Detecting the influence of item chaining on student responses to the Force Concept Inventory and the Force and Motion Conceptual Evaluation

Philip Eaton[*]

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

Barrett Frank and Shannon Willoughby

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

(Received 17 January 2020; accepted 8 September 2020; published 1 October 2020)

Items that are chained, or blocked, together appear on many of the conceptual assessments utilized for physics education research. However, when items are chained together there is the potential to introduce local dependence between those items, which would violate the assumption of item independence required by classical test theory, unidimensional item response theory, and other measurement theories. Local dependence can be divided into two categories: (i) underlying local dependence, which can be adequately modeled with multidimensional measurement theories, and (ii) surface local dependence (SLD), which cannot be modeled using multidimensional measurement theories. The act of chaining items is thought to be one of the many potential sources of SLD between items. Using previous local dependence research results, this study proposes two methods for detecting the presence of local dependence and SLD between items on an assessment. These methods were applied to the Force Concept Inventory (FCI) and the Force and Motion Conceptual Evaluation (FMCE). It was found that the assumption of item independence was violated for both assessments, implying that unidimensional measurement theories may not adequately model either the FCI or FMCE. Further, both detection methods identified the potential for a minimal amount of SLD present for FCI and a significant amount of SLD present for the FMCE. This implies that even multidimensional measurement theories may not be capable of adequately modeling the FMCE when scoring items individually. This result supports the claim made by Thornton *et al.* that the items on the FMCE should be scored in groups; however, the currently proposed grading scheme was found to be inadequate.

DOI: 10.1103/PhysRevPhysEducRes.16.020122

I. INTRODUCTION

Identifying the optimal conceptual instrument is often difficult for both researchers and educators. This is compounded when multiple tools exist for a single domain of physics knowledge. For example, both the Force Concept Inventory (FCI) and the Force and Motion Conceptual Evaluation (FMCE) attempt to probe students' understanding of Newtonian mechanics [1,2]. Deciding between which of these instruments to use can be done by comparing their validity and/or the specific concepts each instrument probes.

The FCI is a 30-item, five-option, multiple-choice conceptual assessment with a suggested duration of 30 minutes [1,3]. Whereas, the FMCE is a 47-item, shared response pool, multiple-choice conceptual assessment with a suggested duration of 35 minutes [2,3]. Each assessment was designed to probe student understanding of Newton's laws and kinematics. The FMCE further probes position and velocity versus time plots as well as conservation of mechanical energy.

Both assessments have been exposed to numerous statistical analyses ranging from unidimensional to multidimensional treatments; see the following studies for a sample of the kinds of analysis that have been performed [2,4–17]. Also, both have been used numerous times in classrooms to assess the understanding of students and/or the effectiveness of new and innovative curricula; see, for example, Refs. [18,19]. Overall, each assessment is designed to take a similar amount of time to complete, probes similar concepts, and has been shown to function well statistically.

Another feature shared by both instruments is the utilization of item chaining. The technique of item (question) chaining, or item blocking, is present on many of the current conceptual assessments used in physics education

[*] philip.eaton@stockton.edu

research (PER) [1,2,20–22]. Items are said to be chained together when groups of items appear in close physical proximity while probing the same concepts or when items use the same figures, response pools, reading prompts, etc. [23,24]. This is done by test developers for numerous reasons, which include, but are not limited to (i) less space is used to print the assessment, (ii) more items can be asked using a single figure, reading prompt, etc., and (iii) false-positive detection possibilities can be added.

The reasons for utilization of item chaining are sensible, and in some cases desirable; however, they can introduce unintended statistical consequences. The most critical issue associated with item chaining is the high potential for loss of local item independence between the chained items. Local item independence, or local independence, is the assumption that items are conditionally independent of each other, and is required for many measurement theories used in PER [24,25]. For example, unidimensional item response theory (IRT) assumes local item independence to estimate item parameters within student response models [25]. Similarly, classical test theory (CTT) assumes the errors of items are independent from one another, which implies the items themselves are independent [24]. Understanding if an assessment breaks this assumption is vital to selecting appropriate measurement theories and models. As a result, the total score reported for an assessment depends on which grading model is used. This then affects the gains measured for students and classes. Thus, understanding which measurement model should be used is critical to generating the accurate gain measurements needed for curricular intervention research.

The loss of local independence, referred to as local dependence (LD), has been investigated in the psychometric field for some time now [26–32]. However, few—if any—of these results have been implemented into assessing the conceptual assessments used in PER.

In 1997, Chen and Thissen proposed separating LD into two categories: surface LD (SLD) and underlying LD (ULD) [30]. Underlying LD occurs when groups of items on an assessment share a common latent trait (e.g., a physics conception) that links items together. As a result, this kind of linking can also be called a *conceptual linking* since students interact with the linked questions through the underlying trait they share. Effects of this nature can be modeled using multiple latent variable models, like multitrait IRT and factor analysis, but are not described within unidimensional models [25,33].

Surface LD occurs when students answer an item based on superficial characteristics of previous items; they are not independently interacting with the items on the assessment. A situation like this could occur for items which are chained together. Item chaining is an example of what may cause items to be possibly linked via SLD, and could cause students to answer items based partially, or entirely, on how they responded to the previous items. Thus, this kind of linking could be referred to as an *artificial linking* as students are no longer using only a shared concept when interacting with the linked questions. Chaining items of similar content together is a prime situation for the existence of SLD. This form of LD is problematic as all commonly used psychometric theories cannot properly account for artificial linking.

An investigation into the presence of SLD between items is one way to test for the potential influences of item chaining. Since SLD is *not* caused by a shared latent trait which links the items, resulting effects of SLD cannot be accounted for using current multidimensional models.

The first article that we are aware of that investigated LD between items on a conceptual assessment in PER was the original validation of the relativity concept inventory (RCI) [34]. The article discussed initial results for the RCI and identified some item pairs that were likely breaking local item independence. However, it did not go into detail about the effects that detected LD would have on parameter estimations, within IRT or CTT.

Recent efforts made toward understanding the impacts of chaining items within PER conceptual assessments can be found in Refs. [8,9]. These studies looked into the effects of "blocking" items on the Force Concept Inventory (FCI) using multitrait item response theory (MIRT) and modular network analysis. The methods used in these studies were then applied to the Force and Motion Conceptual Evaluation (FMCE) in Ref. [35]. It was found that the FCI's and FMCE's factor structures were likely being significantly impacted by the "blocking" of items on each assessment.

Concept inventories have been used extensively in the field of physics education. One main repository for these inventories is found in Ref. [3], which has 95 such assessments listed. The authors of this website encourage users to upload their own class data, as one goal of PhysPort is to allow investigators to compare their data with a national dataset they are building. Many types of data analysis are possible with large datasets, including factor analysis, structural equation modeling, and item response theory. However, each of these types of analysis assumes local item independence. It is, of course, possible to test this assumption, but it would appear that in the field of PER, it is merely assumed to be true and *not actually tested* before the data are analyzed statistically. For example, in 2019 Physical Review Physics Education Research published a special focus issue, "Quantitative methods on PER: A critical examination." Only two of the articles in this collection mention testing for item dependence [9,36]. However, these articles are fairly typical in our field in that they do not provide this level of detail regarding specifically how the data are, or should be, treated before analysis is completed. Further, on the PhysPort website, the "research" tab for the FMCE includes a number of articles related to the validation of the instrument itself, as well as analysis of the results of using the instrument. None of the articles listed on this website and published within the last ten years mention tests of item independence as part of their preanalysis work on the data.

This perusal of the literature in PER suggests that local item dependence is not being treated correctly by the field as a whole. Thus we argue that this current work will help jump-start a much needed conversation among fellow PER researchers that statistical assumptions such as item independence are indeed just that, *assumptions*, and in order for us to gain useful information from the use of concept inventories, we must begin by assessing whether or not our concept inventories are or are not breaking the assumptions used in currently popular types of statistical analyses.

Initially, this study sought to examine the factor structure of the FMCE within an exploratory factor analysis (EFA) framework. Upon implementation of EFA, results similar to those found in Ref. [35] were obtained. It was found that the developed factor structure simply mimicked the blocked items that appear sequentially on the assessment itself. Further, a modular network analysis of the FMCE resulted in a structure which mimics the blocking structure of the assessment [37]. The fact that the correlational structure of items follows the chained item blocks is troubling, and leads to the investigation of the effects of chaining items.

A clear example of item chaining can be seen with items 1 and 4 on the FMCE. Both ask students to analyze a situation where a person pushes a sled across an icy surface (i.e., a frictionless surface). Item 1 prompts students to consider which force is required to push the sled to the right while speeding it up. In item 4 the sled instead moves to the left while still speeding it up. Both of these items use the same response pool and figures while asking extremely similar questions; only the direction of motion changes. It is reasonable to infer that students could be answering item 4 based partially, or entirely, on how they responded to item 1.

Since both assessments were designed to probe multiple concepts and both employ the use of item chaining, it is not unrealistic to expect some form of LD to be present. It is expected that both assessments will display some amount of both ULD and SLD. If local independence is found to be broken, then theories commonly used to analyze assessments will give statistically biased results [29]. For example, the presence of local dependence on an assessment analyzed using IRT or CTT can result in incorrect estimations for item difficulty, item discrimination, test reliability, and student ability. These impacts are likely less important to instructors assessing their own class. However, as these measures are used to assess both the quality of the assessment and the gains of the students, any errors in their estimations are of critical importance to researchers investigating the impact of new curricular interventions.

