.0036	0.0017	(0.0017)
Strict invariance	0.9475	0.9474	0.0640	(0.0648)	**−0.0116**	0.0065	(0.0065)
Partial strict invariance[a]	0.9555	0.9554	0.0589	(0.0598)	−0.0036	0.0014	(0.0014)

[a]Partial strict single-factor model: Freely estimated the variance of the single factor for the focus group (Females).

EW5M model for this sample. Thus, the latent mean differences between the groups can be used to give a measure of how much each group's performance differed on the factors of the EW5M model. Under strict invariance the latent factor variances are set to 1 for both groups and the reference group's (the male's) latent variable mean are set to 0. Because of this, the estimated latent means for the focus group (females) can be interpreted in a similar manner as a Cohen's d, while controlling for structural differences.

The latent mean differences under strict invariance can be found in Table IV. These differences align well with taking an aggregate score for each factor, see Table III. Further, these results suggest female performance is catching up to male performance from pre- to post-test across all factors of the EW5M model. The factor which has the most significant closing of the initial gender gap is factor F3, which measures Newton's third law. This suggests that by the time an introductory physics course in completed, females and males are performing more similarly compared to the beginning of instruction, particularly with the concept of Newton's third law.

It would be interesting to investigate if these gender gaps continue to close as students continue in their physics schooling, and is suggested as a followup study. Additionally, the comparison of latent means demonstrates another method for statistically assessing the effectiveness of new pedagogical techniques, while controlling factor scores for structural differences between groups (genders in this case).

TABLE III. The conceptual meaning for each factor (F1–F5) can be found in Table I. The d column reports the Cohen's d between the groups; positive values indicate females outperformed males. Cohen's d values in boldface are typically considered medium for effect sizes (0.5–0.8), values in plain text are considered small in size (0.2–0.5), and values in italics are considered negligible (0.0–0.2).

	Pretest: $N_{\text{Male}} = 6238$ and $N_{\text{Female}} = 2874$		
Factor	M Mean (St. Dev.)	F Mean (St. Dev.)	d
Full FCI	0.507 (0.243)	0.332 (0.196)	**−0.677**
F1	0.658 (0.280)	0.436 (0.228)	**−0.691**
F2	0.588 (0.252)	0.401 (0.216)	**−0.678**
F3	0.461 (0.359)	0.317 (0.325)	−0.359
F4	0.361 (0.351)	0.204 (0.263)	−0.431
F5	0.237 (0.347)	0.132 (0.259)	−0.291
	Post-test: $N_{\text{Male}} = 6338$ and $N_{\text{Female}} = 2955$		
Factor	M Mean (St. Dev.)	F Mean (St. Dev.)	d
Full FCI	0.664 (0.228)	0.535 (0.221)	−0.495
F1	0.782 (0.250)	0.616 (0.274)	**−0.543**
F2	0.676 (0.245)	0.515 (0.237)	**−0.570**
F3	0.711 (0.312)	0.669 (0.332)	*−0.112*
F4	0.600 (0.332)	0.490 (0.316)	−0.291
F5	0.419 (0.402)	0.319 (0.365)	−0.223

TABLE IV. EW5M latent means are given with strict invariance; male latent means are fixed to zero and estimated female latent means give an indication of group performance differences for each factor. The single-factor (S-F) model results are given with strong invariance. All of the latent means are statistically significant ($p < 0.001$).

	Pretest		Post-test	
EW5M	Latent mean	Std. error	Latent mean	Std. error
F1	−0.913	0.027	−0.733	0.027
F2	−0.923	0.027	−0.763	0.027
F3	−0.508	0.027	−0.185	0.028
F4	−0.608	0.028	−0.406	0.025
F5	−0.464	0.031	−0.320	0.027
S-F	−0.670	0.019	−0.556	0.021

B. Single-factor model results

The results show good model fit for configural invariance pre- and post-test between the genders with all of the reported fit indexes being within acceptable ranges (pre; post: CFI = 0.9750; 0.9627, TLI = 0.9732; 9600, RMSEA = 0.0490; 0.0558). This suggests that a single-factor structure is consistent between the genders. Although the fit of this single-factor model is adequate, it is significantly worse than the fit of the EW5M model. This suggests that the EW5M model should be used when interpreting the FCI over a single-factor model.

