2019

Exploring the Structure of Misconceptions in the Force Concept Inventory with Modified Module Analysis

James Wells
Science Department of Claremont McKenna, Pitzer

Rachel Henderson
Michigan State University

John Stewart
West Virginia University, jcstewart1@mail.wvu.edu

Gay Stewart
West Virginia University, gbstewart@mail.wvu.edu

Jie Yang
West Virginia University

See next page for additional authors

Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis

James Wells,[1] Rachel Henderson,[2] John Stewart,[3,*] Gay Stewart,[3]
Jie Yang,[3] and Adrienne Traxler[4]

[1]*W. M. Keck Science Department of Claremont McKenna, Pitzer, and Scripps Colleges, Claremont, California 91711, USA*
[2]*Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan 48824, USA*
[3]*West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506, USA*
[4]*Wright State University, Department of Physics, Dayton, Ohio 45435, USA*

(Received 17 May 2019; published 3 September 2019)

Module analysis for multiple-choice responses (MAMCR) was applied to a large sample of Force Concept Inventory (FCI) pretest and post-test responses ($N_{pre} = 4509$ and $N_{post} = 4716$) to replicate the results of the original MAMCR study and to understand the origins of the gender differences reported in a previous study of this dataset. When the results of MAMCR could not be replicated, a modification of the method was introduced, modified module analysis (MMA). MMA was productive in understanding the structure of the incorrect answers in the FCI, identifying 9 groups of incorrect answers on the pretest and 11 groups on the post-test. These groups, in most cases, could be mapped on to common misconceptions used by the authors of the FCI to create distractors for the instrument. Of these incorrect answer groups, 6 of the pretest groups and 8 of the post-test groups were the same for men and women. Two of the male-only pretest groups disappeared with instruction while the third male-only pretest group was identified for both men and women postinstruction. Three of the groups identified for both men and women on the post-test were not present for either on the pretest. The rest of the identified incorrect answer groups did not represent misconceptions, but were rather related to the blocked structure of some FCI items where multiple items are related to a common stem. The groups identified had little relation to the gender unfair items previously identified for this dataset, and therefore, differences in the structure of student misconceptions between men and women cannot explain the gender differences reported for the FCI.

DOI: 10.1103/PhysRevPhysEducRes.15.020122

I. INTRODUCTION

The "gender gap," gender differences between the scores of men and women on the Force Concept Inventory (FCI) [1] and other instruments developed by physics education research (PER), has been extensively studied (see the review by Madsen, McKagan, and Sayre [2]). For the FCI, a substantial number of studies have suggested that some of the gender differences observed resulted from different response patterns of men and women to a subset of the items in the instrument; see Traxler *et al.* for an overview of this research [3]. The origin of these differential response patterns is, however, unknown.

A. Research questions

The purpose of this study is to apply module analysis for multiple-choice responses (MAMCR) introduced by Brewe *et al.* [4] to a large sample of FCI responses known to contain a subset of items that produce substantially different response patterns for men and women in order to determine if the structure of the misconceptions of men and women differ on these items. When the MAMCR method did not yield productive results for the large sample in this study, the reasons for the failure of MAMCR were explored and a modification to the algorithm was proposed called "modified module analysis." The modified algorithm was used to explore gendered differences in the patterns of incorrect answers on the FCI.

In general, network analysis uses the term "community" and "module" interchangeably to represent connected (under some definition) subsets of a network. We adopt the term community instead of module in anticipation of the "igraph" package [5] in the "R" software system [6] becoming the primary network analysis tool in PER.

*jcstewart1@mail.wvu.edu

In igraph, algorithms to detect structure within a graph are called "community detection" algorithms.

This study explored the following research questions:

RQ1 Are the results of module analysis for multiple-choice responses replicable for large FCI datasets? If not, what changes to the algorithm are required to detect meaningful communities of incorrect answers?

RQ2 How do the communities detected change as network-building parameters are modified? Do these changes support the existence of a coherent non-Newtonian conceptual model?

RQ3 How is the incorrect answer community structure different between the pretest and the post-test?

RQ4 How is the incorrect answer community structure different for men and women? Do the differences explain the gender unfairness identified in the instrument?

This work extends the module analysis technique to a larger dataset, explores alternate choices during that analysis, and contrasts structure between pre- and post-test data. Structural clues in the community structure are examined to explain unresolved questions about gender differences in answer choices [3].

B. Previous studies of the FCI

The FCI, either in aggregate or disaggregating by gender, is one of the most studied instruments in PER. The present study examines item-level structure disaggregated by gender. The structure of the incorrect answers is examined to identify coherent patterns of incorrect answers.

1. Exploratory analyses of the FCI

Many studies have examined the structure of the FCI, primarily using exploratory factor analysis (EFA). These studies began soon after the publication of the FCI when Huffman and Heller [7] failed to extract the factor structure suggested by the authors of the instrument [1], identifying only one factor for a sample of university students. A later work by Scott, Schumayer, and Gray [8] applied EFA to FCI post-test scores and found an optimal model with 5 factors; however, one of the factors explained much of the variance. The result that a single factor explains the majority of the variance is fairly robust and is further supported by the high Cronbach α values reported [9]. Scott and Schumayer [10] replicated their 5 factor analysis using Multidimensional Item Response Theory (MIRT) on the same sample. Semak *et al.* [11] reported optimal models with 5 factors on the pretest and 6 factors on the post-test for calculus-based introductory physics students. Stewart *et al.* also performed EFA using MIRT and reported 9 factors as optimal [12].