If the local dependence in an assessment can be entirely described by conceptual linking (ULD), then multidimensional analysis methods can, and should, be used to properly assess the quality of the instrument. However, if the linkings are artificial, then there will be some level of inherent error built into the results of multidimensional analysis methods. If there is a large amount of SLD present, many items that are artificially linked, then this error may lead to incorrect conclusions. In this case, special grading criteria (e.g., testlet grading methods) will need to be generated and assessed before research should use the assessment to examine new pedagogical techniques.

This study investigated the extent to which local dependence and artificial linking (SLD), assumed to be caused by item chaining, were present on the FCI and FMCE through considering the following research questions:

RQ 1: To what extent is the assumption of local item independence valid for each assessment?

RQ 2: To what extent is SLD present on either of the assessments?

RQ 3: By examining the item pairs identified in research question 2, is it correct to assume that item chaining is responsible for the SLD inferred?

The rest of this work is organized as follows, First, the data used in this study are specified in Sec. II. Then a detailed methodology is discussed in Sec. III, followed by a presentation of the methodology's results in Sec. IV. The implications of the results are considered in Sec. V. Lastly, the limitations of the study, a summary of the study, and suggestions for the future direction of PER can be found in Secs. VI–VIII, respectively.

II. DATA

The data for the FCI and the FMCE came from students taking algebra- or calculus-based first-year introductory mechanics (i.e., Physics I) before and after instruction had taken place. PhysPort supplied the data for both assessments [3]. The supplied data for both assessments had incomplete demographic information, so complete details of the demographic breakdown of the data are unknown.

The data for the FCI originally contained 22 029 students. However, after removing any students with blank entries in their pre- or postinstruction response vectors, the sample was left with 19 745 matched student responses. Similarly, the original FMCE data contained 19 708 pre- and/or postinstruction student responses. Many of these student repressions were only for before or after instruction, not both. After removing students that did not take both a pre- and postinstruction administration, and any student with blank responses, the FMCE sample was left with 10 084 matched student responses. Test statistics for each of these samples can be found in Table I.

The models used in this analysis required the data to be graded dichotomously, meaning questions are sorted between two different categories for each student. In this case questions are answered either correctly or incorrectly. In the response vectors this is represented by a "1" for correct and "0" for incorrect.

TABLE I. Test statistics for each of the samples used in this study. The data are matched pre- to postinstruction. The mean is represented by μ and the standard deviation by σ. The scores for the FMCE are calculated using the blocked grading proposed in Ref. [17].

Assessment	N	Pre μ	Pre σ	Post μ	Post σ
FCI	19 745	0.437	0.213	0.608	0.221
FMCE	10 084	0.317	0.244	0.537	0.296

Dichotomous scoring is common for the FCI, but is not recommended for the FMCE. It has been proposed that the FMCE should be graded in a blocked fashion [2,17]. However, since this study is looking into the LD between individual items, each item will be graded separately.

III. METHODOLOGY

The following section contains a detailed methodology of how conceptual and artificial links (ULD and SLD) between items can be detected. A brief explanation of how this is done can be found in Sec. III A and a technical discussion of the methods used can be found in Sec. III B.

A. Brief methodology

Identifying the possible presence of conceptual and artificial linking (ULD and SLD) within an assessment can be done using simulations. Both conceptual and artificial linking can be modeled in a simple manner. These simple models are used to simulate student responses to questions on an assessment that incorporated either of these linkings. Each model contains a single parameter which characterizes the amount of linking present between the items. The ULD parameter has an understood range of values for typical assessments. Thus, if the ULD model using a parameter at the maximum of this range, or higher, cannot account for all of the local dependence detected between a pair of items, then some amount of SLD must be present. This is how the possible presence of SLD can be detected. When these parameters are set to zero, the linking is "turned off," and when increased from zero simulate stronger links between items.

The artificial linking model is parametrized by a probability, π_{LD}. This represents the probability that the independent question in the linked pair informs the correctness of the other, dependent question, regardless of the relative difficulty or content of the questions. This parameter ranges from 0 (i.e., no linking) to 1 (entirely linked). For example, within an assessment where questions 3 and 4 are artificially linked, question 4 is assumed to be the dependent question. Thus, there is a probability of π_{LD} that a student will have the same result on question 4 as question 3 (i.e., both correct or incorrect). This form of dependence is assumed to be how item chaining impacts student responses to the questions.

The ULD parameter, or ULD weight, ranges from 0 (i.e., no linking) to an unbounded maximum. However, for typical assessments, ULD weights above 1.5 are rarely observed [32]. This weight corresponds to the strength of the underlying concept (or trait) that links the question pair together. As the ULD weight parameter increases, the apparent ability of a student also increases for these linked questions. This models the manner in which an underlying concept (e.g., Newton's third law) could be used by students to assist them in answering conceptually linked questions on an assessment. It is important to remember that these conceptual linkings can be modeled using multiple latent trait methods.

The presence of local dependence is quantified by three measures: (i) Cramer's V of Pearson's χ^2, (ii) Cramer's V of the G^2 statistic, and (iii) the tetrachoric correlation. The expected values of these statistics for an assessment that contains totally independent items can be calculated through simulations. Thus, it is safe to assume that any deviations from these values are caused by the presence of local dependence. This methodology allows for the identification of pairs of questions that are linked by some form of LD, but does not differentiate between conceptual and artificial linking.

However, since the ULD weights for typical conceptual linkings are bound between 0 and 1.5, any measures of LD which are not explained by ULD weights in this range are assumed to be, at least partially, a manifestation of artificial linking. That is, if an assessment's statistics of LD are found to be significantly larger than those generated using this range of ULD weights, it can be concluded that conceptual linking is not the only source of LD between the questions. This implies that some amount of artificial linking is present to account for all of the detected LD.

This describes how local dependence and artificial linking can be detected on the FCI and FMCE. For more technical details of how this was carried out, see the following section of the methodology. However, those not interested in the technical specifics can skip the following section and go to Sec. IV.

B. Technical methodology

1. Item response theory

Item response theory is a latent trait theory that attempts to measure students' ability scores through their interactions with items (i.e., questions) on an assessment. Lord and Novick proposed two assumptions that must be met for a mathematical IRT model to be viable [38]. The first of these assumptions pertains to the mathematical models themselves, and simply asserts that the model must describe the data well [38]. For example, a function that predicts a decreasing probability of a student answering a problem correctly as their ability increases would not match the qualitative or quantitative nature of how questions are answered. A function of this nature would break the first assumption of IRT, and would not be an adequate IRT model.

One model which satisfies this assumption is the two-parameter logistic (2PL) model. The 2PL model returns the probability that a student with an ability of θ will respond correctly to item i given the item's discrimination index α_i and intercept index d_i. Mathematically, the 2PL model can be given as

$$P(X_i = 1|\theta, \alpha_i, d_i) = \frac{1}{1 + e^{-D(\alpha_i\theta + d_i)}}.$$

The constant D is taken to be 1.702 to make the metric of the ability scale more closely relate to the traditional normal ogive metric (i.e., $\theta = 1$ is approximately 1 standard deviation in student ability). Typically, the item discrimination is factored out of the parentheses in the exponent and then the substitution $\delta_i = -d_i/\alpha_i$ is made, where δ_i is called the item difficulty index [25]. The item difficulty index is equal to the student ability required such that the probability a student will have responded correctly is 50%. Item discrimination relates to the slope of the 2PL curve at a student ability equal to the item difficulty. Lastly, $X_i = 1$ indicates a correct response for item i, and $X_i = 0$ indicates an incorrect response.

The second assumption states that item responses are locally independent from one another, meaning students respond to an item without being influenced by the other items on the assessment [38]. This assumption is used in parameter estimation techniques to find student ability scores and the item parameters for an assessment. Specifically, using the assumption of location item independence, an assessment's likelihood function can be written as the multiplicative product of all of the items' individual likelihood functions. Item and student parameters are found through maximizing the assessment's likelihood function with a given set of student responses. This parameter estimation technique is referred to as *maximization of the likelihood*. All of the item and student parameter estimations performed in this study used the R package MIRT [39,40].

2. Local dependence

The assumption of local item independence can be mathematically represented in the following manner:

$$P(X_i = 1, X_j = 1|\theta) = P(X_i = 1|\theta) \cdot P(X_j = 1|\theta).$$

This means the probability of getting items i and j correct simultaneously is equal to the probability of getting each item correct individually multiplied together. The same principle applies to getting both items incorrect, and one correct and the other incorrect. Any deviations away from this relation are an indication that students are not answering items i and j in a completely independent manner, which is to say local dependence exists between the two items.

Local dependence is separated into two categories: surface local dependence and underlying local dependence [30]. By definition, ULD results from unmodeled latent variables (i.e., modeling a multitrait assessment using a single trait), which can link multiple items together [30]. This could occur on an assessment that is designed to assess multiple conceptions (e.g., Newton's three laws, kinematics, etc.). These conceptions can be thought of as being linked to a global conception (e.g., Newtonian mechanics), but will appear in the statistics as different traits from a latent trait perspective. On the other hand, SLD occurs between pairs of items that contain highly similar content and/or are in close proximity on an assessment (e.g., chaining items of similar content together, using the same figure, response pool, reading prompt for a set of items, etc.).