Next, strong invariance can be seen to have had acceptable model fit (pre; post: CFI = 0.9731; 0.9591, TLI = 0.9720; 9575, RMSEA = 0.0500; 0.0575) and did not have a significant change in fit from configural invariance (pre; post: ΔCIF = −0.0020; −0.0036, ΔRMSEA = 0.0010; 0.0017).

Strict invariance is shown to not have adequate fit with the sample post instruction (CFI = 0.9475, TLI = 0.9474, RMSEA = 0.0640) and had significant change in fit from strong invariance (pre; post: ΔCIF = −0.00128; −0.0116, ΔRMSEA = 0.0096; 0.0065). This result is not surprising since strict invariance is seldom observed outside of simulations.

Modification indexes can be consulted for how to update the strict invariance model, thus creating a partial strict invariance model. It was found that the most impactful modification, pre- and post-test, would be to allow the female's single-factor variance to be freely estimated. This partial invariance model was found to have adequate model fit with the data (pre; post: CFI = 0.9687; 0.9555, TLI = 0.9677; 9554, RMSEA = 0.0537; 0.0589) and did not significantly differ from the fit of strong invariance (pre; post: ΔCIF = −0.00053; −0.0036, ΔRMSEA = 0.0037; 0.0014)

The latent means for this model can be compared, but, since strict invariance does not hold, cannot be interpreted in the same manner as a Cohen's d effect size. Under strong

invariance the latent means were estimated to be −0.670 and −0.556 for females pre- and post-test, respectively. These only serve to suggest that females are underperforming compared to males at the beginning and end of introductory physics instruction.

IV. DISCUSSION

The FCI under the EW5M factorization and the single-factor model was found to have strict invariance and partial strict invariance, respectively. This implies males and females are relating the items on the FCI in a similar manner, however, these relations could be due to surface features of the assessment. A previous study of the FCI on a different set of data found that this kind of dependence, if it was present, was likely small enough to not significantly impact factor analysis results [12]. Assuming that study's results apply to the sample used in this study suggests males and females are likely using the same latent traits when taking the FCI.

This implies that differences observed between the performances of the groups is not likely due to different, group specific factor structures of the assessment. Any differences in performance between the groups observed on the FCI while using the EW5M model and enforcing strict invariance could be a detection of true group differences; similarly for the single-factor model and the partial strict invariance.

When comparing the latent means of each group (i.e., the estimate of the center for each factor's scale) it was found that females underperformed compared to males across all factors on the EW5M model, under strict invariance. Recall, strict invariance fixes the male's latent means to zero and both groups' variances to one. Thus, the latent means estimate for the focus group in this manner gives an effect sizelike measure of the differences in group performance; see Table IV.

To control for the number of items on each factor, the latent means can be divided by the number of items on each factor. This gives each factor's latent mean difference per item on the pretest and post-test (pre; post) as: F1 = −0.130; −0.104, F2 = −0.092; −0.076, F3 = −0.127; 0.046, F4 = −0.101; −0.068, and F5 = −0.155; −0.107. From these results the most significant differences pretest exist on factors F1, F3, and F5, which represent Newton's first law plus kinematics, Newton's third law, and Force Superposition and Mixed Methods, respectively. After instruction it appears that factors F1 and F5 still show large differences per item, while F3's difference has almost entirely disappeared. This suggests that instruction is effectively closing the gender gap on Newton's third law, but not for Newton's first law plus kinematics and Force Superposition and Mixed Methods. The major factors females appear to be struggling with compared to males in this sample is Newton's first law plus kinematics and problems combining the simultaneous application of force superposition and Newton's first and second law plus kinematics.

Considering the change of the latent means per item from pre- to post-test gives differences of F1 = 0.026, F2 = 0.018, F3 = 0.081, F4 = 0.033, and F5 = 0.048. This shows that, of the five factors, Newton's first law plus kinematics and Newton's second law plus kinematics (F1 and F2, respectively) made the least progress towards equal group performance within this sample, particularly Newton's second law plus kinematics.

