Assessing the longitudinal measurement invariance of the Force Concept Inventory and the Conceptual Survey of Electricity and Magnetism

Yang Xiao,[1,2] Guiqing Xu,[1] Jing Han,[2] Hua Xiao,[1] Jianwen Xiong,[1,†] and Lei Bao[2,*]

[1] *South China Normal University, Guangzhou, Guangdong 510006, China*
[2] *The Ohio State University, Columbus, Ohio 43210, USA*

(Received 22 January 2020; accepted 12 June 2020; published 14 July 2020)

Concept inventories (CIs) are commonly used in pre-post instruction to study student conceptual change. For consistency in assessment interpretation, a CI's assessment construct is desired to maintain invariance across different test times. In this study, the longitudinal measurement invariance (LMI) analysis under the confirmatory factor analysis framework was used to examine the stability of the factor structure between pretest and post-test of two commonly used CIs, i.e., the Force Concept Inventory (FCI) and Conceptual Survey of Electricity and Magnetism (CSEM). A number of existing and modified models were examined in this paper. The results confirmed that all factor models of the FCI fitted well with both pre- and post-test data. For CSEM, acceptable fits were obtained with a reduced version of the CI. When reliability analysis was performed for the factors of these models, most modified models were found to be more reliable than the existing models. The modified models were further tested in LMI analysis, in which a sequence of models with increasingly restrictive parameter constraints was examined. For the FCI, LMI analysis demonstrated the existence of partial strict invariance, i.e., common factor structures, factor loadings, and item thresholds, and equally observed residual variances for all the items except items 2 and 29. For the CSEM, after excluding 10 items, a reduced version was found to hold the strict invariance criteria. These findings reveal that changes in scores of the whole FCI and the reduced CSEM can be attributed to changes in the latent constructs measured by the CIs, which confirms these two CIs as reliable instruments to study students' conceptual change over time in introductory physics courses.

DOI: 10.1103/PhysRevPhysEducRes.16.020103

I. INTRODUCTION

In physics education, the Force Concept Inventory (FCI) [1] and the Conceptual Survey of Electricity and Magnetism (CSEM) [2] have been widely used in pre- and postinstruction measurement to evaluate students' learning gain [3]. Popular methods to analyze the change in pre- and postmeasurement include pre-post score change [4] and the normalized gain [5–8].

From pre- to post-test, students' conceptual understandings often change significantly, which influence how students interact with the context and content of test items [9]. As a result, the latent construct measured by a concept inventory (CI) can also vary from pre- to post-test, which can pose a question on the consistency of data interpretation between pre- and post-test. Therefore, it is necessary to evaluate to what extent a CI can maintain invariance on its latent construct between different test times [10]. In empirical research, the changes of latent construct measured by a CI between pre- and post-test have been observed for several CIs [11–13]. The change of latent construct is also supported by the existing theories on conceptual change, which assume that students' conceptual understanding is restructured through instruction [14–17]. For example, the results of FCI at pretest show predominantly non-Newtonian views, while at post-test a significant portion of Newtonian views often emerge, which can also be in mixed states between Newtonian and non-Newtonian views depending on the instruction [1,9,12].

To study the possible changes of the latent constructs across test times, one may use the longitudinal measurement invariance (LMI) analysis [18,19], which evaluates the degree to which measurements of an instrument on a population at different test times yield measures of the same psychometric attributes. It has been suggested that the establishment of measurement invariance across test conditions is essential for evaluation of the effectiveness of a specific teaching method (e.g., tests of pre-post score

[*] Corresponding author.
bao.15@osu.edu
[†] Corresponding author.
jwxiong@scnu.edu.cn

change assessed by a CI) [20]. However, there exists limited research on evaluating the LMI of the CIs in science education, which may adversely impact interpretation of pre-post test data that are frequently used to evaluate the effectiveness of teaching [21]. This study contributes to supplement the literature on LMI analysis of the FCI and CSEM, and to establish empirical baselines for comparing the latent abilities or observed scores assessed by these CIs between pre- and post-test.

II. LITERATURE REVIEW

A. Students' conceptual structure change from fragmentation to integration

Over the past five decades, a large number of studies have demonstrated that students come into science classes holding a variety of preinstruction conceptions, which are inconsistent with the scientific concepts (see a comprehensive review in Ref. [16]). These preinstruction conceptions have also been labeled as *misconceptions* [22], *alternative conceptions* [23], *facets* [24], expressions of *phenomenological primitives* (*p* prims) [14], *children's initial explanatory framework* [17], etc., based on different theoretical perspectives of conceptual change, e.g., see Refs. [14,17,25].

Theories of conceptual change are generally characterized by one of two prominent but competing theoretical perspectives: (i) knowledge as theory, and (ii) knowledge as elements [15,26]. From the knowledge-as-theory perspective, researchers advocate that students' preinstruction conceptions could be considered as coherent theorylike thinking, e.g., see Refs. [17,25]. On the other hand, researches holding the knowledge-as-elements perspective argue that students' preinstruction conceptions are diverse and have no theoretical structure in any deep sense [15].

The debate about the structure of students' preinstruction conceptions has been going on for a long time, and recent studies have tried to find a middle ground between the two perspectives [9,27–31]. For example, Bao and his colleagues [9,29] revealed that students tend to progress from the "*consistent naive model states*" to the "*mixed model states*" (a combination of the correct and incorrect answers in different contexts), and finally to the "*consistent expert model states.*" Similarly, the knowledge integration perspective also assumes that, as students progress from lower to higher conceptual understanding levels, their knowledge structure becomes more integrated [30,32–35]. Recently, researchers in learning progression also characterized students' thinking into several gradually sophisticated levels, which may be completely inaccurate, entirely correct, or someplace in between [27,31].

What is in agreement among these previous studies is that students' preinstruction conceptions do not change easily under traditional instruction. Therefore, specifically designed teaching strategies are required to help students achieve conceptual change [36,37]. However, educators hold different opinions regarding to what degree should students' preinstruction conception be reconstructed. For instance, some researchers argue that students' conceptual structure should be fundamentally or radically reconstructed after effective instruction [14,38]. By contrast, other researchers consider the coexistence of preinstruction conceptions and scientific ones (e.g., student model state and model space [9], the conceptual profile change model [39], and the multiple knowledge system model [40,41]).

To sum up, although a consensus on the structures of students' preinstruction conceptions has yet to be achieved, it has been agreed that such structures are different from that of the experts and it can be reconstructed through instruction [14–17]. The structures of students' conceptual understanding can be analyzed based on student responses across a variety of items in a CI [42], thus, the latent construct measured by a CI can change from pretest to posttest. In other words, the possible pre-post changes of latent constructs of CI may impact how students' assessment results are interpreted. In the next section, the possible pre-post changes of latent constructs of FCI and CSEM will be discussed in detail.

B. The latent construct of FCI and CSEM

In science education, CIs have been widely used to assess student's conceptual changes from pre- to postinstruction [43–45]. Typically a CI is a research-based instrument in the multiple-choice form designed to assess a number of key concepts within the target content area or subject matter [46]. For each question or item, the answer choices include one correct option and several incorrect options (distractors), which are designed to elicit common student misconceptions.

Among the many CIs in physics, the FCI is the most widely used, which also provides an example for the development of other CIs [1]. The latent construct measured by the FCI has also been extensively studied. For example, Huffman and Heller [47] found different factor structures through principal component analysis of data, which showed a two-factor structure for a sample of 145 high school students and a one-factor structure for a sample of 750 university students. Recently, Scott, Schumayer, and Gray [48] again explored the factor structures of FCI through exploratory factor analysis (EFA) of post-test data from 2150 college students. Their results suggested that while a unidimensional construct is sufficient to explain the variance of the FCI data, a five-factor structure was superior [48].

In recent studies, statistical models from the item response theory (IRT) family have also been used to explore the latent construct measured by the FCI. Many studies have confirmed the unidimensionality of FCI using different IRT models, e.g., Rasch model [49], two-parameter logistic IRT model [48], and the three-parameter logistic IRT model [50].

Multidimensional IRT (MIRT) models have also been applied to study the possible multidimensional latent construct of the FCI [51,52]. For example, Scott and Schumayer [51] used the MIRT model to analyze a related dataset from Scott, Schumayer, and Gray [48] and confirmed that the five-factor structure was optimal.

In addition to the exploratory approach adapted by the studies mentioned above, some recent studies took the confirmatory approach to assess the multidimensional construct of the FCI, e.g., see Refs. [52,53]. For example, Eaton and Willoughby [53] applied confirmatory factor analysis (CFA) on a dataset from 20 822 students. The results confirmed the expertlike multidimensional model proposed by the creators of the FCI [1]. Meanwhile, results of CFA on smaller sample sizes suggested that the models of multidimensional construct proposed by Scott *et al.* [48] and Eaton and Willoughby [53] had more stable performances than the model proposed by Hestenes *et al.* [1]. Taking a confirmatory MIRT approach, Stewart *et al.* [52] tested a set of construct models using post-test data from 4716 college students. The optimal model suggested that the FCI can differentiate students' understanding among several related concepts, including Newton's 1st and 2nd laws, one-dimensional and three-dimensional kinematics, and addition of forces.