Currently, it is not understood how to completely distinguish between the two types of LD when analyzing an assessment. Research into the effects of LD has been performed, and research into possibly distinguishing between the two types of LD on an assessment is ongoing [29,30,32,41].

Of the two, ULD is less concerning since it can be modeled using higher-dimensional models. That is, an assessment could be designed to independently measure both kinematics and Newton's three laws, which would be properly described by a four-trait multidimensional IRT model or a four-trait factor analysis model [25,42]. Alternatively, SLD can only be addressed by altering the structure of an assessment (i.e., moving items around, removing items, changing the wording of an item, etc.).

Mathematically, a simple model for SLD which links items m and n can be given as

$$\begin{aligned}
&\text{with a probability of } \pi_{\text{LD}}\text{:}\\
&X_n = \begin{cases} 1 & \text{if } X_m = 1 \\ 0 & \text{if } X_m = 0, \end{cases}\\
&\text{with a probability of } 1 - \pi_{\text{LD}}\text{:}\\
&X_n = \begin{cases} 1 & \text{with} \quad P(X_n = 1|\theta, \alpha_n, d_n) \\ 0 & \text{with} \quad P(X_n = 0|\theta, \alpha_n, d_n), \end{cases}
\end{aligned} \tag{1}$$

where $X_m = 1/0$ and $X_n = 1/0$ are the correct/incorrect responses to item m and n, respectively, and π_{LD} represents the degree, or severity, of SLD that has formed between the two items. For example, if $\pi_{\text{LD}} = 0.2$, then 20% of the time a student will answer item n based entirely on how they responded to item m. Thus, a student does not interact with item n in the manner assumed by IRT and CTT. This could occur on an assessment when multiple items that probe the same concept are asked sequentially (i.e., the items are blocked together). In this case, if the wording of the items is highly similar, then students may treat all of these items as if they were a single item. This effect is apparent with items

1–7 on the FMCE, where all of these items use the same response pool and figures while probing the same conception. Situations of this nature are prime locations for students to “use the test against itself” and respond to some items while being influenced by how they answered other items in this block.

Items with close proximity and similar wording and/or content will result in a higher likelihood of developing SLD. Because of the nature in which SLD is likely to occur, this analysis can effectively identify if the act of chaining items of similar content on an assessment superficially influences how students are responding.

Underlying LD can be modeled via a simple bifactor model which links multiple items through a shared underlying trait. As a result of this, however, a student’s apparent ability on the linked items will be a combination of their unidimensional ability and the unmodeled underlying trait ability. In the following model, θ_1^* is a student’s unidimensional ability score and θ_2^* is their ability score on the underlying trait in question. The strength of the ULD between the paired items can be represented by a ULD weight, wt_{ij}. This effect can be modeled in the following manner:

$$\begin{bmatrix} \theta_{\mathrm{item}\,1} \\ \theta_{\mathrm{item}\,2} \\ \theta_{\mathrm{item}\,3} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & \mathrm{wt}_{12} \\ 1 & \mathrm{wt}_{22} \\ 1 & 0 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} \theta_1^* \\ \theta_2^* \end{bmatrix} = \begin{bmatrix} \theta_1^* + \mathrm{wt}_{12}\theta_2^* \\ \theta_1^* + \mathrm{wt}_{22}\theta_2^* \\ \theta_1^* \\ \vdots \end{bmatrix}. \tag{2}$$

From Eq. (2), the effective student ability for item i ($\theta_{\mathrm{item}\,i}$) will be the ability a student uses to answer item i. This effective student ability has the potential to be significantly different from students’ actual unidimensional ability θ_1^* if the ULD weight is large. The effective student ability will be the ability with which a student will answer item i. This will artificially make the item appear more or less difficult compared to if the ULD were not present. The larger the ULD weight, the more significant the unmodeled trait is in determining how a student is responding to the items. On the other hand, if $\mathrm{wt}_{ij} = 0$, for all i and j, then no ULD exists and the items are all locally independent, provided no SLD exists between the items. The assessment can then be assumed to be unidimensional.

From a multitrait IRT perspective, the weights in the presented model of ULD can be interpreted as the ratio of the underlying trait item discrimination and the unidimensional trait discrimination, $\mathrm{wt}_{i2} = \alpha_{i2}/\alpha_{i1}$. This can be seen in the 2PL multitrait model:

$$\begin{aligned} P(\theta, \alpha_{i1}, \alpha_{i2}, d_i) &= \frac{1}{1 + \exp[-D(\alpha_{i1}\theta_1^* + \alpha_{i2}\theta_2^* + d_i)]} \\ &= \frac{1}{1 + \exp\{-D[\alpha_{i1}(\theta_1^* + \frac{\alpha_{i2}}{\alpha_{i1}}\theta_2^*) + d_i]\}} \\ &= \frac{1}{1 + \exp\{-D[\alpha_{i1}(\theta_1^* + \mathrm{wt}_{i2}\theta_2^*) + d_i]\}} \\ &= \frac{1}{1 + \exp[-D(\alpha_{i1}\theta_{\mathrm{item}\,i} + d_i)]}, \end{aligned} \tag{3}$$

where α_{i1} is the unidimensional trait’s item discrimination and α_{i2} is the ULD trait’s item discrimination. Equation (3) demonstrates how a multidimensional assessment could, incorrectly, be modeled unidimensionally by ignoring ULD. Consequentially, estimations of item parameters for the items influenced by the ULD will be inaccurate in a unidimensional framework.

Since student IRT ability scores are estimated using all items on an assessment, they will be relatively robust to the effects of ULD, and also SLD, provided “enough” items on the assessment are locally independent. The more items on the assessment that are locally independent, the more robust the estimated student abilities will be as a result. Thus, ULD and SLD can be expected to significantly impact linked-item parameter estimations, while leaving the estimated student abilities relatively unchanged [43,44].

The structure of ULD can be investigated using a multiple latent variable model like factor analysis or multitrait IRT, which both attempt to model underlying latent trait structures. This underlying latent trait structure has been well explored for the FCI; see Refs. [6,10,11]. These factor models can then be assumed to represent the ULD that exists within the FCI. It should be noted that these models can be influenced by the presence of SLD, and disentangling the effects of ULD from those of SLD is not currently well understood [32].

3. Detecting local dependence

A common way to detect LD between a pair of items is through the utilization of contingency tables. A contingency table displays the number of occurrences for a particular combination of events. For items m and n, the contingency table records the number of times the items were answered correctly and incorrectly simultaneously or one was answered correctly while the other was not. This is displayed as follows:

		Item n	
		0	**1**
Item m	**0**	O_{00}	O_{10}
	1	O_{01}	O_{11}

where O_{pq} is the observed number of occurrences when items m and n were answered correct/incorrect ($p = 0/1$ and $q = 0/1$).

Many useful statistics can be obtained from these tables when each element is derived from IRT-probability models. For instance, the expected number of occurrences for the contingency table above can be estimated from these probability functions. Since the parameters used in the IRT models are estimated assuming all of the items on the assessment are locally independent, any deviations between the observed and expected contingency tables can be assumed to be a result of LD between the item pairs.

Within the literature two statistics are commonly used to characterize the deviations between the observed and estimated contingency tables, Pearson's χ^2 and the logarithmic ratio G^2 statistic. Pearson's χ^2 can be calculated in the following manner:

$$\chi^2 = \sum_{p=0}^{1}\sum_{q=0}^{1}\frac{(O_{pq} - E_{pq})^2}{E_{pq}},$$

and the G^2 statistic is calculated using

$$G^2 = -2\sum_{p=0}^{1}\sum_{q=0}^{1} O_{pq} \ln\left(\frac{E_{pq}}{O_{pq}}\right),$$

where O_{pq} and E_{pq} are the observed and expected number of occurrences from the contingency table for the pair of items being investigated [45]. Both of these statistics will depend on the size of the sample being used and compare the observed and expected number of observations from the contingency table.

Since these statistics are sample-size dependent, it is often useful to employ a Cramer's V standardization to control for the sample size. This is given by

$$V_{\chi^2} = \sqrt{\frac{\chi^2}{n(k-1)}},$$

where n is the total number of observations and k is the number of rows in the contingency table; for this study, $k = 2$. A similar expression can be written for V_{G^2}.

Each of these statistics measure variations between observed and expected values of the contingency table. Thus, values closer to zero indicate good agreement between observations and expected results. Since the model's parameters are estimated assuming no LD, any deviations of V_{χ^2} and V_{G^2} away from 0 are an indication of potential LD linking between the items, which is unaccounted for by the model. These statistics have been found to be good indicators that LD exists [29–31,41].