2. Gender and the FCI

In an extensive review of gender differences on physics concept inventories [2], men outperformed women by 13% on pretests and 12% on post-tests of conceptual mechanics: the FCI and the Force and Motion Conceptual Evaluation [13].

Many reasons have been explored to explain these differences. Differences in high school physics class election [14–16] may cause differences in college physics grades [17,18]; FCI scores correlate with physics grades. Cognitive differences have also been advanced as explanations of academic gender differences [19–22] with women scoring generally higher on verbal reasoning tasks and men scoring generally higher on spatial reasoning tasks. Psychocultural factors have also been explored as explanations of academic performance differences including mathematics anxiety [23,24], science anxiety [25–27], and stereotype threat [28]. For a more detailed discussion about the many sources that may influence the overall gender differences on physics conceptual inventories, see Henderson *et al.* [29].

3. Item fairness and the FCI

In addition to student-centered explanations for conceptual inventory gender differences, bias in the individual FCI items has been investigated as a source of these gender differences. McCullough and Meltzer [30] randomly gave students the original FCI or a version where each problem's context was modified to be more stereotypically familiar to women and found significant differences in performance on multiple items. Multiple studies have reported item unfairness in unmodified items in the FCI [3,31,32]. Recent research has suggested that other commonly used conceptual physics instruments do not contain a substantial number of unfair items [33]. Traxler *et al.* provide a thorough summary of research into the item fairness of the FCI [3].

C. Misconceptions and the FCI

Since the early 1980s, student difficulties, most commonly known as "misconceptions" or "alternate conceptions or hypotheses," have been extensively studied within physics classrooms. The early work done by Clement and colleagues [34–36] qualitatively analyzing the "alternate view of the relationship between force and acceleration" that are grounded in students' experiences has influenced much of the research examining conceptual understanding in physics. Halloun and Hestenes [37,38] further explored this idea by collecting a taxonomy of "common sense concepts" that conflict with student understanding of Newtonian mechanics. Hestenes, Wells, and Swackhamer developed the FCI [1] with the intent of measuring student conceptual understanding of Newtonian theory, specifically analyzing student misconceptions pre- and post-instruction [39].

The authors of the FCI provided a detailed description of the misconceptions measured by the instrument [1]. A summary of those misconceptions follow.

Impetus.—Dating back to pre-Galilean times, the impetus model involves the idea that an object has a "motive power" that can explain why an object remains in motion

regardless of any external forces [1,38]. Students with this misconception do not fully understand Newton's 1st law. For example, FCI items 6 and 7 describe a ball moving in a circle and ask about the path the ball will take after it exits a circular path. Selecting the circular trajectory after exiting the track demonstrates the misconception that the ball has a circular impetus.

Active force.—The misconception that motion implies force involves the idea that an object in motion must be experiencing a force. This misconception involves a naive understanding of the difference between velocity and acceleration [1,34] and demonstrates that Newton's 2nd law is not well understood. For example, item 11 asks about the forces on a hockey puck traveling on a frictionless surface after it has been kicked. The motion implies force misconception would predict that there is a force in the direction of motion; response 11C describes the forces on the puck as "a downward force of gravity, an upward force exerted by the surface, and a horizontal force in the direction of motion" [1].

Action-reaction pairs.—The misconception that the larger object exerts a greater force on a smaller object stems from the "dominance principle" [1,38]. This misconception demonstrates that Newton's 3rd law is not well understood. For example, items 4 and 15 describe a small car pushing a large truck and ask the student to describe the forces between the two objects. The dominance principle misconception would predict that the truck exerts a larger force on the car than the car exerts on the truck. In addition to the dominance principle, there are other misconceptions related to the naive understanding of Newton's 3rd law. Items 15, 16, and 28 also test the misconception that the most active agent produces the greatest force.

Concatenation of influences.—This misconception involves the idea that forces influence with "one force winning out over the other" [1]. This misconception demonstrates that the superposition principle for Newtonian forces is not well understood. For example, items 8 and 9 describe a hockey puck sliding horizontally at a constant speed on a frictionless surface. These items ask for the path that the hockey puck would take and the speed of the puck after it receives a swift kick. The misconception of "one force winning" would predict that the last force (i.e., the swift kick) determines the motion and speed of the puck.

Gravity.—The misconception that gravity is not a force stems from the Aristotelian physics idea that heavier objects tend to move toward the center of Earth and lighter objects tend to move away from the center of Earth [1,38]. For example, FCI item 1 describes two metal balls of different weights that are dropped at the same time; the item asks about the amount of time it takes for the two balls to hit the ground. The gravity misconception predicts that the heavier ball falls faster.

The above descriptions are only a few examples of the misconceptions measured by the FCI; others can be found in Hestenes and Jackson's detailed taxonomy [40].

Recently, quantitative studies have been used to begin to further understand the misconception structure of the FCI. Scott and Schumayer [41] applied EFA to all 150 responses, 5 per item, on the FCI pretest. The two most important factors each contained responses from the majority of the items in the FCI; contained both incorrect and correct responses; and mixed conceptually very different correct reasoning. Eaton, Vavruska, and Willoughby [42] replicated this work for both pretest and post-test data; no consistent theme could be identified for multiple factors in their study. The failure of these studies to identify an intelligible factor structure containing items requiring related Newtonian reasoning may indicate that factoring the incorrect and correct responses together in the same analysis is not productive.

Scott and Schumayer provided additional analysis of two of their factors using network analytic techniques [43]. As in this work, the network was constructed using the correlation matrix; however, only correlations within the factors identified in their early factor analysis were considered. This work reported node centrality measures, but did not use the community detection methods of MAMCR.