Similar studies have also been conducted on the CSEM, which is designed to assess introductory students' conceptual understandings of electricity and magnetism [2]. In the original paper, an 11-factor model was established, which is too many for a 32-item CI. Recently, two additional studies have tried to analyze the latent factor structures of the CSEM [54,55]. Zabriskie and Stewart [55] explored the factor structures of the CSEM through a MIRT approach using two datasets from different institutions, which produced two different optimal factor models, a nine-factor model and an eight-factor model, for the two populations, respectively. The results suggested that the optimal factor models identified in their study were not generalizable. In Eaton *et al.*'s study, a six-factor model identified using EFA was found to be more general as it was also validated with another population sample using CFA [54].

C. Changes of the latent constructs between pre- and post-test

In recent studies, the pre-post changes of the latent constructs of FCI have been partially explored [12,13]. For example, Semak *et al.* [13] applied EFA on 427 matched pre-post FCI data and identified a five-factor structure for the pretest data and a six-factor structure for the post-test data. The six-factor structure of the post-test was also suggested to be more aligned with that of an expert. In another study, Eaton *et al.* [12] examined the changes of non-Newtonian views as measured by the FCI from pre- to postinstruction. It was found that the coherent non-Newtonian views were similar in both pre- and post-test, which corroborated with the existing literature on that the construct of non-Newtonian views has been well established before instruction. Similar studies have also been conducted with other CIs. For example, Davenport [11] applied EFA on pre-post data from the Force Motion Conceptual Evaluation (FMCE) [56] and found that the FMCE data from pre- and post-test yielded six-factor structures, however, the nature of the factor structures was slightly different.

Summarizing the related literature, it appears that the latent constructs of a CI often change between pre- and post-test, which is consistent with existing theories of conceptual change. There is also evidence showing that constructs representing student naïve views often maintain from pretest to post-test, which are well documented by empirical studies in the literature. What often changes are the constructs representing expertlike knowledge, which are the learning goals of instruction. To thoroughly evaluate the possible pre-post changes of the latent constructs of CIs, longitudinal measurement invariance (LMI) is introduced in the next section.

D. Evaluation of the longitudinal measurement invariance

Measurement invariance refers to the degree to which measurements of an instrument under different conditions yield measures of the same psychometric attributes [57]. In cross-section analysis, measurement invariance across different groups can be used to examine whether the latent construct measured by a specific instrument is the same or not across population samples, which provides a function for detecting possible structural bias. In the longitudinal analysis (e.g., pre- and post-test), measurement invariance of tests conducted at different times with the same population is evaluated, which is referred to as LMI.

The most widely used method to evaluate measure invariance is the confirmatory factor analysis (CFA) [20,58,59]. In CFA, a student's response on an item is modeled as the sum of the product of a latent variable and a factor loading, an item intercept, and some residual error for the item [60]. Following this relationship, Widaman and his colleagues suggested a four-level scale to evaluate measurement invariance, i.e., configural, metric, scalar, and strict invariance [18,19]. The configural invariance only requires the same measurement pattern with freedom in factor loadings across different conditions. The metric invariance, which is also called weak factor invariance, further requires identical factor loadings across different conditions on the basis of configural invariance. However, the configural and metric invariance are insufficient to guarantee the comparability of observed or latent scores across time or groups. To do this, the scalar and strict invariance must hold across different conditions. The scalar invariance, also called strong factor invariance, requires both invariant factor loadings and invariant intercepts across time. The strict factor invariance further requires identical item residual variances over scalar invariance.

The establishment of measurement invariance has been suggested as necessary evidence of validation for assessment instruments in science education [20]. This study is conducted to examine whether the latent constructs measured by the FCI and CSEM may change between pre- and post-test under the CFA framework. While the latent constructs of the FCI and CSEM are still undetermined, their factor structures are first reexamined using the dataset collected in this study. After establishing construct validities of the two CIs, LMI analysis is conducted with the pre-post data of each of the CIs. In addition, since student knowledge has also developed from pretest to post-test, the influence of student knowledge states on LMI will also be explored. Specifically, the following questions are studied:

1. What are the best-fit models of latent construct measured by the two CIs at pre-test and post-test?
2. To what extent do the LMI hold for each of the two CIs from pre- to post-test?
3. To what extent do the LMI vary in each of the two CIs as a result of changes in students' understandings from pre- to post-test.

III. METHODOLOGY

A. Participants and instruments

This study investigates the factor structures and LMI of the FCI (1995 version) and CSEM (form H). The dataset of the FCI was collected in a large state university with a U.S. ranking of top 60 and an acceptance rate of 52% ($N_{\mathrm{FCI}} = 474$). The dataset of the CSEM was collected in another large state university with a U.S. ranking of the top 100 and an acceptance rate of 60% ($N_{\mathrm{CSEM}} = 499$).

For both the FCI and CSEM samples, students were enrolled in the related calculus-based introductory physics courses. These students took the FCI or CSEM as a pretest during the first week of the course and as a post-test during the week before the final exam. For the dataset of FCI, students' mean score was 33.77% (Cronbach's $\alpha = 0.760$) at pretest and 45.56% (Cronbach's α=0.825) at post-test. For the dataset of CSEM, students' mean score was 27.38% (Cronbach's $\alpha = 0.543$) at pretest and 45.70% (Cronbach's $\alpha = 0.748$) at post-test. The main goal of this study is to evaluate the LMI of the two CIs. The pre-post test scores provided here give the background information of the student population, since the LMI is a result of both the characteristics of the population and the instrument. Hence, it is important to understand the limitation of this type of study that any analysis outcomes from any study are dependent on both the population and the instrument, and should not be considered as a feature of the instrument only.

B. Models of factor structures

For the FCI, three models of factor structures identified and validated in previous studies [48,53] were fitted to the pre- and post-test data (see the first three models in Table I, namely, HWS6, SSG5, and EW5). Among these, the HWS6 was adapted from the expertlike measurement model proposed by the creators of the FCI [1], the SSG5 was proposed by Scott *et al.* [48] through EFA, and the EW5 was suggested by Eaton and Willoughby [53] through expert considerations of the items on the FCI. In the model fit computation, some items have been removed due to poor model fit. All the factor names in these three models were analogous to those given in Eaton and Willoughby [53].

To represent more complete factor structures of the full FCI, three modified models have been proposed in this study, in which the items removed in previous studies are reassigned to different dimensions based on experts' considerations (see the last three models in Table I). For easy comparisons, the factors in the modified models are given almost identical names based on the original models. To establish clear factor structures, some cross-loading items in the new models, e.g., item 26 in the F2 "2nd law" and F6 "superposition" of the HWS6 model, are only loaded to one factor based on experts' views. Meanwhile, some items, e.g., items 1, 2, and 3 in the F5 "forces" of the HWS6 model are reassigned into a different more suitable dimension according to experts' considerations.

Regarding the factor structures of CSEM, there exists only one study, which identified an EFA driven 6-factor model labeled as Eaton6 shown in Table II [54]. Among the factors, F1 concerns students' understandings of action or reaction pairs in the context of electricity and magnetism. F2 addresses the electric forces on a charge given a charge distribution or an external electric field. F3 contains two items, which probe students' understandings of the magnetic field due to current-carrying wires. F4 includes questions that probe students' understandings of the magnetic force on charged particles. F5 is related to students' understandings of electric force or field in the context of an electric potential difference. F6 concerns relations between electric fields, electric forces, and electric potentials. With this six-factor model, item 19 is cross loaded in both F5 and F6, which is an undesired outcome, and will be addressed in a modified model.

When the Eaton6 model was fitted to the pre- and post-test CSEM data of this study, the overall fit statistics were acceptable, while the reliabilities of the fourth and fifth factors was found questionable with the pretest data (see more details Sec. IV. C). Hence, a slightly modified model, Eaton5M, was proposed (Table II). In the modified model, the original F5 of Eaton6 is removed due to low reliability. For the three items associated with F5 in Eaton6, item 18 is excluded from the new model due to a negative estimate of the factor loading, and items 19 and 20 are reassigned into

TABLE I. The six factor models of FCI examined in this study. Note that in the HWS6 model, item 29 was excluded; in the SSG5 model, items 1, 2, 3, and 29 were excluded; and in the EW5 model, items 1, 2, 3, 9, 14, and 26 were excluded.