It was recently demonstrated that the tetrachoric correlation can be used to detect LD independent of IRT models [32]. The tetrachoric correlation is a special case of the polychoric correlation used when the sample is dichotomous. The tetrachoric correlation is calculated numerically, but can be approximated as

$$r_{\text{tet}} \approx \cos\left(\frac{\pi}{1 + \sqrt{\frac{O_{00}O_{11}}{O_{10}O_{01}}}}\right),$$

where the argument of cos is in radians. Note that, if $r_{\text{tet}} = 0$, then $O_{00}O_{11} = O_{10}O_{01}$, which implies that there was no preference to answering both items correct/incorrect simultaneously. For example, if $r_{\text{tet}} < 0$, then students tended to answer one item correctly and the other incorrectly, and vice versa. Other types of correlations could be used to detect LD; however, the tetrachoric correlation tends to be more sensitive to correlations for dichotomous data [46].

Correlations are expected between items on assessments that probe a single concept (i.e., a unidimensional assessment). This is due to students answering items based on their latent ability, and not randomly (which would yield a correlation of zero). As a result, more difficult items will often be answered incorrectly together and visa versa for easier items. This generates nonzero correlations between the items on a single-conception assessment. When items are linked by ULD and/or SLD, the tetrachoric correlation will be artificially inflated. For this reason, item pairs with larger correlations than typical could potentially be linked via LD.

For brevity, V_{χ^2}, V_{G^2}, and the tetrachoric correlation together will be referred to as the "statistics of LD" for the remainder of the article.

4. Simulation specification

As it currently stands, no models exist that can differentiate between ULD and SLD for student responses. In order to understand the effects ULD and SLD may have on the statistics of LD, simulations using existing models were performed. This simulation methodology was proposed by Chen and Thissen [30], and was further used by Houts and Edwards [32] with minor variations. The main goal of these simulations was to identify whether LD is present and then how much of it can likely be modeled by ULD alone. This then allowed for the testing of statistics of LD, above which ULD can no longer reasonably account for all of the LD. That is, these simulations assume that all of the LD present on a theoretical assessment is varying levels of either ULD or SLD. This allows for a comparison between the statistics of LD for these simulations and actual student responses.

In order to generate a baseline to test for LD, 200 assessments of 30-items each were generated by randomly sampling 2PL item parameters. For each generated assessment, a class of 1000 students were assigned randomly sampled latent abilities. The student and item statistics were sampled in the following manner:

- $\theta \sim$ normal distribution(mean $= 0$, SD $= 1$)
- $\alpha \sim$ normal distribution(mean $= 1.7$, SD $= 0.3$)
- $d \sim$ normal distribution(mean $= 0$, SD $= 1$).

Locally independent dichotomous data were constructed for each of the randomly sampled class and assessment pairs. From these data the tetrachoric correlation matrix of the items for each simulated assessment was calculated. This tetrachoric correlation matrix served as a baseline for comparison when testing item pairs for LD; see Sec. III B 5.

Item characteristics and student abilities were then estimated from the locally independent simulated data. This enabled a fair comparison of the locally independent simulations and the LD simulations (see below) while controlling for possible differences due to the numerical estimation of the item parameters. The Cramer's V standardization of Pearson's χ^2 and G^2 (V_{χ^2} and V_{G^2}) were then calculated using these estimated values.

The statistics of LD used in this study are all bivariant in nature. As such, the calculation of these statistics is independent of the number of items on an assessment [32]. This was tested by running the simulations for 20-item assessments and 30-item assessments, and the resulting statistics of LD were found to be independent of the number of items on an assessment. All simulation results presented in this study used 30 items for each assessment.

Surface local dependence simulation.—To simulate SLD, items 3 and 4 of the simulated assessments were linked using Eq. (1). Simulations were run for π_{LD} values that ranged from 0 to 1 in steps of 0.01. For each value of π_{LD}, 200 simulated assessments were generated using the same criteria as discussed previously, while modifying item 4's responses as per the SLD model. It is important to note that the data for item 3 remained unchanged as a result of the SLD model used for this study. Student abilities and item parameters were then estimated for each of the modified simulated datasets, and statistics of LD were calculated. This resulted in a distribution of 200 values for each of the statistics of LD for every π_{LD} value.

Underlying local dependence simulation.—Similar to the SLD simulations, items 3 and 4 in the randomly generated assessments were treated as a pair of items linked by ULD, as modeled by Eqs. (2) and (3). For simplicity, the ULD weights for each item were taken to be the same. Simulations were run for ULD weight values that ranged from 0 to 5 in steps of 0.1. For each value of ULD weight, 200 simulated assessments were generated. Student abilities and item characteristics were then estimated and statistics of LD were calculated for each assessment. This resulted in a distribution of 200 values for the statistics of LD for each ULD weight values.

5. Identifying likely LD pairs

Identification of item pairs on an assessment that are potentially linked by LD was done using two different methodologies. The first method involved the development of cutoff values for the statistics of LD found from the simulations. Stricter cutoff values were generated from the ULD simulations to test for the potential presence of SLD between a pair of items. Since the simulation results are independent of the number of items on the assessment, the cutoff values presented in Table II can be used as they appear to test for LD and/or SLD on any assessment. This methodology requires a high level of LD between items for the pair to be flagged. The other method discussed uses one-tail t-testing to compare the simulation statistics of LD to those for an assessment. It should be noted that these methodologies yield different item pairs with the t-test method generally flagging more item pairs than the cutoff value method. This will be expanded upon in Sec. IV.

TABLE II. The proposed cutoff values for the statistics of LD for some given ULD weight values. The cutoff values for a ULD of "0" were used to detect the presence of general LD within an item pair.

ULD weight	0	1.5	2.0	2.5
r_{tet}	0.613	0.851	0.895	0.927
V_{χ^2}	0.060	0.455	0.564	0.650
V_{G^2}	0.060	0.478	0.596	0.693

Using cutoff values.—A pair of items is assumed to be linked by LD if their statistics of LD are significantly larger than those of the baseline model. To test for LD, cutoff values for the statistics of LD were generated from the results of the baseline simulation. If a pair of items had statistics of LD significantly above the generated cutoff values, then the pair was said to have LD between them. Generation of the cutoff values will be discussed below.

Items identified in this manner would violate local independence needed for IRT and CTT [24,25]. These item pairs would thus be a source of error for any unidimensional IRT models and CTT statistics of the assessment. As a result, IRT and CTT may not accurately model the assessment and thus multidimensional models should be considered, such as MIRT and factor analysis. However, these statistical frameworks model only ULD and do not accurately model SLD [30].

Differentiating between whether the LD is caused by SLD or ULD is currently not well understood for small to moderate severity. Coincidentally, midrange ($\pi_{\mathrm{LD}} = 0.4$–0.7) and high-end ($\pi_{\mathrm{LD}} = 0.7$–0.8) SLD severity result in particularly large statistics of LD. In order for ULD to result in similar statistics of LD, unusually large ULD weights must be used. These large ULD weights manifest in a 2PL model as larger slope parameters than typically found for conceptual assessments [32]. Given values of the statistics of LD, it can be inferred how likely it is for all of the LD between a pair of items to be explained with reasonable ULD weights. If it is not likely that ULD can account for all of the LD, then it can be assumed that some SLD must be present.

If it is reasonable to assume that ULD is the sole cause of any LD detected, then multiple latent trait models—like MIRT or factor analysis (FA)—can be used to properly model the assessment. Item pairs whose LD are unlikely to be a result of solely ULD imply that there is likely some SLD present for the item pair. Since SLD is not modeled in MIRT or FA, any item pairs in an assessment identified as likely possessing SLD would be a source of error. As a result, MIRT or FA would not be appropriate for the assessment.

To test for the existence of LD and to possibly distinguish between SLD and ULD, cutoff values for the statistics of LD were proposed using the distributions from the baseline and ULD simulations. Then, these cutoff values were compared to item pair statistics of LD for both the FCI and FMCE.

The cutoff values used in this study were taken to be the upper 95% confidence value of the statistics of LD's distributions generated by the simulations. The upper 95% confidence values are given by

$$\mathrm{Cutoff}(\mathrm{wt}) = \mu_{\mathrm{sim}}(\mathrm{wt}) + 1.667\sigma_{\mathrm{sim}}(\mathrm{wt}),$$

where wt is the ULD weight value being used to generate the cutoff values and μ_{sim} and σ_{sim} are the mean and standard deviation of the statistics of LD given a ULD weight. Four sets of cutoff values were generated for this study. One set was generated from the baseline simulations to test for the presence of LD, represented using $\mathrm{wt} = 0$. Three sets of cutoff values for the statistics of LD were generated to test for SLD using different ULD weights: $\mathrm{wt} = 1.5$, 2.0, and 2.5. These weights used for the analysis are larger than detected for typical conceptual assessments [32], but any ULD weight above 1.5 could be used to generate reasonable cutoff values to test for possible SLD. All of the cutoff values for ULD weights ranging from 0 to 5 in steps of 0.1 can be found in Table VII located in the Appendix.

The proposed cutoff values were compared to the distributions of the statistics of LD from student data. These distributions were generated by randomly sampling 200 classes of 1000 students for both the FCI and FMCE. From here, the mean and standard deviation for each of the item pairs were calculated and representative normal distributions were used to test the significance of the proposed cutoff values. If 95% of an item pair's distribution was found to be above the corresponding cutoff value, then the item pair was concluded to likely have LD and/or SLD.