D. Theories of knowledge

Many researchers have investigated students' conceptual understanding by exploring the misconceptions outlined above. Early research explored the overall common difficulties and beliefs that students had about Newtonian mechanics [44–50]. More recently, researchers have designed systematic studies to explore student understanding and the epistemological development of Newton's laws of motion [13,51–54]. For example, Rosenblatt and Heckler developed a new assessment to investigate student understanding of the relationship between force, velocity, and acceleration [52]. This study found that understanding the relationship between velocity and acceleration was necessary to understanding the relationship between velocity and force; however, the reverse was not necessarily true.

Modeling coherent patterns of student wrong answers as misconceptions is only one of many ways to explain patterns of reasoning about mechanics. Other important theories include knowledge in pieces [55,56] and ontological categories [57–59]. The knowledge-in-pieces framework posits that student knowledge is formed of a number of granular pieces of reasoning that are activated either individually or collectively to produce a solution. These reasoning pieces have been explored by many authors and have been conceptualized as phenomenological primitives (*p* prims) [55,56], resources [60–62], or facets of knowledge [63]. In the knowledge-in-pieces framework, misconceptions become coherently activated sets of *p* prims. Unlike the misconception view, the knowledge-in-pieces view identifies positive intellectual components which a instructor can activate to encourage the knowledge construction process.

The relation of the misconception view and the knowledge-in-pieces framework is complex. For a careful and accessible exploration of the relation and differences, see Scherr [64]. We adopt the definitions from this work. The misconception model refers to "a model of student thinking in which student ideas are imagined to be determinant, coherent, context-independent, stable, and rigid" [64]. The knowledge-in-pieces model views student ideas "as being at least potentially truth-indeterminate, independent of one another, context-dependent, fluctuating, and pliable" [64].

The ontological categories framework is substantially different than either the misconception view or the knowledge-in-pieces view; the ontological categories framework proposes that incorrect student answers result from a misclassification of a concept. For example, misclassification of the concept of force as a substance that can be used up might lead a student to predict an object would come to rest after the applied force is removed. A substantial amount of research has also investigated how students' conceptual knowledge changes over time [65].

The framework chosen, knowledge-in-pieces, misconceptions, or ontological categories, has different consequences for instruction or curriculum design in how they draw out and make use of student ideas [61]. However, it is less clear that this difference is measured by conceptual inventories. Incoherence in student answers for the same concept might suggest a knowledge-in-pieces view, where different problem contexts can trigger different p prims even if a physicist would see the scenarios as isomorphic. However, the FCI was not designed to measure this effect, and as such, a separate instrument designed around the knowledge-in-pieces or ontological categories frameworks is likely required to fully explore either framework.

The quantitative method in the present work identified small segments of incorrect reasoning. Intrinsically, the method applied identifies incorrect reasoning applied across items with multiple contexts suggesting the structures identified are better described as misconceptions using the above definitions. In addition, the FCI was strongly developed within the misconception view and, therefore, the current work will primarily employ the misconception description of novice understanding [66,67]. We note in Sec. III where alternative frameworks seem relevant. Ultimately, while we call groups of incorrect answers identified by network analytic techniques "misconceptions," this work is purely quantitative and cannot distinguish between the various theoretical frameworks developed to explain incorrect answering patterns.

E. Background studies

This work drew heavily from three previous studies which will be referenced as study 1, study 2, and study 3 in this work.

1. Study 1: Module analysis

In study 1, Brewe, Bruun, and Bearden introduced module analysis for multiple-choice responses (MAMCR) to analyze concept inventory data at the level of individual responses to the items [4]. Unlike many analysis techniques applied to FCI data, which consider only a student's overall score or only the correct answers to individual items, MAMCR considers each answer choice a student selected in order to provide a fine-grained examination of students' misconceptions of Newtonian physics and to allow instructors to target specific errors.

MAMCR is based on network analytic techniques [68,69]. A network is represented by a graph where nodes are connected to one another by edges. Edges can be weighted, where the value of the weight represents some aspect of the interaction. Network analysis is a highly successful and versatile set of methods which have been applied to a variety of problems including the probability of homicide victimization among people living in a disadvantaged neighborhood [70], the mapping of functional networks in the brain from electrical signals [71], passing patterns of soccer teams in the World Cup [72], and the response of plants to bacterial infection [73].

Study 1 examined the FCI post-test scores of 143 first-year physics majors at a Danish university. The sample was 78% male and scored relatively highly on the exam: pretest $65 \pm 22\%$ and post-test $81 \pm 18\%$.

To analyze the FCI, each response was assigned to a node in the network; for example, if a student selected the choice "D" on FCI item 1, then 1D would be a node. An edge was added for each time a student selected two responses; for example, if a student selected 1D and 2E, then an edge was drawn connecting 1D and 2E. The correct responses were removed from the network leaving the network of incorrect responses.

In order to find connected responses in the network, a community detection algorithm (CDA) was applied to the network. There are many different types of CDAs [74]; in study 1, the Infomap algorithm was chosen [75].

Study 1 identified nine modules, each representing a separate misconception in student thinking. Two modules could be clearly interpreted: module 1 "the impetus model" and module 2 "more force yields more results." The other seven modules were more difficult to interpret.

In study 1, Brewe *et al.* emphasized that the results of this study should be generalized with care. The group of students tested was small, unusually high scoring, and had limited diversity. Likewise, there were several choices made during the process of applying MAMCR to the data which could have been made differently. Both the choice of sparsification method (described in Sec. III) and the decision of how to group responses that cluster together only on some of the one-thousand applications of Infomap was somewhat arbitrary, as was the interpretation of the meaning

of the modules. As will be seen in Sec. III, our dataset required different choices to be made.