A study performed by Traxler *et al.*, examined differential item functioning (DIF) of the individual items on the FCI assuming a single-factor structure [7]. The single-factor model in the present study was shown to adequately fit student response data and had partial strict invariance between males and females. Assuming the results of the present study also apply to the sample used in Traxler *et al.* implies any detected DIF is likely due to group differences and not due to group dependent factor structure differences.

The two DIF methods used by Traxler *et al.* commonly identified items 14, 21, 22, 23, and 27 as having significantly different item difficulties, all with a small effect size. Comparing these to the EW5M model in Table I, shows all but item 23 belong to the Newton's second law plus kinematics factor. This suggests that Newton's second law plus kinematics may have been the major concept on which the males and females had significant performance differences within Traxler *et al.*'s samples.

Following the detection of DIF, Traxler *et al.* generated a purified FCI (i.e., all items with detectable DIF were removed iteratively) and showed that it managed to significantly reduce the gender gap for the exploratory group—sample 1 in that study. However, when this purified assessment was applied to the other samples, no significant reduction to the gender gaps were observed. This suggests that the observed DIF in that study were likely sample dependent and should be assessed using a different set of data.

Similarly, the results of the present study are assuredly sample dependent to some degree and should not be taken as generally applicable until more analyses have been performed on various samples. Future research should be directed to assess the measurement invariance and the DIF present on the FCI, as well as other major conceptual assessments used in physics education research. This would supply more evidence for or against the presence of measure invariance across gender on the FCI, and other assessments. Further, these analyses would help to direct the focus of gender research to identify any significant conceptual reasoning differences that may be present at the beginning of instruction. This could help drive pedagogical research whose focus would be to narrow any detected gender gaps.

V. LIMITATIONS

The results of this study are likely sample and factor model dependent. Because of this, the results presented

here should not be taken as generally applicable until more measurement invariance studies have been performed. A larger collection of measurement invariance studies would supply evidence to better support or refute the hypothesis of measurement invariance across genders being present on the FCI.

Further, the present study used the EW5M model and a single-factor model. As such, these results cannot be applied to different factor models for the FCI. More studies would need to be performed to assess the measurement invariance of the FCI under different factorizations.

VI. CONCLUSIONS

Using confirmatory factor analysis and measurement invariance techniques, it was found that males and females are likely interacting with the Force Concept Inventory in a similar manner from a latent construct perspective. Strict measurement invariance of the FCI under the Eaton and Willoughby five-factor modified model was demonstrated to hold for the sample used in this study. Whereas, strong invariance and only partially strict invariance were obtained for a single-factor model of the FCI. However, the lack of full strict invariance is common for real-world assessments.

These models suggested that females in the sample underperformed compared to males across all factors used to model the FCI pre- and post-instruction. However, the EW5M model under strict invariance indicated that these gaps did close slightly over the course of instruction. In fact, the gender gap almost vanished entirely for the factor said to probe Newton's third law. The factors interpreted as probing Newton's first law plus kinematics and Superposition and Mixed Methods contained the largest disparity between the genders per item. Further, Newton's first and second law plus kinematics were shown to have the smallest change in performance gap over the course of instruction. As a result, it is suggested that targeted pedagogical research be performed which specifically targets these concepts to close these gender gaps further.

This study supplies the first piece of evidence that the gender gap in performance on the FCI is likely not a result of different, gender specific factor structures of the assessment. Thus, the results of studies that have attempted to characterize the detected gender gap have gained some evidence that their results were likely not being influenced by the latent variable structure of the FCI. However, future studies should verify their data does possess measurement invariance between the genders to strengthen any conclusions made about detected performance gaps and subsequent changes in these gaps.