Using t-testing.—An alternative method for identifying likely pairs of items that break local independence utilized t-testing for significance. This method compares randomly sampled student responses with the simulation results via a two-sample pooled t-test [47]. To test for the presence of LD, one can use the statistics of LD that resulted from the baseline simulation. If the distributions of an item pair's statistics of LD are significantly larger than the baseline's distributions, then the item pair is likely not locally independent. The possible presence of SLD can be inferred in a similar manner by using the ULD simulations with weights larger than 1.5. All of the t-testing done in this study used $\alpha = 0.001$ for the significance level.

To reiterate, these analyses only identify pairs of items that are *likely to possess* LD and/or SLD. An item pair that is determined to likely possess SLD implies that some amount of SLD is required to explain the statistics of LD observed between items. This does not suggest that all of the LD is accounted for by SLD alone, but that some amounts of ULD and SLD are likely.

IV. RESULTS

Presented below are the pairs of items flagged by the cutoff and t-testing methodologies for the FCI and FMCE. Item pairs were first tested for LD then subsequently tested for SLD.

A. Using cutoff values

Proposed cutoff values for the statistics of LD were generated using the procedure described in Sec. III. For reference, the cutoff values from the locally independent simulation can be found in the column labeled "0" within Table II. These results are independent of the total number of items, so the cutoff values in Table II can be used to test for the presence of LD in any assessment. For brevity, the results for the three individual statistics of LD will be given in a

$$(V_{\chi^2}\ \text{results}, V_{G^2}\ \text{results}, r_{\mathrm{tet}}\ \text{results})$$

format.

The postinstruction FCI was found to possess (17, 18, 9) pairs of items that are not locally independent, involving a total of (17, 18, 11) individual items. This represents 2%–4% of the total item pairs possible for the FCI and involves 30%–60% of the individual items. The preinstruction FCI was found to contain (21, 21, 8) pairs of items as not being locally independent, including (18, 18, 10) individual items. This accounts for 1.8%–5% of the total item pairs on the FCI and 33.33%–60% of the total items.

The postinstruction FMCE results found (249, 256, 222) item pairs as not being locally independent, involving a total of (45, 42, 45) individual items. These results account for 21%–24% of the total number of item pairs on the FMCE and 90%–96% of the individual items. The preinstruction results for the FMCE were more severe with (386, 390, 203) item pairs being identified as losing local item independence. This involved (47, 47, 40) individual items, representing 19%–30% of the total possible item pairs on the FMCE and 85%–100% of the items. For a summary of these results for the FCI and the FCME, see Table III.

From these results it can be seen that both assessments contain LD between items. This is expected since both assessments were originally constructed to measure multiple conceptions. Provided no SLD is present on either assessment, they both can be properly modeled using a multiple latent variable theory (such as MIRT or FA).

To test for the possible presence of SLD, cutoff values were generated using ULD weights of 1.5, 2.0, and 2.5, which can be found in the last three columns of Table II. In conjunction with the distributions of the statistics of LD

TABLE III. The number of item pairs identified via t-testing as potentially breaking local item independence for the FCI and FMCE pre- and postinstruction. The N column represents the number of item pairs detected. The % column denotes the percentage of the total item pairs made up by the values in the N column.

	V_{χ^2}		V_{G^2}		r_{tet}	
Assessment	N	%	N	%	N	%
Pre FCI	21	2.4	21	2.4	8	0.92
Post FCI	17	2.0	18	2.1	11	1.3
Pre FMCE	386	17.9	390	18.0	203	9.4
Post FMCE	249	11.5	256	11.8	222	10.3

from the student data, these cutoff values were used to identify item pairs where the LD is unlikely to be explained solely by ULD.

TABLE IV. The pairs of items identified as possibly linked by SLD on the FMCE via the cutoff values method. The numbers in the table separated by a colon indicate the question pairs whose statistics of LD are significantly larger than the cutoff. Items listed in regular text indicate pairs detected on both the pre- and postinstruction assessments, items in parenthesis indicate pairs flagged only on the preinstruction assessment, and bold for only the postinstruction assessment.

Cramer's V of Pearson's χ^2

ULD weight	1.5	2.0	2.5
	32:34 36:38	**36:38**	

Cramer's V of G^2

ULD weight	1.5	2.0	2.5
	32:34 **36:38**	**36:38**	

Tetrachoric correlation

ULD weight	1.5		2.0	2.5
	1:2	1:4	**1:2**	14:17
	3:7	8:9	1:4	(24:26)
	8:10	8:11	**8:9**	**36:38**
	9:12	11:12	(11:12)	
	11:13	14:16	**11:13**	
	14:17	14:18	**14:16**	
	14:19	16:17	14:17	
	16:18	16:19	16:18	
	17:18	(17:19)	16:19	
	18:19	22:23	22:23	
	(22:26)	24:26	24:26	
	27:28	**27:29**	(27:28)	
	32:34	36:38	**27:29**	
	46:47		32:34	
			36:38	

For the FCI, zero item pairs were flagged for potentially being linked through SLD for both pre- and postinstruction administrations. This implies that the detected LD on the FCI is likely a result of only ULD, which can be modeled using a multiple latent variable model. Thus, the act of chaining items on the FCI is not significantly affecting how students respond to items. Note that this result disagrees with previous literature, see Ref. [9], which found that item blocking was significantly affecting the results for the FCI. This suggests that more research should be performed to fully understand the effects item chaining has on this instrument.

The FMCE had many pairs of items flagged as potentially being linked through SLD. That is, the likelihood that ULD alone can account for the observed statistics of LD is very low, and thus there is likely a combination of both ULD and SLD linking the items. Table IV shows the items flagged by the cutoff value methodology for the FMCE.

For the postinstruction FMCE data, (2, 2, 25) pairs of items were identified as possibly being linked through SLD when using the smallest ULD weight (wt = 1.5). When considering the larger ULD weights, (1, 1, 13) and (0, 0, 2) item pairs were identified; see Table IV. For the preinstruction data, (1, 1, 21) pairs of items were identified as possibly being linked through SLD when using the smallest ULD weight. When considering the larger ULD weights, (0, 0, 9) and (0, 0, 2) item pairs were identified.

B. Using *t*-testing

The results presented using the cutoff values method were sufficient for revealing that both the FCI and FMCE do not posses local item independence. Consequentially, the results of the t-test method add little extra information to what has already been revealed about the existence of LD. The results presented below for the t-testing method investigated only the possible presence of SLD.

For the FCI, two pairs of items were identified as possibly being linked by SLD, items 5:18 and items 25:26. These item pairs were flagged using a ULD weight

TABLE V. The pairs of items identified as possibly linked by SLD on the FMCE via the t-testing method. The numbers in the table separated by a colon indicate the question pairs whose statistics of LD are significantly larger than the cutoff. Items listed in regular text indicate pairs detected on both the pre- and postinstruction assessments, items in parenthesis indicate pairs flagged only on the preinstruction assessment, and bold for only the postinstruction assessment. All of the found pairs have a significance of $p < 0.001$.

Cramer's V of Pearson's χ^2 and G^2

ULD weight	1.5		2.0	2.5
	(1:4)	**30:32**	32:34	**36:38**
	30:34	32:34	36:38	
	36:38	46:47		

TABLE VI. The pairs of items flagged on the FMCE and FCI as likely being linked with some SLD via the polychoric correlation and the t-testing method. All of the item number listed were found to have paired with the items number of the far left, for the varying ULD weight values. Items listed in regular text indicated pairs detected on both the pre- and postinstruction assessments, items in parenthesis indicate pairs flagged only on the preinstruction assessment, and bold for only the postinstruction assessment. All of the found pairs have a significance of $p < 0.001$.

FMCE–Tetrachoric correlation			
	ULD weight values		
Item	1.5	2.0	2.5
1	2, **3**, 4, 14, 16, (17), 18, 19	2, 4	**2**, 4
2	1, **3**, 4, 14, 16, 17, 18, 19	1, **4**, **14**	**1**
3	**1**, **2**, **4**, **6**, 7	7	**7**
4	1, 2, **3**, 14, 16, 18, 19	1, **2**	1
6	**3**, **7**		
7	3, **6**	3	**3**
8	9, 10, 11, 12, 13, (14), (16), 18, 21, 27, (28), (29)	9, 10, 11, 12, **13**	9, **10**, 11
9	8, 10, 11, 12, **13**, 28	8, **10**, 11, 12	8, 12
10	8, 9, **11**, 13	8, **9**, **13**	**8**
11	8, 9, **10**, 12, 13, (18), **21**, 27, 28, **29**	8, 9, 12, 13, **27**	8, 12, 13
12	8, 9, 11, 13, **27**, 28	8, 9, 11, **13**, **28**	9, 11
13	8, **9**, 10, 11, 12, **27**, **29**	**8**, **10**, 11, **12**	11
14	1, 2, 4, (8), 16, 17, 18, 19, (20), **23**, **24**, **25**, **26**	**2**, 16, 17, 18, 19	16, 17, 18, 19
16	1, 2, 4, (8), 14, 17, 18, 19, 20, **23**, **25**	14, 17, 18, 19	14, 17, 18, 19
17	(1), 2, 14, 16, 18, 19, (20), **23**, **24**, **26**	14, 16, 18, 19	14, 16, 18, (19)
18	1, 2, 4, 8, (11), 14, 16, 17, 19, 20, 21, **22**, 23, **24**, **25**	14, 16, 17, 19, (20), **23**	14, 16, 17, 19
19	1, 2, 4, 14, 16, 17, 18, 20, **23**, **25**	14, 16, 17, 18	14, 16, (17), 18
20	(14), 16, (17), 18, 19	(18)	
21	8, **11**, 18		
22	**18**, 23, 24, 25, 26	23, 24, 25, 26	23, (26)
23	**14**, **16**, **17**, 18, **19**, 22, 24, 25, 26	**18**, 22, 24, 26	22
24	**14**, **17**, **18**, 22, 23, 25, 26	22, 23, **25**, 26	26
25	**14**, **16**, **18**, **19**, 22, 23, 24, 26	22, **24**	
26	**14**, **17**, 22, 23, 24, 25	22, 23, 24	(22), 24
27	8, 11, **12**, **13**, 28, 29	**11**, 28, 29	28, **29**
28	(8), 9, 11, 12, 27, 29	**12**, 27, 29	27, (29)
29	**11**, **13**, 27, 28	27, 28	**27**, (28)
30	**31**, 32, 34	**32**, 34	
31	**30**, **32**, **34**		
32	30, **31**, 34	**30**, 34	34
34	30, **31**, 32	30, 32	32
36	38	38	38
38	36	36	36
40	42		
42	40		
44	**45**		
45	**44**		
46	47	**47**	**47**
47	46	**46**	**46**