The current work will conclude that MAMCR does not scale to larger datasets and that the modules identified in study 1 were the result of the small sample size and some of the decisions made in applying the algorithm; as such, the results of study 1 will not be further discussed in this work.

2. Study 2: Item fairness and the FCI

In study 2, Traxler *et al.* [3] explored item-level gender fairness of the FCI using classical test theory [76], item response theory [77], and differential item functioning (DIF) [78,79] analysis. An item is fair to men and women if men and women of equal overall ability score equally on the item. Using three samples, a graphical analysis identified five FCI items that were substantially unfair to women: item 14 (bowling ball falling out of an airplane), items 21 through 23 (sideways-drifting rocket with engine turning on and off), and item 27 (a large box being pushed across a horizontal floor). A further DIF analysis, which controlled for the student's overall post-test score, identified eight items on the FCI as substantially unfair. These eight items included the five items identified in the graphical analysis along with items 9, 12 (the trajectory of a cannon ball shot off of a cliff), and 15. Many of the unfair items had been identified as unfair in previous studies [3,30–32]. Two of these were unfair to men: item 9 (speed of a puck after it receives a kick) and item 15 (a small car pushing a large truck). Overall, study 2 demonstrated that eliminating all unfair items on the FCI to create a fair instrument reduced the gender gap by 50% in the largest sample.

Study 2, however, could not identify the source of the unfairness. The distribution of student responses was analyzed. Focusing on the five items that were identified with the graphical analysis and the DIF analysis, incorrect female responses were predominately one of the distractors in each of the FCI items; however, the distractors chosen by the male students were less uniform in all five FCI items. Overall, study 2 concluded that no physical principle or common misconception could explain the unfairness identified in these FCI items; however, this conclusion was drawn from a qualitative inspection of the items. The current study builds on the work in study 2 by performing a quantitative analysis of the incorrect responses of men and women.

3. Study 3: Multidimensional Item Response Theory and the FCI

In study 3, Stewart *et al.* examined the correct answer structure of the FCI using both exploratory and confirmatory methods [12]. The study in the current work applies network analytic methods to understand the incorrect answer structure of the FCI. This structure might be influenced by features of the FCI which produce correlations between the correct answers. If a consistent misconception is being applied, it would form an alternate incorrect answer to sets of related correct answers. In study 3, exploratory factor analysis (EFA) suggested that the practice of "blocking" items produced correlations between the items within the block. A block of items is a sequence of items which all refer to a common stem or where one item refers to a previous item. Blocking has also been called "item chaining" in previous studies [80]. The FCI contains item blocks $\{5,6\}$, $\{8,9,10,11\}$, $\{15,16\}$, $\{21,22,23,24\}$, and $\{25,26,27\}$. Study 3 reported that often the factors identified by EFA strongly loaded on items in the same block, suggesting that blocking was generating correlations among the items in the block. Study 3 concluded that exploratory methods such as EFA were not productive in understanding the physical concepts measured by the FCI.

Study 3 went on to produce a detailed model of the reasoning required to solve the FCI. MIRT was used to test alternate models and allowed the identification of an optimal model. This model allowed the identification of groups of items with very similar solution structure: $\{5,18\}$, $\{6,7\}$, $\{17,25\}$, and $\{4,15,28\}$. The items in each of these groups require the same set of logical reasoning for their solutions; as such they differ by only surface features and experts would view them as equivalent problems. The misconception view asserts that misconceptions are context-independent and, therefore, students should hold the same misconception for each item in the group. Study 3 only included the first item in a block in the analysis and it is likely that item 16 should be added to the last block which represents Newton's 3rd law items. This mapping of item blocks and groups with similar solution will be important to understanding the incorrect answer structure presented in this work.

II. METHODS

A. Instrument

The FCI is a 30-item instrument designed to measure a student's facility with Newtonian mechanics [1]. Each item includes one correct response and four incorrect responses. The instrument includes items involving Newton's three laws as well as items probing an understanding of one- and two-dimensional kinematics. The instrument does not cover many topics in Newtonian mechanics; for example, conservation of energy and momentum are not covered. The instrument was also constructed with distractors representing common student misconceptions. The instrument was revised after its initial publication [81]; this study uses the revised instrument. The revised instrument is available at PhysPort [82].

B. Sample

The data for this study were collected at a large southern land-grant university serving approximately 25 000 students. Overall university undergraduate demographics were 79% White, 5% African American, 6% Hispanic, and other groups each with 3% or fewer [83].

The sample was collected in the introductory calculus-based mechanics class serving primarily physical scientists and engineers. The sample has been analyzed previously by Traxler *et al.* (study 2) [3]; it is referenced as sample 1 in that work. The sample contains 4716 complete FCI post-test records (3628 men and 1088 women) and 4509 complete pretest records (3482 men and 1027 women). Table II in study 2 reports basic descriptive statistics. On the pretest, men have an average percentage score of 43%, women 32%. On the post-test, men have an average percentage score of 73%, women 65%. The course in which the sample was collected was presented using the same pedagogy and managed by the same lead instructor for the period studied. A more thorough discussion of the sample and the instructional environment may be found in study 2.

C. Analysis methods

Initial replication of study 1 was performed with the Infomap software available from mapequation.org [84]. All other statistical analysis was performed in the R statistical software system [6]. This work failed to replicate the study 1 results and proposes a modified analysis method; as such, the analysis method is a result of the work and the various network techniques employed are described as they are used.