Model	Factor		Items
HWS6	F1	Kinematics	12, 14, 19, 20, 21
	F2	2nd Law	9, 22, 26, 27
	F3	1st Law	6, 7, 8, 10, 23, 24
	F4	3rd Law	4, 15, 16, 28
	F5	Forces	1, 2, 3, 5, 11, 13, 18, 25, 30
	F6	Superposition	17, 25, 26
SSG5	F1	Identification of forces	5, 11, 13, 18, 30
	F2	1st law with 0 force	6, 7, 8, 10, 12, 16, 24, 29
	F3	2nd law with kinematics	19, 20, 21, 22, 23, 27
	F4	1st law with canceling forces	16, 17, 25
	F5	3rd law	4, 15, 28
EW5	F1	1st law + kinematics	6, 7, 8, 10, 20, 23, 24
	F2	2nd law + kinematics	9, 12, 14, 19, 21, 22, 27
	F3	3rd Law	4, 15, 16, 28
	F4	Force identification	5, 11, 13, 18, 30
	F5	Mixed	17, 25, 26
HWS6M	F1	Kinematics	1, 2, 3, 12, 14, 19, 20, 21
	F2	2nd Law	9, 22, 27
	F3	1st Law	6, 7, 8, 10, 23, 24
	F4	3rd Law	4, 15, 16, 28
	F5	Identification of forces	5, 11, 13, 18, 25, 29, 30
	F6	Superposition	17, 25, 26
SSG5M	F1	Identification of forces	5, 11, 13, 18, 30
	F2	1st law with 0 force	6, 7, 8, 10, 12, 24, 29
	F3	2nd law with kinematics	1, 2, 3, 9, 14, 19, 20, 21, 22, 23, 27
	F4	Superposition	17, 25, 26
	F5	3rd law	4, 15, 16, 28
EW5M	F1	1st law + kinematics	6, 7, 8, 10, 20, 23, 24
	F2	2nd law + kinematics	1, 2, 3, 9, 12, 14, 19, 21, 22, 27
	F3	3rd law	4, 15, 16, 28
	F4	Identification of forces	5, 11, 13, 18, 30, 29
	F5	Superposition	17, 25, 26

TABLE II. The two factor models of CSEM examined in this study. Note that in the Eaton6 model, items 1, 2, 14, 22, 28, 30, and 32 were excluded and in the Eaton5M model, items 1, 2, 13, 14, 18, 21, 22, 28, 30, and 32 were excluded.

Model	Factor		Items
Eaton6	F1	Newton's 3rd law	4, 5, 7, 24
	F2	$\vec{F}_E = q\vec{E} +$ Superpos.	8, 6, 9, 12, 3, 17, 16
	F3	$\vec{B}$ by I	23, 26
	F4	$\vec{F}_B = q\vec{v} \times \vec{B}$	21, 27, 25, 29, 31, 16, 13
	F5	$\vec{E} = -\nabla V$	20, 18, 19
	F6	q in fields	11, 15, 10, 19
Eaton5M	F1	Newton's 3rd law	4, 5, 7
	F2	$\vec{F}_E = q\vec{E} +$ Superpos.	8, 6, 9, 12, 3
	F3	$\vec{B}$ by I	23, 26
	F4	Electromagnetic interaction	27, 25, 29, 31, 24
	F5	q in fields	11, 15, 10, 19, 20, 17, 16

F5, which targets the concept of force caused by an electric field (as suggested by the creators of the CSEM [2]).

In addition, items 16 and 17 are reassigned to F5, since these were loaded poorly on F2 in Eaton6 [54], and were also suggested by the creators of the CSEM as probing student understandings similar to those measured by the other items in F5 of Eaton5M [2]. With these changes, F5 in Eaton5M is conceptually interpreted as relations between electric fields, electric forces, and electric potentials, which is similar to the original F6 in Eaton6 but with a few more items included.

Finally, F4 in the new model targets student understandings of electromagnetic interaction, which keeps four items (27, 25, 29, 31) from the original F4 in Eaton6, and adds item 24 from the original F1. Item 24 is moved to F4 from the original F1 because it probes student understandings of electromagnetic interaction regarding the forces between two current-carrying wires. Items 16 and 13 are removed from the new F4 since they probe understandings of electrostatics instead of magnetism [54].

To sum up, a total of 10 items are excluded in the new model, while 7 items were excluded in the original Eaton6 model. Then, the LMI analysis will only be conducted for the reduced version of CSEM using the Eaton5M model (see more details in Table II). Future studies are suggested to be conducted to reexamine these models with fewer items excluded using different datasets.

C. Analysis

All analysis is conducted using the R packages *lavann* [61] and *semTools* [62]. Specifically, the *lavann* package is used to conduct the CFA and following LMI analysis, and the *semTools* is used to compute the composite reliability [(CR), much like the Cronbach's α] for a CI and its different subscales. For continuous data, the maximum likelihood (ML) estimator is commonly used for CFA. In this study, the weighted least squares mean and variance-adjusted (WLSMV) estimator is used for the ordered categorical data collected by the CIs. It has been shown that, for categorical data, the WLSMV estimator is superior to the ML estimator in terms of both the model rejection rates and the appropriate estimation of factor loadings [63,64]. The comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA) are used to determine an acceptable model fit: CFI ≥ 0.95, TLI ≥ 0.95, RMSEA ≤ 0.08 [65].

After the factor structures of the two CIs are identified for both pre- and post-test data, a set of longitudinal CFAs are applied to examine the LMI between pre- and post-test. Within the longitudinal CFA approach, data of different tests are integrated into one model in which the residuals of the same item at different test times are allowed to covary over time. Because all items in the FCI or CSEM are scored dichotomously, the data must be considered as ordered categorical. For categorical data, the factor loadings and thresholds must be varied in tandem [66], which further makes the steps of LMI testing different from those with continuous data. That is, the step of metric measurement invariance testing is dropped in the procedure of testing LMI with categorical data [67]. Accordingly, Edossa *et al.* [67] have clearly described the procedure of testing LMI with categorical data, which includes a sequence of models with increasingly restrictive constraints, i.e., configural invariance, strong invariance, and strict invariance. The necessary parameter restrictions in the testing procedure for LMI with categorical data are summarized in Table III.

Historically, the Satorra-Bentler-scaled chi-square statistic (SB-χ^2) has been used to test measurement invariance by comparing the difference of the fit between models [68]. However, the χ^2 difference test has been criticized due to its sensitivity to sample size [69]. As a result, different statistics have been recommended and used in this study [70], with which a condition of ΔCFI > 0.01 and ΔRMSEA > 0.015 between two consecutive models (e.g., the configural and strong invariance model) is considered unacceptable to establish measurement invariance.

IV. RESULTS

A. Factor structures of the FCI

The six factor models of the FCI (in Table I) were fitted to the pre- and post-test data separately. The fit statistics are presented in Table IV. All the six models show adequate fit to the data in both pre- and post-test judging by $\chi^2/\mathrm{df}(\leq 3)$, CFI ($\geq 0.95$), TLI ($\geq 0.95$), and RMSEA ($\leq 0.08$) [65]. For the pre-test data, it was found that the EW5 model showed the best fit to the data judging by all fit statistics, i.e., lowest

TABLE III. Testing for longitudinal measurement invariance with categorical data. The asterisk (*) indicates that the parameter is freely estimated across test time points. Fixed = the parameter is fixed to equity in the pre- and post-test. Fixed at $0/0$ = factor means parameters are fixed at 0 at both pre- and post-test. Fixed at $0/*$ = factor means parameters are fixed at 0 at pretest and freely estimated at post-test. Fixed at $1/1$ = the residual variances are fixed at 1 at both pre- and post-test. Fixed at $1/*$ = the residual variances are fixed at 1 at pre-test and freely estimated at post-test.

	Factor loadings	Thresholds	Residual variances	Factor means
Configural invariance	*	*	Fixed at $1/1$	Fixed at $0/0$
Strong invariance	Fixed	Fixed	Fixed at $1/*$	Fixed at $0/*$
Strict invariance	Fixed	Fixed	Fixed at $1/1$	Fixed at $0/*$

TABLE IV. Fit statistics for the six models of FCI applied to the pre- and post-test datasets separately. The reliability estimates for the subscales and the whole FCI were computed using CR. Values of CR below the minimally acceptable level (0.65) were underlined.

	Test							RMSEA	Composite reliability						
Model	occasion	χ^2	df	χ^2/df	P	TLI	CFI	(90% Upper CI)	F1	F2	F3	F4	F5	F6	Total
HWS6	Pre	432.195	360	1.201	0.000	0.972	0.975	0.021 (0.027)	0.622	0.439	0.723	0.777	0.767	0.800	0.873
	Post	633.086	360	1.759	0.000	0.954	0.959	0.040 (0.045)	0.629	0.608	0.786	0.785	0.829	0.826	0.905
SSG5	Pre	270.463	241	1.122	0.093	0.985	0.987	0.016 (0.025)	0.844	0.679	0.609	0.636	0.819	/	0.857
	Post	385.329	241	1.599	0.000	0.967	0.971	0.036 (0.042)	0.882	0.748	0.691	0.740	0.815	/	0.887
EW5	Pre	318.193	289	1.101	0.114	0.987	0.989	0.015 (0.024)	0.742	0.621	0.777	0.844	0.800	/	0.869
	Post	512.296	289	1.773	0.000	0.957	0.962	0.040 (0.046)	0.804	0.677	0.785	0.882	0.826	/	0.902
HWS6M	Pre	438.984	389	1.128	0.041	0.982	0.984	0.016 (0.024)	0.670	0.395	0.723	0.777	0.806	0.800	0.875
	Post	638.421	389	1.641	0.000	0.958	0.963	0.037 (0.042)	0.672	0.450	0.786	0.785	0.822	0.826	0.902
SSG5M	Pre	441.160	395	1.117	0.054	0.983	0.985	0.016 (0.023)	0.844	0.697	0.709	0.800	0.777	/	0.875
	Post	647.553	395	1.639	0.000	0.958	0.962	0.037 (0.042)	0.882	0.743	0.746	0.826	0.785	/	0.902
EW5M	Pre	457.787	395	1.159	0.016	0.977	0.979	0.018 (0.025)	0.742	0.657	0.777	0.814	0.800	/	0.875
	Post	651.720	395	1.650	0.000	0.958	0.962	0.037 (0.042)	0.804	0.700	0.785	0.812	0.826	/	0.902

χ^2/df with $p = 0.114$, highest CFI and TLI, and lowest RMSEA. From pre- to post-test, model performance became worse for all six models according to all fit statistics. That is, these models seem to provide more accurate descriptions of the latent constructs of the FCI in the pretest than that in the post-test. For post-test data, the best performing model was the SSG5 judging by χ^2/df, CFI, TLI, and RMSEA.