FCI–Tetrachoric correlation			
	ULD weight values		
Item	1.5	2.0	2.5
5	18		
18	5		
25	26		
26	25		

of 1.5 and the tetrachoric correlation for pre- and postinstruction data. None of the other ULD weights or statistics of LD flagged any item pairs, pre- or postinstruction. It can be inferred that either the FCI has a small amount of SLD or these items are linked through a strong underlying trait.

The results of the *t*-test analysis for the FMCE can be found in Tables V and VI. The item pairs identified using the Cramer's V of Pearson's χ^2 and G^2 were identical. These statistics flagged 5 pairs, 2 pairs, and 1 pair of items when using ULD weights of 1.5, 2.0, and 2.5, respectively. The tetrachoric correlation flagged 78, 36, and 23 pairs of items for the ULD weights postinstruction. Similarly, for the preinstruction data 77, 35, and 23 item pairs were flagged. Because of the number of flagged item pairs on the FMCE, it is unlikely that all pairs can be explained via strong underlying traits alone.

V. DISCUSSION

Discussed herein is the likelihood that SLD invalidates multivariate models for the FCI and FMCE. Further, a discussion of unidimensional scoring and possible solutions is presented.

A. Multivariate models of the FCI and FMCE

As was presented in Sec. IV, most of the LD flagged item pairs on each assessment can be explained using only ULD.

For the FCI only two item pairs were flagged as potentially being linked, in part, via SLD. These were found using only the most lenient of testing criteria presented in this article. Thus, it can be assumed that these items are linked either by small π_{LD} values or by a very strong underlying trait. This implies that the error introduced to a multiple latent variable model of the FCI by these item pairs will likely be small. Researchers concerned about this error should perform their own investigation to test for the possible presence of SLD between items 5 and 18 and items 25 and 26.

Of the two instruments, the FMCE was found to contain far more SLD. For the most lenient testing criteria, 78 pairs of items were flagged as likely containing some amount of SLD (compared to the two found for the FCI). As a result, multiple latent variable models that assume local item independence may not accurately represent the FMCE.

Recall that in some cases item chaining is the practice of using the same figures, response pools, reading prompts, etc. for groups of items [23]. From this it can be seen that of the two item pairs flagged for the FCI, only the item pair 25∶26 actually meets this criteria. Thus, only this item pair on the FCI can be assumed to be impacted by item chaining. However, on the FMCE 78 item pairs were identified as likely being linked in part by SLD. Of these 78 identified items, more than half meet the criteria of being chained or blocked items. This implies that item chaining is having a significant impact on how the items on the FMCE are functioning.

In a previous study of the FMCE, Yang *et al.* obtained exploratory FA and MIRT models [35]. Since both exploratory FA and MIRT assume only ULD exists between items, the possible presence of SLD may significantly impact the results of these methodologies. In fact, a comparison of the flagged item pairs in Table VI and the results of the factor analysis presented in Ref. [35] reveals that many of the factors identified may be linked through a combination of ULD and SLD. Similarly, many of the major links present in the partial correlation networks are flagged in this analysis as likely containing some level of SLD. This, however, does not imply that the results presented in Ref. [35] are incorrect. The results of this study imply that some of the observed correlations used to generate the factor models and network structures are likely being artificially inflated due to SLD. To test the validity of the proposed models, further exploration into the effects of SLD on multivariate models is recommended.

B. Scoring the FCI and FMCE

Both assessments were found to violate the assumption of local item independence. This implies that unidimensional models, which assume local item independence, should not be used to analyze or score either of these assessments. The effects of LD on the results of unidimensional IRT models have been explored in detail; see Ref. [23]. Therein, Yen details the effect LD can have on total scores, assessment validity, IRT test information, and IRT item parameter estimations. The exact details of these effects are outside the scope of this study. From the results presented in Ref. [23] it can be inferred that unidimensional IRT will not be accurate if LD is present in an assessment.

Similarly, CTT statistics are affected by the presence of LD within an assessment. For example, the effects of SLD as a proxy of LD on classical item difficulty and discrimination are shown in Figs. 1 and 2. Within each figure the SLD influenced values are plotted versus the original locally item independent values for varying severity of SLD. Details concerning CTT statistics can be found in Ref. [48]. From these figures it is apparent that as SLD severity increases, the induced error of observed classical measures also increases. Because of the detection of potential SLD for the item pairs listed in Tables IV–VI, it can be assumed that the CTT statistics measured for these items are likely inaccurate. It can then be inferred that aggregate total scores made up of SLD linked item will not accurately reflect student knowledge. As these preliminary results show, the presence of LD on an assessment can drastically impact the observed CTT statistics for an assessment.

To address the possible effects of LD, Yen suggests using locally independent testlets (grouped items that are graded together) in place of locally dependent item blocks [23]. Testlets can be formed by grouping items that share LD,

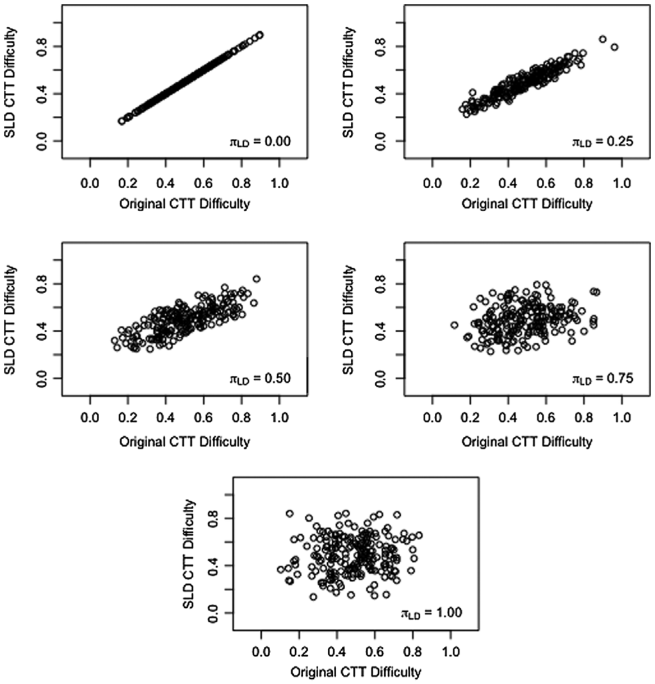

FIG. 1. Plots of the classical test theory item difficult indices of the SLD modified results versus the original item difficulties. Beginning from the top left, these plots are for $\pi_{\mathrm{LD}} = 0.00$, 0.25, 0.50, 0.75, and 1.00. Notice as π_{LD} gets larger, the classical difficulty becomes increasingly affected.

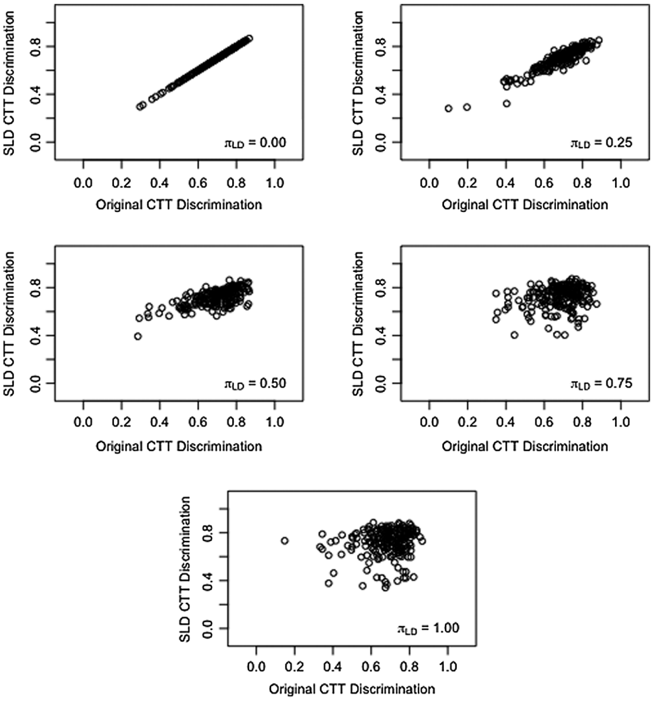

FIG. 2. Plots of the classical test theory item discrimination indices of the SLD modified results versus the original item discrimination. Beginning from the top left, these plots are for $\pi_{\mathrm{LD}} = 0.00$, 0.25, 0.50, 0.75, and 1.00. Notice as π_{LD} gets larger, the classical discrimination becomes increasingly affected.

then a grade can be assigned to each testlet individually. If each of the testlets is locally independent, then IRT can be performed by treating each of the individual testlets as “items.” This would result in a reliable measure of student ability, while controlling for the effects of LD.