III. RESULTS

A. Module analysis

Figure 1 outlines the original and modified analysis steps. The original module analysis method presented in study 1 first formed a bipartite network, a network that

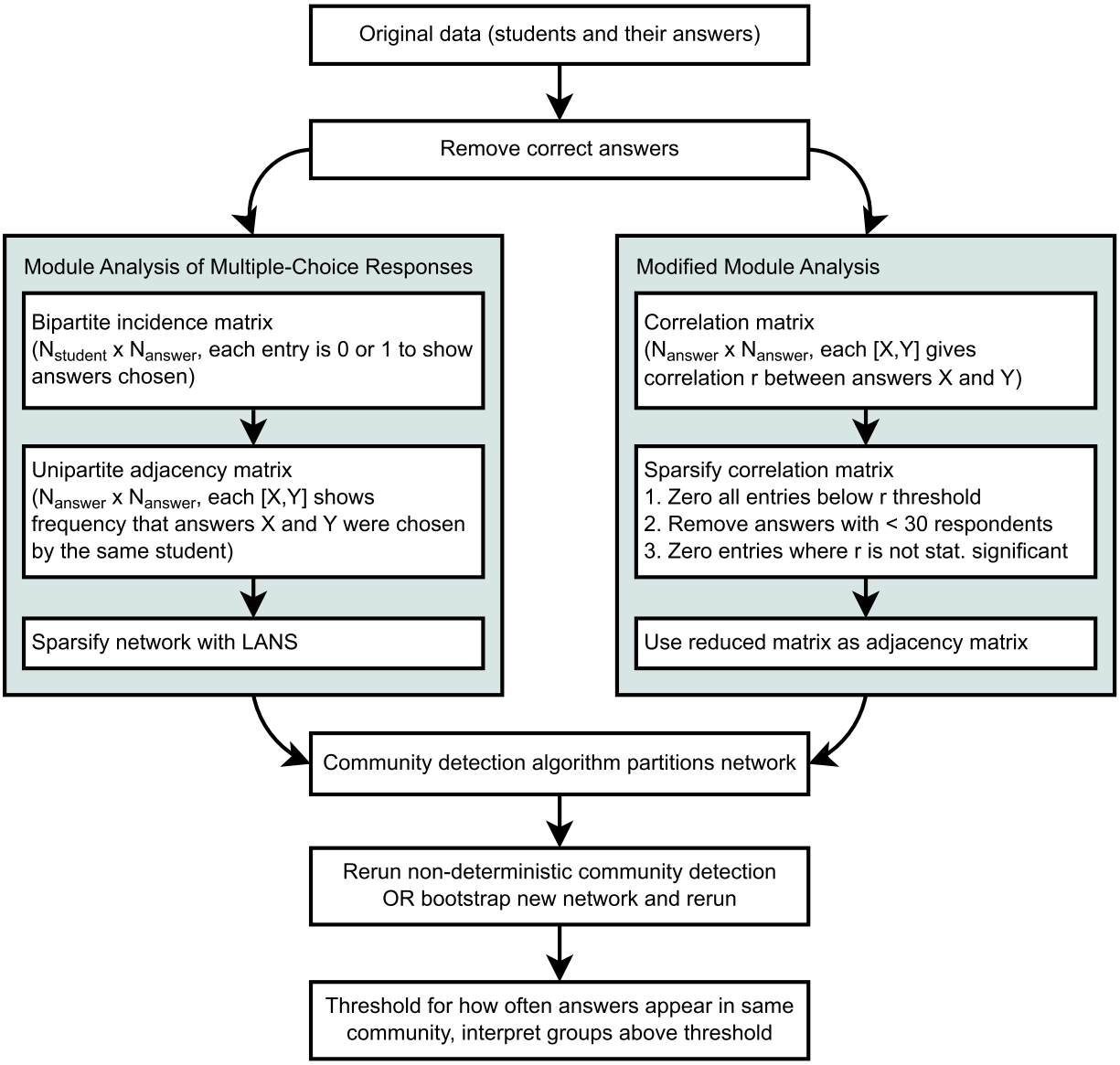

FIG. 1. Workflow of analysis for the original module analysis method (left branch) and our modified version (right branch).

includes two types of nodes where all edges connect nodes of different types. This network included nodes representing students and nodes representing FCI responses. The bipartite network is then projected onto a unipartite network containing only nodes representing FCI responses. Edges in this network connect different responses of the same student. Edge weights represent the number of students who selected the pair of responses connected by the edge. For example, if 40 students selected FCI responses 1A and 2B, where the number is the item number and the letter is the response within the item, there would be an edge between node 1A and node 2B with weight 40. While the bipartite network can be used to extract additional properties of the network [85], this was not done in study 1. As such, we began with the unipartite network. The unipartite network can be represented by a two-dimensional matrix, called the adjacency matrix, $\mathrm{adj}(X, Y)$, where X and Y are FCI item responses (for example, $X = 1\mathrm{A}$). The value $\mathrm{adj}(X, Y)$ is the number of students who selected response X and response Y. In the above example, $\mathrm{adj}(1A, 2B) = 40$. The network representing the post-test responses of women on the FCI post-test is shown in Fig. 2. Because of the differences identified between men and women in study 2 for this sample, all results are reported disaggregated by gender. The network in Fig. 2 is fairly representative of the pretest and post-test networks for men and women. Figure 2 uses a node placement algorithm that places more densely connected nodes close to one another. As in study 1, only incorrect responses were included in the network. The correct responses are highly correlated and are often the most commonly selected responses. If they are included in the network, they form a tightly connected community that prevents exploration of the incorrect answers. Figure 2 is presented as an example of a network based on the adjacency matrix; in what follows, we will propose a modification to this network converting it to a correlation network. It is the correlation network in Fig. 4 that is used in the primary analysis in this work. For interested readers, an enlarged version of the network in Fig. 2 is presented in the Supplemental Material [86].