In contrast to the model fit statistic, the total and component reliability estimates increased from pre- to post-test, which indicates that the influence of measurement error in test scores is smaller after instruction. The total reliability estimates of all six models are in the range of good (≥0.80) to excellent (≥0.90) [71]. The reliabilities of the three modified models are slightly better than those of the three original models from the literature due to the removal of certain items in the original models, which decreases the number of total items.

However, the reliability estimates for some factors of the three original models show unacceptable values of composite reliability. The HWS6 model has the worst performance in reliability: the values of CR for factors F1 and F2 are below the minimally acceptable level in both pre- (0.622 and 0.629 for F1 and F2, respectively) and post-test (0.439 and 0.608 for F1 and F2, respectively). The other two models only show some unreliable measures in the pretest: the SSG5 model has two unreliable factors, F3 (0.609) and F4 (0.636), and the EW6 model has one, F2 (0.621). After examining the items that contribute to reliability issues of these factors, it has been found that these items mainly address Newton's second law and its resulting kinematics (see Table I for details). These items have been adjusted in the three modified models. After the adjustment, all the reliability estimates of all three modified models have achieved the minimally acceptable level except for F2 in HWS6M. As a result, the subsequent LMI analysis of FCI will focus on the two models containing all items with reliable measurement for all subscales, i.e., the SSG5M and EW5M model.

B. Longitudinal measurement invariance of the FCI

The procedure of LMI evaluation of FCI consists of a sequence of three models with increasingly restrictive parameter constraints, i.e., configural invariance, strong invariance, and strict invariance. The model fit deterioration between two consecutive models are examined using changes of CFI and RMSEA. In this part of the study, two of the modified models, SSG5M and EW5M, are examined, and the results are shown in Table V.

SSG5M is a five-factor model for the FCI. The first step of LMI testing is to examine configural invariance. For the configural invariance model, all the factor loadings and thresholds are estimated without constraints, and the residual variances are fixed for identification purposes. For SSG5M, the configural invariance model produces an acceptable model fit ($\chi^2/\mathrm{df} = 1.211$, $\mathrm{TLI} = 0.980$, $\mathrm{CFI} = 0.981$, $\mathrm{RMSEA} = 0.021$). This result suggests that FCI shares similar factor structures between pre- and post-test. In the second step, the strong invariance model is applied, in which the factor loadings and thresholds are set to be identical between pre- and post-test, and the residual variances of the indicators are estimated without constraints. The result still indicates an acceptable model fit ($\chi^2/\mathrm{df} = 1.296$, $\mathrm{TLI} = 0.971$, $\mathrm{CFI} = 0.973$, $\Delta\mathrm{CFI} = -0.008$, $\Delta\mathrm{RMSEA} = 0.004$). The third step applies the strict measurement invariance model, in which the factor loadings, thresholds, and residual variances are constrained to be identical between pre- and post-test. The result (Table V, SSG4M strict invariance model) shows an adequate fit with $\chi^2/\mathrm{df} = 1.409$, $\mathrm{TLI} = 0.960$, $\mathrm{CFI} = 0.962$, $\Delta\mathrm{CFI} = -0.011$,

TABLE V. Model fit statistics of the SSG5M and EW5M models of the FCI for longitudinal measurement invariance restrictions.

Model	χ^2	df	χ^2/df	p	TLI	CFI	ΔCFI	RMSEA (90% Upper CI)		ΔRMSEA
SSG5M										
Configural	1979.854	1635	1.211	0.000	0.980	0.981	/	0.021	(0.024)	/
Strong	2145.040	1655	1.296	0.000	0.971	0.973	−0.008	0.025	(0.028)	0.004
Strict	2373.779	1685	1.409	0.000	0.960	0.962	−0.011	0.029	(0.032)	0.004
Partial Strict[a]	2330.547	1684	1.384	0.000	0.963	0.965	−0.008	0.028	(0.031)	0.003
EW5M										
Configural	2019.595	1635	1.235	0.000	0.977	0.979	/	0.022	(0.025)	/
Strong	2151.781	1655	1.300	0.000	0.971	0.973	−0.006	0.025	(0.028)	0.003
Strict	2415.401	1685	1.433	0.000	0.958	0.960	−0.013	0.030	(0.033)	0.005
Partial Strict A[a]	2376.142	1684	1.411	0.000	0.960	0.962	−0.011	0.029	(0.032)	0.004
Partial Strict B[b]	2351.849	1683	1.397	0.000	0.961	0.963	−0.010	0.029	(0.032)	0.004

[a] Freely estimated variance for item 29 at the post-test.
[b] Freely estimated variance for items 2 and 29 at the post-test.

ΔRMSEA = 0.029. However, this model appears to have a significant deterioration of model fit based on the change of CFI (ΔCFI = −0.011). For all the models, the residual invariance changes are quite small, as indicated by ΔRMSEA.

Following the suggestion from Cheung and Rensvold [72], the modification indices (MIs) from the strict invariance model are evaluated along with the factor-ratio test, which can be used to evaluate which indicators of the SSG5M model are responsible for the small residual invariance changes. The results indicate that the residuals of item 29 in factor F2, "1st law with 0 force," differ between pre- and post-test (MI = 35.116). Therefore, a partial strict invariance model with freely estimated variance for item 29 at the post-test is fitted, which is less restrictive than the restrictions imposed in the fully strict invariance model. The partial strict invariance model reveals little degrading of fit compared to the strong invariance model ($\chi^2/\mathrm{df} = 1.384$, TLI = 0.963, ΔCFI = −0.008, ΔRMSEA = 0.003).

The final results of partial strict LMI analysis of FCI across pre- and post-test using the factor structures suggested by the SSG5M model is presented in Fig. 1, which shows that the standardized factor loadings ranged from 0.222 to 0.908 at the pretest and from 0.211 to 0.930 at the post-test. The intercorrelations between the five factors were moderate, which ranged from 0.307 to 0.851 at pretest and from 0.389 to 0.745 at post-test. It is noted that from pre- to post-test the majority of intercorrelations between the five factors show slight increases except for correlations between F2 and F3 and between F4 and F5. Meanwhile, from pre- to post-test the correlations of the same factors in different times range from 0.534 to 0.963, which are larger than the correlations between different factors in different test times for each of the five factors.

EW5M is also a five-factor model for FCI. The differences between EW5M and SSG5M are the assignments of four items, i.e., items 12, 20, 23, and 29 (see Table I). Despite these differences, the LMI analysis of EW5M yields very similar results. Again, the configural invariance model produces an acceptable fit statistic ($\chi^2/\mathrm{df} = 1.235$, TLI = 0.977, CFI = 0.979, RMSEA = 0.022), and the strong measurement invariance does not show a meaningful deterioration in a model fit ($\chi^2/\mathrm{df} = 1.300$, TLI = 0.971, ΔCFI = −0.006, ΔRMSEA = 0.003). The strict measurement invariance model results in a significant deterioration of model fit (ΔCFI = −0.013). Again, item 29 is responsible for the lack of residual invariance between pre- and post-test (MI = 28.355). However, after releasing the variance for item 29 from the post-test data, the partial strict invariance model (model A) still shows significant deterioration of model fit (ΔCFI = −0.011). According to the MIs from the partial strict invariance model A, item 2 is identified as responsible for the lack of residual invariance (MI = 24.181), which leads to the making of a second partial strict invariance model (model B), in which the variances are freely estimated for both item 2 and item 29 at post-test in addition to the restrictions imposed in the fully strict invariance model. The results of model B do not show a meaningful degrading of fit compared to the strong invariance model ($\chi^2/\mathrm{df} = 1.397$, TLI = 0.961, ΔCFI = −0.010, ΔRMSEA = 0.004). Following the five-factor EW5M model, the FCI is also found to maintain a partial strict measurement invariant between pre- and post-test.

The final EW5M mapping with factor loadings and correlations is shown in Fig. 2, which reveals that the majority of the standardized factor loadings show slight increases from pretest to post-test except for item 29. The intercorrelations between the five factors are moderate, which ranged from 0.319 to 0.875 at pretest and from 0.395 to 0.787 at post-test. From pre- to post-test, the majority of intercorrelations between the five factors show slight increases except for the correlations between F1 and F2 and between F3 and F5. Finally, from pre- to post-test, the correlations of the same factor in different times range from 0.534 to 0.935, which are larger than the correlations among different latent factors in different test times for each of the five factors.