Testlets for the FCI could be formed using one of the three models tested in Ref. [6], where each testlet would contain a single factor. This would result in students receiving one score for their understanding of each Newtonian concept represented by these models. These testlets could then be used to generate an IRT model of the FCI which would supply a student with their “Newtonian” ability.

Because of the potential presence of SLD on the FMCE, multiple scoring options may be considered: (1) forming testlets from LD linked items, (2) rearranging the items which appear on the FMCE, and (3) splitting the FMCE into two smaller assessments. Each of these methods contain drawbacks, and it is possible that the best option is to simply rewrite the instrument.

The testlet scheme for the FMCE proposed by Thornton *et al.* is well in line with Yen’s suggestion; however, some of the suggested testlets are not locally independent and thus do not meet all the specifications indicated by Yen [17,23]. The required modifications to ensure these testlets are locally independent would result in the FMCE being made up of one dominant testlet and many smaller testlets.

Rearranging the items on the FMCE would entail generating unique physical descriptions, response options, and figures for each of the items on the assessment. The separated items could then be randomized to ensure students are being primed for concepts as minimally as possible. The resulting assessment would need to undergo extensive analysis to fully elucidate its statistical properties. However, considering the initial reasons for chaining the items, and the added false-positive detection benefits, this “fix” may make the assessment something completely different than what the creators of the FMCE originally intended [2,16,17].

The splitting of FMCE could be done in a similar manner to how both the FCI and the Conceptual Survey of Electricity and Magnetism were separated in Refs. [49,50]. However, due to the potential SLD found in this study, any estimated characteristics for each individual item are likely not accurate, and will need to be reexamined after any of these suggested changes are made. It should be noted that these new assessments may not contain the same false-positive detection abilities as the original form of the FMCE. This would suggest that splitting the assessment in half may not be recommended; see Ref. [16].

Ultimately, a single-number grading scheme could be readily created for the FCI using current research results. However, given the extent of LD found and the large possibility of SLD being present, the FMCE should not be graded using a single number until a new grading scheme has been proposed and studied.

Considering the LD present on each of the assessments, only multivariate models should be used to assign scores for either of the assessments. Until the SLD which is likely present on the FMCE is better understood, any multidimensional measures for the FMCE should be treated as being inaccurate. Thus, the results of this study suggest that the FCI, and not the FMCE, be used in conjunction with a multidimensional model to probe the Newtonian understanding of students.

VI. LIMITATIONS

The data used in this study were a mixture of algebra- and calculus-based introductory physics courses. Some of the LD found in this study may belong more to one of these groups over the other. However, since the FCI and FMCE are intended to be used for both courses, the presence of potential LD suggests that both assessments should be modeled multidimensionally. Despite this, the potential presence of SLD on the FMCE is alarming and researchers should be aware of this critical limitation.

The interpretations made in this study assumed that LD was generated only through ULD and SLD. There could be effects other than ULD and SLD that generate some of the LD present on the FCI and FMCE. As it currently stands, the literature into LD does not offer any other classifications aside from ULD and SLD. The interpretations made in this study ignored the possibility of LD sources other than ULD and SLD.

The models used for SLD and ULD are extremely simplistic in nature and may not fully capture the effects. This may be particularly true for models that contain more traits, and thus better explain the ULD possibly linking items. In a future study, simulations will be performed using FMCE specific models to better replicate the ULD present, such as the MIRT model presented in Ref. [35]. Any unexplained LD could then be interpreted as potential SLD. Also, multiple models would need to be used to ensure the unexplained LD is likely SLD and not from unmodeled ULD. However, since interactions between questions, and groups of questions, can be broken up into collections of two-question interactions, the simple models used for the simulations here likely encapsulate the fundamental interactions between questions. As a result, the results presented in this study are not expected to change.

VII. SUMMARY

The FCI and FMCE were examined to determine the extent to which local independence was broken on each assessment. Local dependence occurs when multiple items influence one another in a manner that prevents students from responding to the items as though they were independent. Chen and Thissen [30] differentiated the causes of LD by defining two categories: surface local dependence and underlying local dependence.

When multiple items on an assessment share a common conception (or trait), it is said that ULD is linking the items. Since unidimensional IRT does not model multiple traits, its results are affected by ULD. Conveniently, the effects of ULD on an assessment can be accounted for by using multidimensional models when analyzing the assessment, such as factor analysis and multitrait item response theory.

Items which share common wording, figures, answer banks, reading passages, etc. or which are chained (blocked) together are likely linked via SLD. These effects cause students to answer items based entirely on how they responded to the previous SLD linked items. As a result, students do not interact with items independently; item independence is assumed by IRT and CTT. Since SLD is not due to a shared latent trait between the items, the effects of SLD cannot be accounted for by using multidimensional models. If an assessment contains SLD, then all statistics and models related to the assessment should be called into question for correctness and validity. For this reason, SLD is more concerning than ULD.

Simulations utilizing simple SLD and ULD models, Eqs. (1) and (3), constructed distributions of statistics of LD as functions of corresponding LD severity. The statistics of LD used in this study were Pearson's χ^2, G^2, and the tetrachoric correlation. Cutoff values for these statistics of LD were proposed based on the simulation results. These cutoffs were used to identify item pairs where the measured LD was unlikely to result entirely from ULD alone. It was then inferred that some level of SLD must link the items to account for all the LD present. Thus, the item pairs flagged using this methodology should reveal the effect that chaining items has on the FCI, FMCE, and other assessments.

A supplementary methodology utilized t-testing to compare the statistics of LD for the ULD simulations to those measured for the FCI and FMCE. If the statistics of LD for an item pair were found to be significantly larger than those generated by the ULD simulations, then the item pair was said to likely be linked by SLD.

RQ 1: To what extent is the assumption of local item independence valid for each assessment?

The assessments in question were found to contain item pairs that break local item independence. Of the items on the FMCE, anywhere from 85% to 100% where found to likely break the assumption of local item independence. For the FCI, 30%–60% of the items were found to lack local independence. As a result, results of unidimensional models for the FMCE will likely contain a significant amount of error. Models for the FCI, on the contrary, will likely contain some error, but not as much as models for the FMCE.

RQ 2: To what extent is SLD present on either of the assessments?

It was found that the FCI only had two item pairs with LD that could not be explained solely by ULD. This implies that the chaining of items on the FCI is not likely affecting student responses. This result disagrees with previous literature; see Ref. [9]. Further, the results presented in this study imply that the interactions between the FCI and students can likely be modeled using a multitrait model with minimal errors being introduced in the parameter estimations.

In comparison, the FMCE was found to likely have many SLD linked item pairs. This implies that the FMCE will need to be either graded in a special manner or modified to correct for the detected SLD. Suggestions of how to modify the FMCE were presented in Sec. V. Any grading model proposed for the FMCE will need to be studied in detail before it is used in practice or research. Until these studies have been performed, researchers and instructors should be aware that, due to the discovered statistical issues, scores reported using the FMCE will likely be inaccurate unless a corrected scoring procedure is used. This also holds true for reporting Hake gains and the normalized change of student scores on the FMCE [19,51].

RQ 3: By examining the item pairs identified in research question 2, is it correct to assume that item chaining is responsible for the SLD inferred?

On the FCI, only one pair of items (25 and 26) is assumed to be linked due to item chaining as a result of their close proximity. Whereas due to the manner in which the FMCE was constructed (distinct blocks of items that share response pools and figures), most of the item pairs identified as likely being artificially linked can be assumed to, at least partially, be a result of item chaining. Thus, it can be concluded that item chaining is likely having little effect on the FCI, but is having a significant impact on how students are responding to items on the FMCE. These results draw serious attention to what the FMCE is actually measuring due to the artificial influences likely present.

Since the results of the presented analysis may be data dependent, researchers analyzing and/or using any PER conceptual inventories are encouraged to assess the extent of local dependence present within the data being analyzed. This allows for a direct check of the assumption of local item independence. Implementing the ULD or SLD analysis discussed in this study, and subsequently eliminating all instances of detected SLD, will help researchers mitigate any possible effects of SLD. This will result in multitrait models generated from exploratory methods that are less likely to be a consequence of SLD.

VIII. IMPLICATIONS FOR PHYSICS EDUCATION RESEARCH

This article serves as a cautionary warning to all researchers and instructors who currently use the FMCE. Although it is possible that the simple models used in this study may not fully represent the behavior of LD, it is recommended that the future use of the FMCE be paused until more investigations of the instrument are completed. In the meantime, the FCI should be favored over the FMCE for use in research and in the classroom. Until the effects of SLD on multivariate models are better understood, any assessment which is found to likely contain SLD should not be used. That is, if future research into the effects of SLD on student responses finds that SLD significantly impacts the results of educational measures, then it is imperative all previous studies that used the FMCE reconfirm their results.