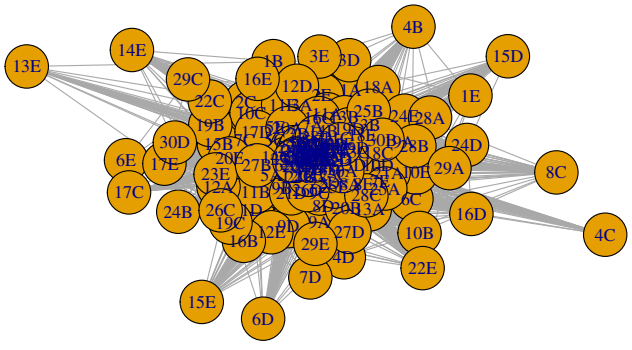

FIG. 2. Unipartite network for the FCI post-test responses of women.

To attempt to replicate the results of study 1, community detection algorithms were applied to the network shown in Fig. 2. First, a complete replication was attempted which employed the "Infomap" software [84] as was originally used in study 1. This software, designed for very large networks, presents such significant installation and use barriers that it seems unlikely that it will ever achieve broad acceptance in PER. A second path to replication using the infomap implementation in the igraph package [5] in R was also attempted.

To extract meaningful structure from a high-density network, the network must generally be simplified without removing important structure. The process of simplifying a network by removing edges is called "sparsification." The network sparsification method used in study 1 was locally adaptive network sparsification (LANS) [87]. The LANS algorithm removes edges based on the distribution of edge weights connected to each node. The probability of selecting an edge with a smaller weight at random is compared to a predetermined significance level and only edges above that level are retained. This method is locally adaptive because it depends only on the edges incident on a single node. A consequence of sparsifying based on the distribution of weights incident on each node is that no node will have its last edge removed, so no connected node is unconnected from the rest of the network. This ensures that local structures important to the global structure of the network are retained.

After sparsifying with LANS (using code from Ref. [88], Supplemental Material), the Infomap CDA was applied. Infomap is based on information theoretic methods. The algorithm records a random walk through the network by assigning codewords to each node, then trying to minimize the length of the description. Nodes visited more often are given shorter codes. Communities where the random walker tends to spend more time are given their own unique codes. The information needed to represent the network is reduced because the codes for individual nodes can be reused within different communities. This process results in communities of nodes that are connected more to each other than to nodes outside the community. Because Infomap is not deterministic, it was run 1000 times and the communities that were most often found were selected as the misconception modules in study 1.

Applying Infomap with LANS sparsification failed to identify meaningful community structure for the large dataset in the current study; Infomap consistently identified only one large community.

To explore the source of the discrepancy with study 1, an alternate implementation of Infomap was employed; this implementation was part of the igraph package in the R software system. A simpler sparsification algorithm was also employed. The LANS algorithm statistically evaluates each edge, but will not remove the last edge connecting a node. This algorithm is a reasonable choice for a network

where every edge is purposeful (such as air travel), but may amplify noise in a network of student responses where some edges are the result of careless mistakes or guessing. As such, the network was also sparsified by imposing a threshold requiring edges to have a minimum weight. Multiple thresholds were tried. The Infomap community detection algorithm used in study 1 identified only one community at all threshold values. Many other CDAs are available in the igraph package; some identified two communities even at very high thresholds. No CDA available in igraph identified more than 2 communities.

The Infomap CDA is based on a random walk algorithm; another class of CDAs is based on maximizing the modularity of the communities within a network. Modularity is a measure that compares, for a given division of a network into communities, how many more intracommunity links exist than expected by chance in an equivalent network [89]. Modularity values range from zero to one, where a network with a modularity of zero means there is no clustering in the network and a modularity of one is a strongly clustered network. The fast-greedy CDA is an implementation of a modularity-based CDA [90,91]. It works by suggesting a random division of the network into communities, then proceeds to move nodes, one at a time, to different communities, keeping any move that increases the modularity of the network. The algorithm is known as "fast greedy" because it prioritizes speed over finding the optimal solution. It is a greedy algorithm because it maximizes the modularity based on local changes, instead of considering the overall structure of the network.

Figure 3 shows the communities identified by the fast-greedy CDA at an edge weight threshold of $N/2$ where N is the number of participants; only 22 nodes remain connected at this threshold. Nodes in different communities are shown with different shading.

There seem to be two likely sources of the differences of the results of this study and study 1: sample size and the LANS algorithm. To investigate sample size, 100 subsamples of 143 students each were drawn from the sample in this study. Applying Infomap using R identified only one community 100% of the time with no sparsification and one community 92% of the time with the requirement that the edge weight be at least $N/10$ where N is the number of students.

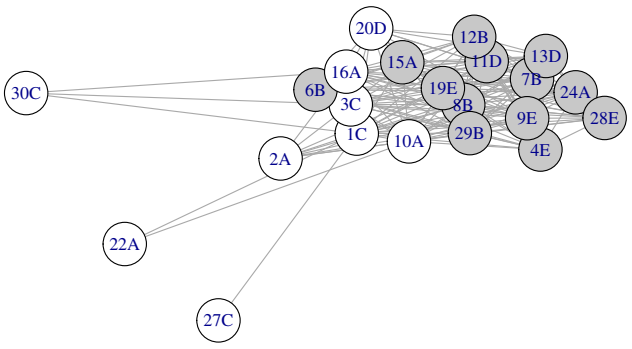

FIG. 3. Communities detected for the adjacency matrix of FCI post-test responses of women with an edge weight threshold of $N/2$ using the fast-greedy CDA.