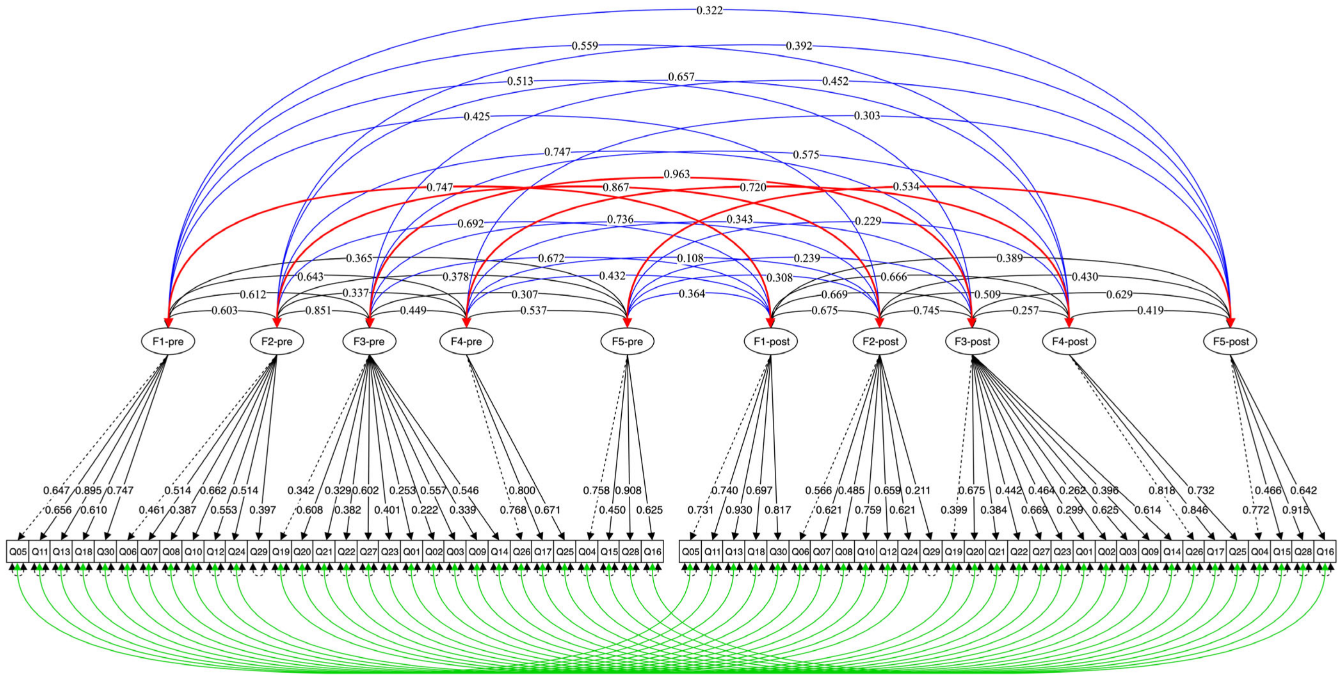

FIG. 1. Partial strict LMI model of the FCI (Freely estimated variance for item 29 at the post-test time). The modified SSG5M model suggested the following five factors: F1 = Identification of Forces, F2 = 1st Law with 0 Force, F3 = 2nd Law with Kinematics, F4 = Superposition, and F5 = 3rd law. Because of the data's longitudinal structure, the latent factors are allowed to correlate across time (the correlation between identical latent factors in the different time was in red curves; the correlation between different latent factors in the different time was in blue curves), and identical items over the two test occasions were allowed to covary except for item 29 (green curves). All parameters are standardized.

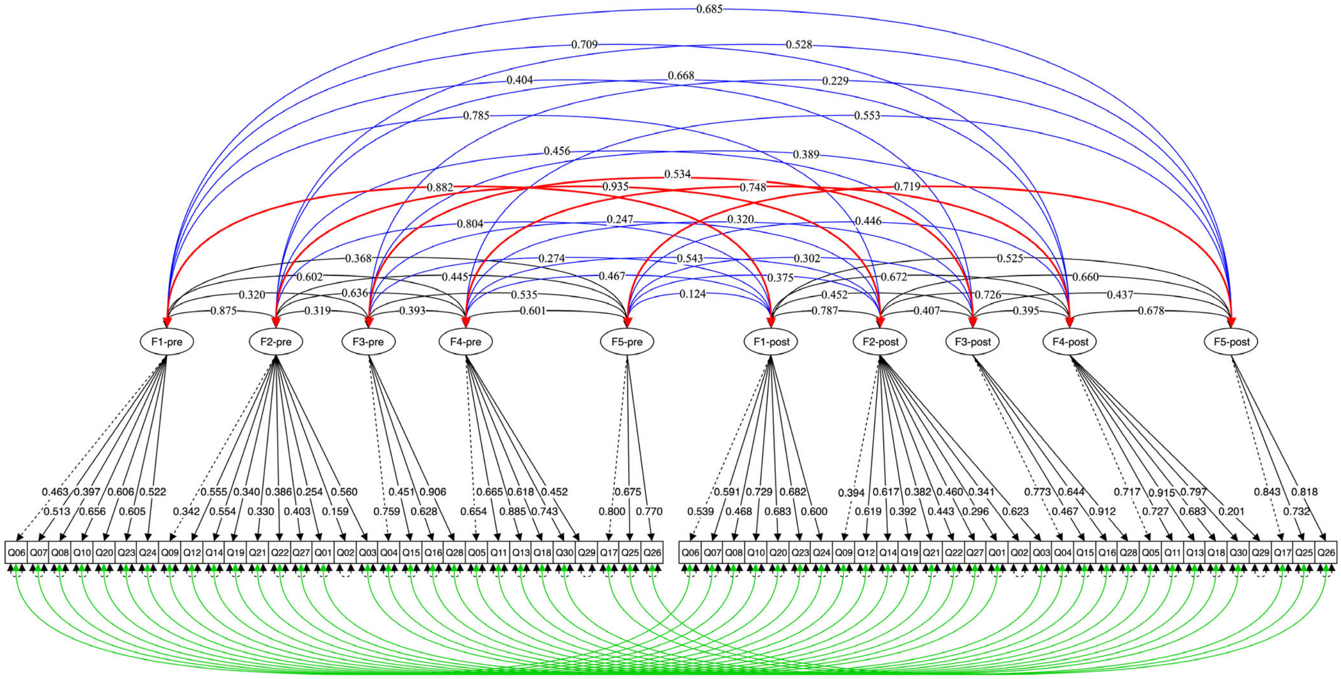

FIG. 2. Partial strict LMI model of the FCI (Freely estimated variance for items 2 and 29 at post-test time). The modified EW5M model suggested the following five factors: F1 = 1st Law + Kinematics, F2 = 2nd Law + Kinematics, F3 = 3rd Law, F4 = Identification of forces, and F5 = Superposition. Because of the data's longitudinal structure, the latent factors are allowed to correlate across time (the correlation between identical latent factors in the different time was in red curves; the correlation between different latent factors in the different time was in blue curves), and identical items over the two test occasions were allowed to covary except for items 2 and 29 (green curves). All parameters are standardized.

C. Factor structures of CSEM

In this part of the study, the factor structures of CSEM are analyzed with the Eaton6 model [54] and the modified Eaton5M model (see Table II). The results of fit statistics are presented in Table VI, which shows that the two models have adequate fits in both pre- and post-test judging by χ^2/df (≤ 3), CFI (≥ 0.95), TLI (≥ 0.95), and RMSEA (≤ 0.08) [65]. In addition, for the Eaton5M model, the p values of chi squares are larger than 0.05 in both pre- and post-test, which indicate that the model fits the data nearly perfectly.

The total reliability estimates also suggest that the modified Eaton5 model is more reliable than the original Eaton6 model. Although the Eaton5M has two fewer items than Eaton6 does, the values of CR of Eaton5M are higher than that of Eaton6 in both pre- (0.710 vs 0.678) and post-test (0.872 vs 0.866). Regarding the individual factors, Eaton5M has larger reliability indexes on most factors (except for F2) than Eaton6 does in both pre- and post-test. For example, the reliability of F1 in the pretest has been improved from a minimally acceptable level (0.648 in Eaton6) to a respectable level (0.819 in Eaton5M) after excluding item 24. These results suggest that for the data in this study, Eaton5M provides better measurement consistency than Eaton6.

However, even after the modification, Eaton5M still has several factors with unacceptable reliability (<0.60) at pretest. For example, three items (17, 16, 20) are added to the factor "q in fields" (F6 in Eaton6 model and F5 in Eaton5M model), but the reliability index is only slightly improved from 0.563 to 0.566 in the pretest. Meanwhile, for F4, the internal consistency has improved from 0.197 to 0.360, but the reliability estimate is still unacceptable. These results suggest that these factors may introduce inconsistency of measures in the pretest, which should be further studied in future research.

Since Eaton5M appears to be a more reliable model over the original Eaton6 model, the subsequent LMI analysis of CSEM will be conducted with the Eaton5M only.