ACKNOWLEDGMENTS

The authors would like to thank PhysPort for allowing us to use their data. This project was funded by Stockton University.

APPENDIX: CUTOFF VALUES

Cutoff values generated using the methodology described Sec. III B 5 using cutoff values are located in Table VII. These cutoff values have been found to be consistent regardless of the number of items included on the modeled assessment. As a result, they can be applied to any assessment as given.

TABLE VII. The proposed cutoff values for ULD weights ranging from 0 to 5 in steps of 0.1. The bold faced rows are the cutoff values used in this study.

ULD weight	V_{χ^2}	V_{G^2}	r_{tet}
0.0	**0.064 225 41**	**0.064 161 94**	**0.620 856 8**
0.1	0.061 365 77	0.061 537 79	0.611 923 7
0.2	0.070 930 6	0.071 402 07	0.612 132 2
0.3	0.076 721 11	0.077 610 15	0.626 537 9
0.4	0.102 402 9	0.104 319 7	0.648 17
0.5	0.121 908	0.124 515 6	0.656 883 3
0.6	0.161 894 8	0.166 214 2	0.701 905 5
0.7	0.196 358 4	0.202 425 9	0.718 134
0.8	0.222 544 4	0.230 823 2	0.727 579 1

(Table continued)

TABLE VII. *(Continued)*

ULD weight	V_{χ^2}	V_{G^2}	r_{tet}
0.9	0.269 752 6	0.280 746 8	0.759 013 8
1.0	0.296 756 4	0.308 299	0.764 501 6
1.1	0.317 920 8	0.332 317 8	0.770 508 9
1.2	0.366 067 5	0.382 443 6	0.803 789
1.3	0.394 626 9	0.413 642 7	0.817 802 6
1.4	0.431 951 8	0.451 091 5	0.836 634 7
1.5	**0.454 700 8**	**0.478 395 2**	**0.851 213 2**
1.6	0.472 099 2	0.495 857 8	0.850 954 1
1.7	0.503 835 3	0.533 338 9	0.872 872 5
1.8	0.517 847 6	0.546 059 9	0.873 516 3
1.9	0.542 508 4	0.573 794 8	0.886 056 9
2.0	**0.563 872 2**	**0.595 884 9**	**0.894 588 2**
2.1	0.575 846 7	0.611 303 4	0.895 417 4
2.2	0.593 591 4	0.630 353 2	0.907 128 1
2.3	0.623 886 4	0.663 031 1	0.917 390 2
2.4	0.632 358 9	0.671 523 6	0.918 420 2
2.5	**0.6504**	**0.692 684 1**	**0.926 625 1**
2.6	0.667 088 1	0.711 463 7	0.934 592 9
2.7	0.678 400 3	0.723 765 6	0.936 116 2
2.8	0.688 925 8	0.735 541 8	0.942 722 6
2.9	0.690 357 8	0.739 041 1	0.940 076 1
3	0.707 513 1	0.758 766	0.950 447
3.1	0.718 474	0.769 870 5	0.949 793 9
3.2	0.728 998 7	0.782 845	0.954 050 1
3.3	0.732 553 2	0.788 292 5	0.960 074 3
3.4	0.742 782 8	0.798 563 8	0.956 894 5
3.5	0.756 979 4	0.813 575 6	0.963 918 9
3.6	0.758 892 5	0.817 718 1	0.960 478 6
3.7	0.760 669 9	0.821 852 4	0.963 336 1
3.8	0.777 562 4	0.841 649	0.970 911
3.9	0.777 625 2	0.840 604 7	0.966 171 2
4	0.801 236 4	0.864 007 6	0.973 19
4.1	0.793 261	0.856 92	0.970 685 6
4.2	0.803 636 9	0.868 948 1	0.972 953 1
4.3	0.805 597 6	0.874 273 3	0.978 052 5
4.4	0.812 301	0.878 827 3	0.974 135 6
4.5	0.811 962 4	0.880 987 4	0.975 841 8
4.6	0.812 054 8	0.881 005 2	0.974 725 1
4.7	0.825 387 8	0.895 317 4	0.977 706 8
4.8	0.826 369 1	0.901 741 4	0.981 786 2
4.9	0.835 909 5	0.909 628 3	0.979 880 9
5.0	0.835 386	0.909 959	0.981 603 6

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

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

[3] S. B. McKagan, PhysPort, 2011, https://www.physport.org.

[4] 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).

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

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

[7] 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).

[8] 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).

[9] 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).

[10] 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).

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

[12] 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).

[13] 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).

[14] G. Davenport, The reliability of the Force and Motion Conceptual Evaluation, Masters Thesis, University of Maine, 2008.

[15] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, Am. J. Phys. **76**, 882 (2008).

[16] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. ST Phys. Educ. Res. **4**, 020101 (2008).

[17] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the Force and Motion Conceptual Evaluation and the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **5**, 010105 (2009).

[18] J. Von Korff, B. Archibeque, K. Alison Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, Am. J. Phys. **84**, 969 (2016).

[19] 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).

[20] D. Hestenes and M. Wells, A mechanics baseline test, Phys. Teach. **30**, 159 (1992).

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

[22] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[23] W. M. Yen, Scaling performance assessments: Strategies for managing local item dependence, J. Educ. Measure. **30**, 187 (1993).

[24] R. F. DeVellis, Classical test theory, Med. Care **44**, S50 (2006).

[25] R. J. de Ayala, *The Theory and Practice of Item Response Theory* (The Guilford Press, New York, NY, 2009).

[26] F. M. Lord, The relation of test score to the trait underlying the test, Educ. Psychol. Meas. **13**, 517 (1953).

[27] A. L. Van den Wollenberg, Two new test statistics for the Rasch model, Psychometrika **47**, 123 (1982).

[28] R. E. Traub, A priori considerations in choosing an item response model, Appl. Item Response Theory **57**, 70 (1983).

[29] W. M. Yen, Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, Appl. Psychol. Meas. **8**, 125 (1984).

[30] W.-H. Chen and D. Thissen, Local dependence indexes for item pairs using item response theory, J. Educ. Behav. Stat. **22**, 265 (1997).

[31] C. R. Houts and M. C. Edwards, The performance of local dependence measures with psychological data, Appl. Psychol. Meas. **37**, 541 (2013).

[32] C. R. Houts and M. C. Edwards, Comparing surface and underlying local dependence levels via polychoric correlations, Appl. Psychol. Meas. **39**, 293 (2015).

[33] L. R. Fabrigar and D. T. Wegener, *Exploratory Factor Analysis: Understanding Statistics* (Oxford University Press, Inc., New York, 2012).

[34] J. S. Aslanides and C. M. Savage, Relativity concept inventory: Development, analysis, and results, Phys. Rev. ST Phys. Educ. Res. **9**, 010118 (2013).

[35] J. Yang, C. Zabriskie, and J. Stewart, Multidimensional item response theory and the Force and Motion Conceptual Evaluation, Phys. Rev. Phys. Educ. Res. **15**, 020141 (2019).

[36] B. Van Dusen and J. Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear modeling, Phys. Rev. Phys. Educ. Res. **15**, 020108 (2019).

[37] J. Wells, R. Henderson, A. Traxler, P. Miller, and J. Stewart, Exploring the structure of misconceptions in the Force and Motion Conceptual Evaluation with modified module analysis, Phys. Rev. Phys. Educ. Res. **16**, 010121 (2020).

[38] F. M. Lord and M. R. Novick, *Statistical Theories of Mental Test Scores* (IAP – Information Age Publishing Inc., Charlotte, NC, 2008).

[39] R Core Tea, R: A language and environment for statistical computing, http://www.R-project.org/.

[40] R. Philip Chalmers *et al.*, MIRT: A multidimensional item response theory package for the Rr environment, **48**, 1 (2012).

[41] M. Wilson, Detecting and interpreting local item dependence using a family of Rasch models, Appl. Psychol. Meas. **12**, 353 (1988).

[42] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. (The Guilford Press, New York, NY, 2015).

[43] C. Magno, Demonstrating the difference between classical test theory and item response theory using derived test data, Int. J. Educ. Psychol. Assess. **1**, 1 (2009).

[44] R. K. Hambleton and R. W. Jones, Comparison of classical test theory and item response theory and their applications to test development, Educ. Meas. **12**, 38 (1993).

[45] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice* (The MIT Press, Cambridge, MA, 1975).

[46] K. D. Kubinger, On artificial results due to using factor analysis for dichotomous variables, Psychol. Sci. **45**, 106 (2003).

[47] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability & Statistics for Engineers & Scientists* (Prentice-Hall, Englewood Cliffs, NJ, 2012).

[48] R. L. Doran, Implications for measurement and evaluation from the trends of science education, Sci. Educ. **60**, 199 (1976).

[49] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the Force Concept Inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[50] Y. Xiao, K. Koenig, J. Han, Q. Liu, J. Xiong, and L. Bao, Test equity in developing short version conceptual inventories: A case study on the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **15**, 010122 (2019).

[51] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).