The igraph package implements many CDA algorithms; for the small network analyzed in this work, most performed similarly. For rest of this work, the fast-greedy CDA described above will be used. Again, the data was subsampled to 143 students to compare with study 1. With no sparsification, the fast-greedy algorithm identified 3 to 6 communities with 3 to 4 communities identified in 92% of the runs. With the edge weight greater than $N/10$ sparsification, fast-greedy identified 2 to 4 communities with 66% of the runs identifying 3 communities. The communities identified made little theoretical sense within the framework of study 3 with very different items in the same communities. As such, while some of the differences in the studies may be attributed to sample size, the choice of CDA also influenced the communities identified at small sample size. At the large sample size of the current study, the various community detection algorithms implemented in igraph give fairly similar results.

B. Correlation analysis

Part of the cause of the failure of MAMCR to find meaningful community structure for large samples can be understood by comparing the adjacency matrix to the correlation matrix. The correlation matrix also defines a network, most usefully when a threshold value is applied. The adjacency matrix which produced the network in Fig. 2 has no obvious clustered structure. The partial correlation matrices reported in study 3 clearly show clustering into distinct communities.

The correlation between item X and item Y is defined as

$$\mathrm{corr}(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{1}$$

where μ_j is the mean of variable j, σ_j is the standard deviation, and $E[Z]$ is the expectation value of the random variable Z. The expectation value is defined as

$$E[X] = \sum_i \frac{X_i}{N}, \tag{2}$$

where i is a participant and N is the number of participants. Equation (1) can be simplified to produce

$$\mathrm{corr}(X,Y) = \frac{E[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y}. \tag{3}$$

For dichotomously scored items, the sum $\sum_i X_i Y_i$ is the X,Y entry in the adjacency matrix, $\mathrm{adj}(X,Y) = \sum_i X_i Y_i$. The correlation matrix is then related to the adjacency matrix by