D. Longitudinal measurement invariance of the CSEM

The LMI analysis is conducted for the reduced CSEM aligned with the Eaton5M model, and the results are shown in Table VII. The configural invariance model produces an acceptable model fit ($\chi^2/\mathrm{df} = 1.056$, $p = 0.125$, $\mathrm{TLI} = 0.993$, $\mathrm{CFI} = 0.994$, $\mathrm{RMSEA} = 0.011$). This result suggests that the reduced CSEM shares similar factor structures between pre- and post-test. The strong invariance model (common factor loadings and thresholds across pre- and post-test) also fits at an acceptable level ($\chi^2/\mathrm{df} = 1.118$, $p < 0.01$, $\mathrm{TLI} = 0.987$, $\mathrm{CFI} = 0.987$, $\mathrm{RMSEA} = 0.015$). The changes of CFI and RMSEA reveal little deterioration in model fit ($\Delta\mathrm{CFI} = -0.007$, $\Delta\mathrm{RMSEA} = 0.004$). Fit results for the strict invariance model (common factor loadings, thresholds and residual variances across pre- and post-test) also indicate an acceptable fit ($\chi^2/\mathrm{df} = 1.192$, $p = 0.125$, $\mathrm{TLI} = 0.976$, $\mathrm{CFI} = 0.978$, $\mathrm{RMSEA} = 0.020$). Again, the changes of CFI and RMSEA indicate no meaningful deterioration in model fit ($\Delta\mathrm{CFI} = -0.009$, $\Delta\mathrm{RMSEA} = 0.005$).

Therefore, following the five-factor solution suggested by the Eaton5M model, the reduced CSEM has been found to hold strict measurement invariance across pre- and post-test (see Fig. 3). As shown in Fig. 3, all the standardized factor loadings have slight increases from pretest to post-test. Unlike the two factor models of FCI, the intercorrelations among the five factors of CSEM all have positive

TABLE VI. Fit statistics for the two models of CSEM applied to the pre- and post-test datasets separately. The reliability estimates for the subscales identified by different models and the whole FCI were computed using CR. Values of CR below the minimally acceptable level (0.65) were underlined, and the unacceptable values of CR (<0.60) were further marked with an asterisk (*).

										Composite reliability						
Model	Test occasion	χ^2	df	χ^2/df	P	TLI	CFI	RMSEA	(90% Upper CI)	F1	F2	F3	F4	F5	F6	Total
Eaton6	Pre	273.866	258	1.061	0.000	0.983	0.986	0.011	(0.022)	$\underline{0.648}$	0.663	0.688	$\underline{0.068}^*$	$\underline{0.197}^*$	$\underline{0.563}^*$	0.678
	Post	340.137	258	1.318	0.000	0.976	0.980	0.025	(0.032)	0.841	0.725	0.753	0.650	$\underline{0.500}^*$	0.689	0.866
Eaton5	Pre	227.643	199	1.144	0.080	0.969	0.973	0.017	(0.027)	0.819	$\underline{0.626}$	0.688	$\underline{0.360}^*$	$\underline{0.566}^*$	/	0.710
	Post	229.260	199	1.152	0.070	0.991	0.992	0.017	(0.027)	0.880	0.658	0.753	0.655	0.730	/	0.872

TABLE VII. Model fit statistics of the Eaton5M model of CSEM for longitudinal measurement invariance restrictions.

Model	χ^2	df	χ^2/df	P	TLI	CFI	ΔCFI	RMSEA	(90% Upper CI)	ΔRMSEA
Configural	882.143	835	1.056	0.125	0.993	0.994	/	0.011	0.017	/
Strong	946.533	847	1.118	0.009	0.985	0.987	−0.007	0.015	0.021	0.004
Strict	1036.036	869	1.192	0.000	0.976	0.978	−0.009	0.020	0.024	0.005

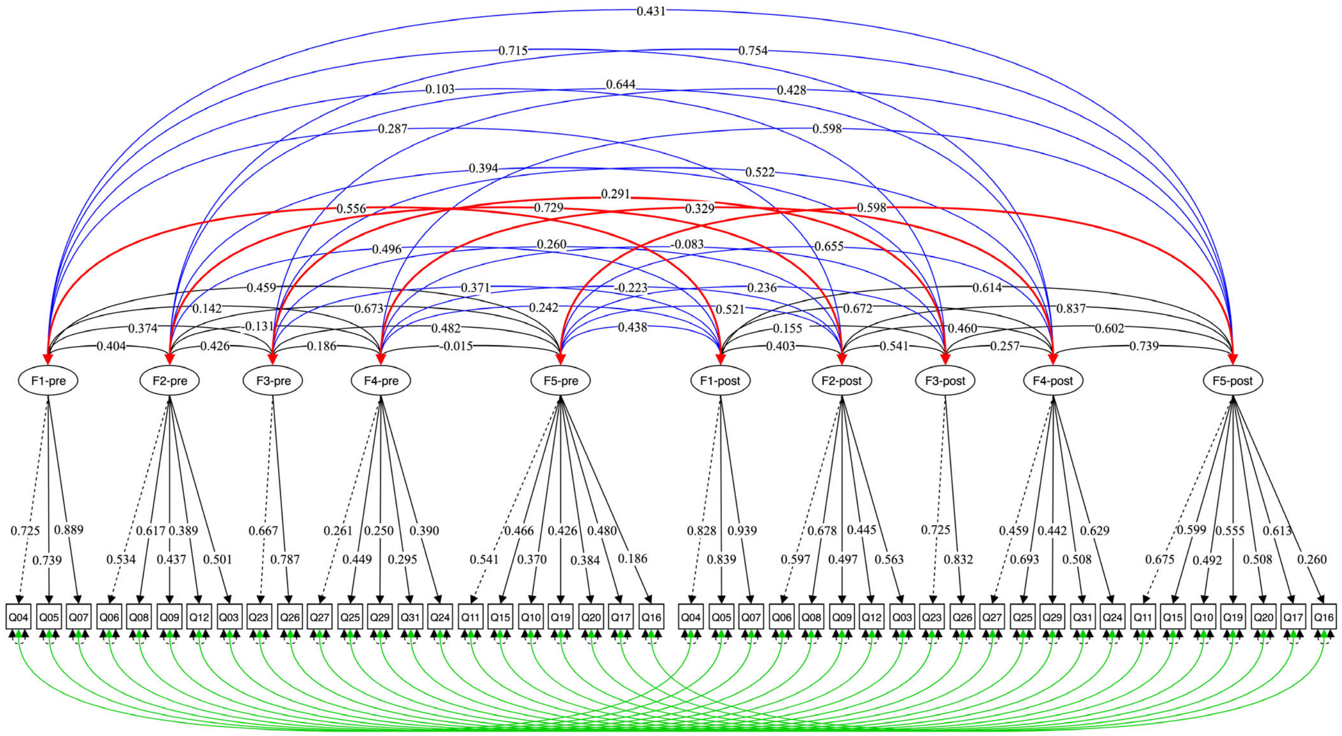

FIG. 3. Strict LMI model of the reduced CSEM. The modified Eaton5 model suggested the following five factors: F1 = Nettons's 3rd law, F2 = $\vec{F}_E = q\vec{E}$ + Superpos, F3 = $\vec{B}$ by I, F4 = Electromagnetic interaction, and F5 = q in fields. Because of the data's longitudinal structure, the latent factors are allowed to correlate across time (the correlation between identical latent factors in the different time was in red curves; the correlation between different latent factors in the different time was in blue curves), and identical items over the two test occasions were allowed to covary. All parameters are standardized.

values at post-test, ranging from 0.155 to 0.837. At pretest, F4 has negative correlations with F2 (−0.131) and F5 (−0.015). From pre- to post-test, the correlations of the same factor in different test times are also moderate, ranging from 0.291 to 0.729.

V. DISCUSSION

A. The factor structures of FCI and CSEM

In this study, the factor structures of FCI and CSEM have been examined using CFA for both pre- and post-test data. In the analysis, several factor models suggested by previous studies are also included. It has been found that, for FCI, three factor models, which were suggested by the creators of the FCI (HWS6 model) [1], by Scott *et al.* (SSG5 model) [48], and by Eaton and Willoughby (EW5 model) [53], all fit the data well. Among the three models, EW5 fits the pretest data best, while the SSG5 model fits best with the post-test data. The results are in agreement with the findings of Eaton and Willoughby, who conducted CFA of the FCI with a large scale post-test dataset [53]. For CSEM, the EFA driven factor model identified by Eaton *et al.* (Eaton6 model) [54] has also been examined. It has found that the Eaton6 model can fit moderately well with the data in this study, which further indicates the generalizability of the Eaton6 model.

However, the reliabilities of certain factors measured by the two CIs have shown unacceptable values of CR. For FCI, the items in the unreliable factors mainly address Newton's second law and kinematics. For CSEM, the items in the unreliable factors mainly address the magnetic forces on charged particles and the relationship between electric potential and electric field. In addition, all the unreliable measures occur in the pretest data only.

Based on the previous models, various modifications have been explored to make the unacceptable factors more reliable, which produce three modified models for FCI (HWS6M, SSG5M, and EW5M) and one for CSEM (Eaton5M). For FCI, the CR values of the factors in the two modified models (SSG5M and EW5M) have achieved the minimally acceptable level (0.65). For CSEM, while the internal consistency of the unreliable factors has improved in the modified models, the reliability estimates are still below the acceptable level. Nevertheless, the modified factor models proposed in the current study seem to perform better than previous models.

B. The stability of latent constructs measured by FCI and CSEM between pre- and post-test

In the current study, the LMI analysis has been conducted for FCI and CSEM to examine the stability of latent constructs measured by these two CIs. For the FCI, the

SSG5M and EW5M models fit well with both pre- and post-test data. For CSEM, the Eaton5M model is the only reliable model, and fits the data very well.