ACKNOWLEDGMENTS

The authors would like to thank Physport for supplying the student response data. This project was funded by Stockton University.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[2] 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).
[3] 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).
[4] 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).
[5] 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).
[6] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[7] 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).
[8] M. Prince, Does active learning work? A review of the research, J. Eng. Educ. **93**, 223 (2004).
[9] 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).
[10] 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).
[11] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).
[12] P. Eaton, B. Frank, and S. Willoughby, Detecting the influence of item chaining on student responses to the Force Concept Inventory and the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **16**, 020122 (2020).
[13] 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).
[14] 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).
[15] 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).
[16] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).
[17] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).
[18] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010137 (2018).
[19] P. Eaton and S. Willoughby, Identifying a preinstruction to postinstruction factor model for the Force Concept Inventory within a multitrait item response theory framework, Phys. Rev. Phys. Educ. Res. **16**, 010106 (2020).
[20] R. Padraic 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).
[21] 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).
[22] T. F. Scott and D. Schumayer, Central distractors in Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **14**, 010106 (2018).
[23] 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).
[24] M. Normandeau, S. Iyengar, and B. Newling, The presence of gender disparity on the Force Concept Inventory in a sample of Canadian undergraduate students, Can. J. Scholarship Teach. Learn. **8**, 9 (2017).
[25] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Proc. **1413**, 171 (2012).
[26] J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).
[27] M. Mears, Gender differences in the force concept inventory for different educational levels in the United Kingdom, Phys. Rev. Phys. Educ. Res. **15**, 020135 (2019).
[28] H. Wu and R. Estabrook, Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes, Psychometrika **81**, 1014 (2016).
[29] S. Sharma, S. Durvasula, and R. E. Ployhart, The analysis of mean differences using mean and covariance structure analysis: Effect size estimation and error rates, Organ. Res. Meth. **15**, 75 (2012).
[30] S. Sharma, S. Durvasula, and R. E. Ployhart, The analysis of mean differences using mean and covariance structure analysis: Effect size estimation and error rates, Organ. Res. Methods **15**, 75 (2012).
[31] R. E. Millsap and J. Yun-Tein, Assessing factorial invariance in ordered-categorical measures, Multivariate Behav. Res. **39**, 479 (2004).
[32] L. R. Price, Differential functioning of items and tests versus the Mantel-Haenszel technique for detecting differential item functioning in a translated test, in *ERIC—Proceedings the Annual Meeting of the American Alliance of Health Physical Education, Recreation and Dance, Boston, MA, April 12–16, 1999* (1999).
[33] Y. Xiao, G. Xu, J. Han, H. Xiao, J. Xiong, and L. Bao, Assessing the longitudinal measurement invariance of the Force Concept Inventory and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **16**, 020103 (2020).
[34] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. (The Guilford Press, New York, 2015).
[35] J.-i. Yasuda, H. Uematsu, and H. Nitta, Validating a Japanese version of the Force Concept Inventory, Lat. Am. J. Phys. Educ. **1**, 6 (2012), http://www.lajpe.org/icpe2011/16_Jun-Ichiro_Yasuda.pdf.
[36] B. P. Chapman, P. R. Duberstein, S. Sörensen, and J. M. Lyness, Gender differences in five factor model personality traits in an elderly cohort, Pers. Individ. Differ. **43**, 1594 (2007).
[37] J. Ock, S. T. McAbee, E. Mulfinger, and F. L. Oswald, The practical effects of measurement invariance: Gender invariance in two big five personality measures, Assessment **27**, 657 (2020).
[38] A. K. Edossa, U. Schroeders, S. Weinert, and C. Artelt, The development of emotional and behavioral self-regulation and their effects on academic achievement in childhood, Int. J. Behav. Dev. **42**, 192 (2018).
[39] L.-t. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Struct. Equation Model.: Multidiscip. J. **6**, 1 (1999).
[40] A. Alexander Beaujean, *Latent Variable Modeling Using R: A Step-By-Step Guide* (Routledge, London, 2014).
[41] F. F. Chen, Sensitivity of goodness of fit indexes to lack of measurement invariance, Struct. Equation Model.: Multidiscip. J. **14**, 464 (2007).
[42] S. B. McKagan, Physport, American Association of Physics Teachers, 2011. URL https://www.physport.org.
[43] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.
[44] Y. Rosseel, lavaan: An R package for structural equation modeling, J. Stat. Softw. **48**, 1 (2012).