The test of configural invariance establishes the equivalence of conceptual constructs for the FCI and the reduced CSEM between pre- and post-test. The results for the FCI are different from prior findings, which show different factors for pre- and post-test through EFA [13]. However, it is noteworthy that the current study is the first to examine the pre-post measurement invariance of FCI under the CFA framework. The different findings may be due to different factor analysis methods applied, i.e., the current study uses CFA, while EFA was used in the previous study [13].

The test of strong invariance establishes the comparability of latent abilities and observed scores between pre- and post-test. Given the importance of intervention studies in understanding the effectiveness of a specific teaching method [6], it is essential to establish strong invariance of the CIs in science education [20]. When the conceptual constructs measured by a CI are not invariant between tests, changes of scores may not reflect change along with the latent constructs of interest [21], but rather changes of the latent structures themselves. For example, such issues have been concerned for students' understanding of mechanics measured by the FCI [12,13], and to some extent, this study provides empirical evidence for these studies.

The test of strict invariance shows that the residual variances are invariant between pre- and post-test for the reduced CSEM and only partial strict invariance can be established for FCI. Specifically, when the SSG5M model is fitted to the FCI dataset, the residuals of item 29 vary between pre- and post-test. For the EW5M model, two items have been documented as responsible for the lack of residual invariance, i.e., item 29 and item 2. However, as a highly constrained model, strict invariance is rarely held in practice [73].

C. Association between students' concepts and the latent constructs measured by a CI

The latent constructs inferred from test data cannot be observed directly. Meanwhile, the structures of students' conceptual understandings can be analyzed based on student responses across a variety of items in a CI [42]. Results of such an analysis are always the outcome of the interactions between the instrument and the students. When the two threads of analysis are in agreement, the latent constructs measured by a CI can be viewed as representative factors underlying students' understandings.

Regarding teaching and learning, many researchers consider that students' conceptual understandings are reconstructed through instruction [14–17]. On the one hand, some researchers argue that students' conception structure should be fundamentally or radically reconstructed after effective instruction [14,38]. Meanwhile, other researchers consider the coexistence of preinstruction conceptions and scientific conceptions in students' conception structure (e.g., student model state and model space [9], the conceptual profile change model [39], and the multiple knowledge system model [40,41]). The latter models resonate with some recent neuroscience studies, which have shown that experts may still have an incorrect preinstruction conception encoded in their brain neural networks that must be inhibited in order to reject incorrect tasks (see a functional magnetic resonance imaging study in Ref. [74], and an event-related potential study in Ref. [75]).

Nevertheless, views on conceptual reconstruction add new meaning for considering longitudinal measurement invariance (LMI) of CIs. Since students' conceptual constructs are expected to change from pretest to post-test, it is then a question as to what extent such changes may exist among different student populations. In this study, the full-FCI LMI analysis demonstrates partial strict invariance for the FCI. These results suggest that, for introductory-level college students (at least for the sample in this study), the latent constructs of FCI do not change from pretest to post-test. However, for the CSEM, only the 22-item reduced test maintains strict invariant between pre- and post-test.

These two commonly used CIs also provide insights into the possible influences from students' prior knowledge on measurement invariance of latent constructs. It is commonly assumed (and also has been confirmed) that introductory-level college students may be more familiar with the mechanics content than that of electricity and magnetism content (see a large-scale investigation in the U.S. for the performance of freshmen college students in these two content areas before physics instruction in Bao *et al.* [76]). That is, students' prior knowledge of mechanics may be more accessible than that of electricity and magnetism. At pretest students may have some level of expertlike understandings in mechanics, and therefore, the pre-post conceptual change can be viewed as redistribution of probabilities for cueing naïve and expert views [9], both of which are included in the factor construct model. As a result, the structure of the factor model would maintain consistency between pre- and post-test. On the other hand, for electricity and magnetism, students typically have little understanding at pretest and their answers reflect significant guessing, while on the post-test, students will start to show some meaningful understanding of the content domains. Therefore, students' conceptual understandings on electricity and magnetism at pre- and post-test should have substantial structural differences, which lead to the weaker LMI on CSEM between pre- and post-test.

This hypothesis can be partially confirmed by the results from this study, which show that the full FCI held partial strict invariance, while the CSEM has to exclude some items (i.e., the reduced CSME) in order to hold strict invariance. To further examine this hypothesis, LMI analysis of a CI can be conducted for student samples with large differences, e.g., high school students vs introductory-level college students, which should warrant attention in future research.

Based on the discussion, it is also suggested that the lack of LMI for a CI should not necessarily invalidate the CI. In the case when students' understandings are significantly restructured, such as on topics with CSEM, the latent constructs measured on the pretest will inevitably be significantly different from that of the post-test. In such cases, the goal of assessment should not aim to maintain LMI, but rather, it is more productive to analyze the differences between pre-post latent constructs in order to make meaningful inferences on the nature of students' conceptual changes.

D. Limitations and suggestions for further study

This research has some limitations that need to be considered when interpreting the outcomes. First, the datasets used for conducting LMI analysis were collected from only two universities in the United States. The results cannot be generally extended to students from other universities and education systems. It would be beneficial to inspect if similar results can be replicated with students from different institutions and education settings. Nevertheless, this study provides a case of empirical analysis of the measurement invariance of CIs from pre- to post-test, which is lacking in current literature. The current study illustrates a way to examine the LMI of CIs in the science education field. Future studies are encouraged to further validate the related models with extended student populations.

Second, the findings of LMI are based on selected previous models [48,53,54], which may not reflect the optimal factor structures of FCI and CSEM. Hence, other models, e.g., the set of theoretical models by Stewart and his colleagues [52,55], may also be explored in LMI analysis for these CIs.

Finally, the CSEM has to exclude some items in order to hold strict invariance. Therefore, the reduced version cannot fully represent the content assessment of the original CSEM.

VI. CONCLUSION

The current study illustrates a way to examine the LMI of FCI and CSEM. For the FCI, the factor models fit well with both pre- and post-test data. For CSEM, acceptable fits are obtained with a reduced version of the CI. However, the combining reliability analysis suggests that the previous models from the literature should be revised. Accordingly, several modified factor models were proposed. Overall, the LMI on the FCI demonstrates the existence of partial strict invariance, while LMI analysis on a reduced version of the CSEM indicates strict invariance.

These findings provide the first piece of empirical evidence that pre-post changes of FCI scores and the reduced CSEM can be attributed to ability changes along the same latent constructs measured by the CIs. These results further establish the CIs' reliability for assessment of students' conceptual changes over time in introductory physics courses.

ACKNOWLEDGMENTS

This work was supported by the NSF Grant No. DUE-1712238, the Humanities and Social Sciences of the Ministry of Education of P. R. China under Grant No. 15YJC880103, and the Guangdong Planning Office of Philosophy and Social Science under Grant No. GD19CJY15. Any opinions expressed in this work are those of the authors and do not necessarily represent those of the funding agencies.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[2] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. V. Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[3] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).

[4] L. Törnqvist, P. Vartia, and Y. O. Vartia, How should relative changes be measured, Am. Stat. **39**, 43 (1985).

[5] B. V. Dusen and J. Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, J. Res. Sci. Teach. **57**, 33 (2020).

[6] 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).

[7] E. D. Scribner and S. E. Harris, The mineralogy concept inventory: A statistically validated assessment to measure learning gains in undergraduate mineralogy courses, J. Geosci. Educ. 1 (2019).

[8] L. K. Weir, M. K. Barker, L. M. McDonnell, N. G. Schimpf, T. M. Rodela, and P. M. Schulte, Small changes, big gains: A curriculum-wide study of teaching practices and student learning in undergraduate biology, PLoS One **14**, e0220900 (2019).

[9] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).

[10] T. C. Pentecost and J. Barbera, Measuring learning gains in chemical education: A comparison of two methods, J. Chem. Educ. **90**, 839 (2013).

[11] Glen Davenport, Detecting conceptual change with latent transition analysis, Ph.D. thesis, University of Connecticut, 2016.

[12] P. Eaton, K. Vavruska, and S. Willoughby, Exploring the preinstruction and postinstruction non-Newtonian world views as measured by the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **15**, 010123 (2019).

[13] 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).

[14] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[15] A. A. diSessa, N. M. Gillespie, and J. B. Esterly, Coherence versus fragmentation in the development of the concept of force, Cogn. Sci. **28**, 843 (2004).

[16] R. Duit and D. F. Treagust, Conceptual change: A powerful framework for improving science teaching and learning, Int. J. Sci. Educ. **25**, 671 (2003).

[17] S. Vosniadou, On the nature of naïve physics, in *Reconsidering Conceptual Change: Issues in Theory and Practice*, edited by M. Limón and L. Mason (Springer Netherlands, Dordrecht, 2002), pp. 61–76.

[18] K. F. Widaman, E. Ferrer, and R. D. Conger, Factorial invariance within longitudinal structural equation models: Measuring the same construct across time, Child Dev. Perspect. **4**, 10 (2010).

[19] K. F. Widaman and S. P. Reise, Exploring the measurement invariance of psychological instruments: Applications in the substance use domain, in *The science of prevention: Methodological advances from alcohol and substance abuse research*, edited by K. J. Bryant, M. Windle, and S. G. West (American Psychological Association, Washington, DC, US, 1997), pp. 281–324.

[20] L.-T. Tsai, C.-C. Chang, and C.-K. Wu, Development of marine science affect scale for junior high school students in Taiwan: Testing for measurement invariance, Eurasia J. Math. Sci. Technol. Educ. **14**, 53 (2017).

[21] Y. Liu, R. E. Millsap, S. G. West, J.-Y. Tein, R. Tanaka, and K. J. Grimm, Testing measurement invariance in longitudinal data with ordered-categorical measures, Psychol. Methods **22**, 486 (2017).

[22] H. Helm, Misconceptions in physics amongst South African students, Phys. Educ. **15**, 92 (1980).

[23] R. Driver and J. Easley, Pupils and paradigms: A review of literature related to concept development in adolescent science students, Studies in Science Education **5**, 61 (1978).

[24] J. Minstrell, Facets of students' knowledge and relevant instruction, in *Research in physics learning: Theoretical issues and empirical studies*, edited by R. Duit, F. Goldberg, and H. Niedderer (IPN, Kiel, 1992), pp. 110–128.

[25] M. T. H. Chi, Conceptual change within and across ontological categories: Examples from learning and discovery in science, in *Cognitive Models of Science*, edited by R. N. Giere (University of Minnesota Press, Minneapolis, MN, 1992).

[26] G. Özdemir and D. B. Clark, An overview of conceptual change theories, Eurasia J. Math. Sci. Tech. Ed. **3**, 351 (2007).

[27] A. C. Alonzo and J. T. Steedle, Developing and assessing a force and motion learning progression, Sci. Educ. **93**, 389 (2009).

[28] A. Alonzo and A. Elby, The nature of student thinking and its implications for the use of learning progressions to inform classroom instruction, in *Learning and becoming in practice: The International Conference of the Learning Sciences, Boulder, Colorado, USA* (2014), pp. 1037–1041.

[29] L. Bao, K. Hogg, and D. Zollman, Model analysis of fine structures of student models: An example with Newton's third law, Am. J. Phys. **70**, 766 (2002).

[30] M.C. Linn, The knowledge integration perspective on learning and instruction, in *The Cambridge Handbook of the Learning Sciences*, edited by R.K. Sawyer (Cambridge University Press, Cambridge, England, 2005), pp. 243–264.

[31] J. T. Steedle and R. J. Shavelson, Supporting valid interpretations of learning progression level diagnoses, J. Res. Sci. Teach. **46**, 699 (2009).

[32] R. Dai, J. C. Fritchman, Q. Liu, Y. Xiao, H. Yu, and L. Bao, Assessment of student understanding on light interference, Phys. Rev. Phys. Educ. Res. **15**, 020134 (2019).

[33] H.-S. Lee, O. L. Liu, and M. C. Linn, Validating measurement of knowledge integration in science using multiple-choice and explanation items, Appl. Meas. Educ. **24**, 115 (2011).

[34] Y. Nie, Y. Xiao, J. C. Fritchman, Q. Liu, J. Han, J. Xiong, and L. Bao, Teaching towards knowledge integration in learning force and motion, Int. J. Sci. Educ. **41**, 2271 (2019).

[35] J. Shen, O. L. Liu, and H.-Y. Chang, Assessing students' deep conceptual understanding in physical sciences: An example on sinking and floating, Int. J. Sci. Math. Educ. **15**, 57 (2017).

[36] Y. Hadzigeorgiou, Young children's ideas about physical science concepts, in *Research in Early Childhood Science Education*, edited by K. Cabe Trundle and M. Saçkes (Springer, Dordrecht, 2015), pp. 67–97.

[37] R. Duit, *Bibliography: Students' and Teachers' Conceptions and Science Education* (University of Kiel, Kiel, Germany, 2009).

[38] Å. Larsson and O. Halldén, A structural view on the emergence of a conception: Conceptual change as radical reconstruction of contexts, Sci. Educ. **94**, 640 (2009).

[39] E. F. Mortimer, Conceptual change or conceptual profile change?, Sci. Educ. **4**, 267 (1995).

[40] J. Solomon, Learning about energy: How pupils think in two domains, Eur. J. Sci. Educ. **5**, 49 (1983).

[41] J. Solomon, Prompts, cues and discrimination: The utilization of two separate knowledge systems, Eur. J. Sci. Educ. **6**, 277 (1984).

[42] P. C. Price, R. Jhangiani, and I.-C. A. Chiang, *Research Methods in Psychology*, 2nd Canadian ed. (BCcampus, Victoria, B.C., 2015).

[43] A. Eryilmaz, Effects of conceptual assignments and conceptual change discussions on students' misconceptions and achievement regarding force and motion, J. Res. Sci. Teach. **39**, 1001 (2002).

[44] D. L. Evans, G. L. Gray, S. Krause, J. Martin, C. Midkiff, B. M. Notaros, M. Pavelich, D. Rancour, T. Reed-Rhoads,

and P. Steif, *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference* (IEEE, Boulder, CO, 2003), pp. T4G–1.

[45] G. Taasoobshirazi and G. M. Sinatra, A structural equation model of conceptual change in physics, J. Res. Sci. Teach. **48**, 901 (2011).

[46] R. S. Lindell, E. Peak, and T. M. Foster, Are they all created equal? A comparison of different concept inventory development methodologies, AIP Conf. Proc. **883**, 14 (2007).

[47] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, Phys. Teach. **33**, 138 (1995).

[48] 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).

[49] 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).

[50] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[51] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. ST Phys. Educ. Res. **11**, 020134 (2015).

[52] 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).

[53] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).

[54] P. Eaton, B. Frank, K. Johnson, and S. Willoughby, Comparing exploratory factor models of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **15**, 020133 (2019).

[55] C. Zabriskie and J. Stewart, Multidimensional item response theory and the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **15**, 020107 (2019).

[56] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[57] J. L. Horn and J. J. McArdle, A practical and theoretical guide to measurement invariance in aging research, Experimental Aging Research **18**, 117 (1992).

[58] A. Abdellaoui, M. H. M. de Moor, L. M. Geels, J. H. D. A. van Beek, G. Willemsen, and D. I. Boomsma, Thought problems from adolescence to adulthood: measurement invariance and longitudinal heritability, Behavior genetics **42**, 19 (2012).

[59] Z. J. Ng, E. S. Huebner, A. Maydeu-Olivares, and K. J. Hills, Confirmatory factor analytic structure and measurement invariance of the emotion regulation questionnaire for children and adolescents in a longitudinal sample of adolescents, J. Psychoeduc. Assess. **37**, 139 (2017).

[60] A. W. Meade and G. J. Lautenschlager, Same Question, Different Answers: CFA and Two IRT Approaches to Measurement Invariance, in *The 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL, 2004* (2004), p. 19.

[61] Y. Rosseel, Lavaan: An R package for structural equation modeling and more, J. Stat. Softw. **48**, 1 (2012).

[62] T. D. Jorgensen, S. Pornprasertmanit, A. M. Schoemann, and Y. Rosseel, SemTools: Useful tools for structural equation modeling (R package version 0.5-3, 2020), https://CRAN.R-project.org/package=semTools.

[63] A. Beauducel and P. Y. Herzberg, On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA, Struct. Equ. Model. Multidiscip. J. **13**, 186 (2006).

[64] M. Rhemtulla, P. É. Brosseau-Liard, and V. Savalei, When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions, Psychol. Methods **17**, 354 (2012).

[65] L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Struct. Equ. Model. Multidiscip. J. **6**, 1 (1999).

[66] B. Muthén and T. Asparouhov, Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus, Mplus Web Notes **4**, 1 (2002); http://www.statmodel.com/examples/webnote.shtml#web4.

[67] 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).

[68] A. Satorra and P. M. Bentler, A scaled difference chi-square test statistic for moment structure analysis, Psychometrika **66**, 507 (2001).

[69] F. F. Chen, K. H. Sousa, and S. G. West, Testing measurement invariance of second-order factor models, Struct. Equ. Model. **12**, 471 (2005).

[70] F. F. Chen, Sensitivity of goodness of fit indexes to lack of measurement invariance, Struct. Equ. Model. Multidiscip. J. **14**, 464 (2007).

[71] R. F. DeVellis, *Scale Development: Theory and Applications* (Sage Publications, Thousand Oaks, CA, 2012).

[72] G. W. Cheung and R. B. Rensvold, Testing factorial invariance across groups: A reconceptualization and proposed new method, J. Manag. **25**, 1 (1999).

[73] N. Kibrislioğlu Uysal and Ç. Akin Arikan, Measurement invariance of science self-efficacy scale in PISA, Int. J. Assess. Tools Educ. **5**, 325 (2018).

[74] S. Masson, P. Potvin, M. Riopel, and L.-M. B. Foisy, Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity: Brain activation related to scientific expertise, Mind Brain Educ. **8**, 44 (2014).

[75] Y. Zhu, L. Zhang, Y. Leng, R. Pang, and X. Wang, Event-related potential evidence for persistence of an intuitive misconception about electricity, Mind Brain Educ. **13**, 80 (2019).

[76] L. Bao, T. Cai, K. Koenig, K. Fang, J. Han, J. Wang, Q. Liu, L. Ding, L. Cui, and Y. Luo, Learning and scientific reasoning, Science **323**, 586 (2009).