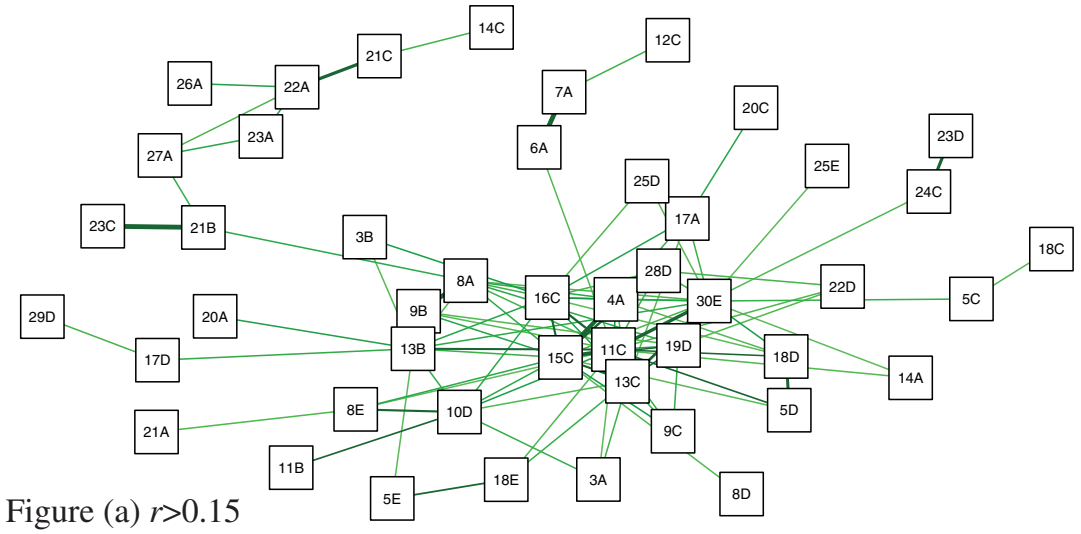

Figure (a) $r>0.15$

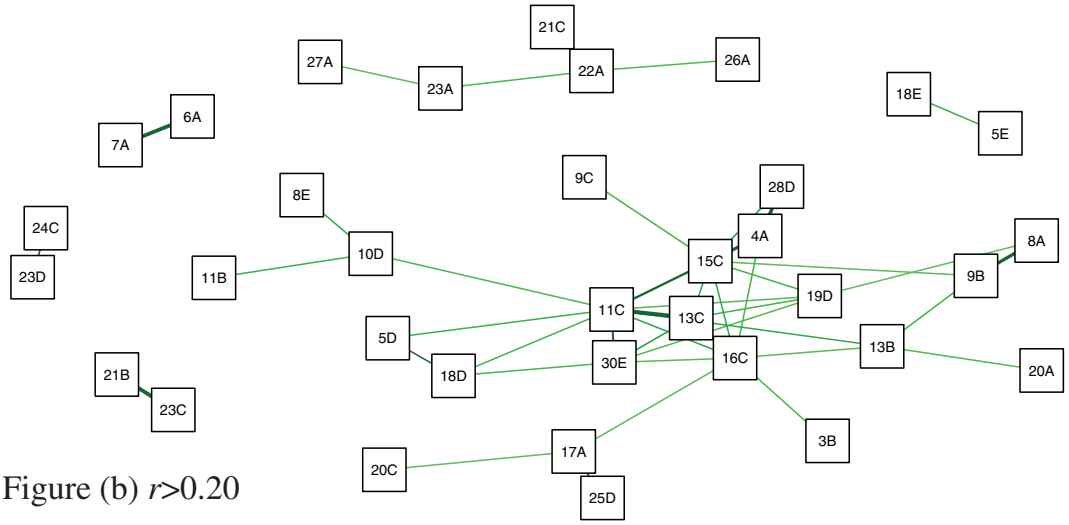

Figure (b) $r>0.20$

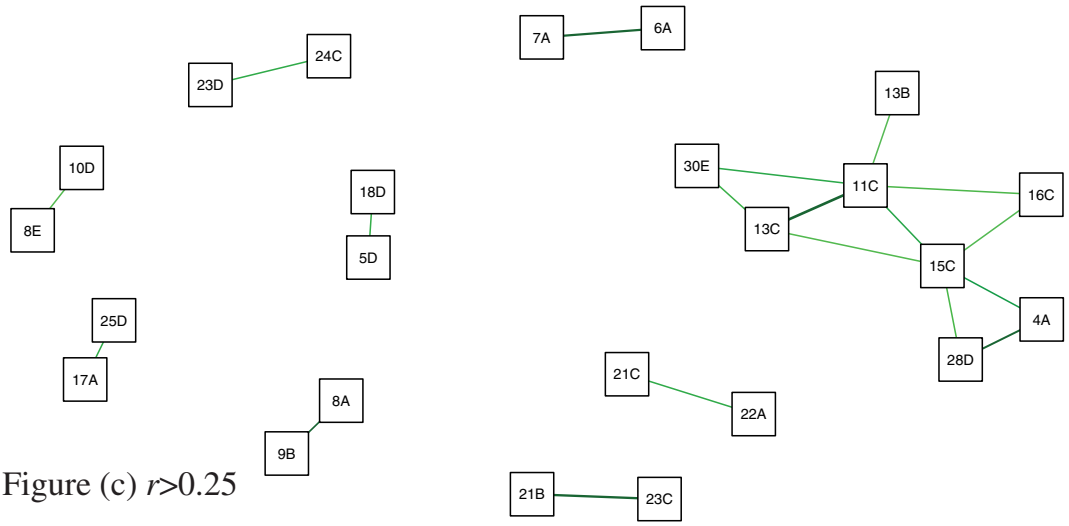

Figure (c) $r>0.25$

FIG. 4. Post-test correlation matrices of women at varying levels of r.

$$\mathrm{corr}(X,Y) = \frac{\mathrm{adj}(X,Y) - N\mu_X\mu_Y}{N\sigma_X\sigma_Y}. \quad (4)$$

A pair of items can have a large $\mathrm{adj}(X,Y)$ in a number of ways: (a) purposeful association, students preferentially select the two items together, or (b) accidental association, many students select both items so on average the items get selected together often. By subtracting the product of the means, the correlation matrix eliminates the second case and only has large values for purposefully selected pairs. This suggests the adjacency matrix contains many more edges that are the result of random chance than the correlation matrix. The correlation matrix also has the substantial advantage of the existence of significance tests for entries allowing the discarding of nonsignificant edges.

With this observation, we propose a modification of MAMCR, called modified module analysis, that investigates the community structure of the correlation matrix. The remainder of this work investigates this proposal. The differences between MAMCR and MMA are presented schematically in Fig. 1.

To explore this proposal, the correlation matrix was calculated for all incorrect answers. Nodes with too few participants to be statistically reliable were eliminated; for this work, nodes with fewer than 30 responses were removed. Edges were removed where the correlation, r, between the two nodes was not significant at the $p = 0.05$ level where a Bonferroni correction was applied to reduce the type I error rate. As with the adjacency matrix, a threshold was then applied to simplify the network. For this work, only positive correlations were considered; future work will investigate networks with positive and negative correlations. Figure 4 shows the correlation matrix for the post-test results of women retaining only entries with $r > 0.15$, $r > 0.20$, and $r > 0.25$. The representation in Fig. 4 was produced by the qgraph package in R [92]. The width of the line is proportional to the size of the correlation. Node placement is for visual effect only.

C. Modified module analysis

The correlation matrices in Fig. 4 show a clear clustered structure. MMA was applied to understand these structures. The communities detected for the $r > 0.20$ correlation matrix are shown in Fig. 5. Figure 5 shows the communities identified by a single application of the fast-greedy CDA. To understand the stability of these structures, the algorithm was applied multiple times. Because some communities were only identified in some applications of the CDA, the communities identified by the multiple applications of the algorithm presented later in the paper do not fully align with those in Fig. 5.

Both the CDA and the sample itself contain randomness, and therefore, some of the community structure in Fig. 5 may result from chance. To determine the part of the community structure not resulting from random

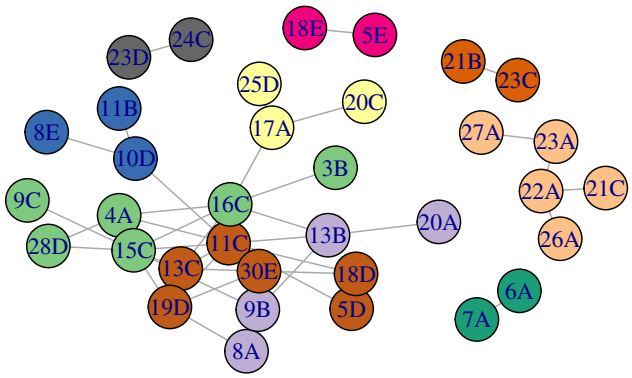

FIG. 5. Communities detected in the FCI correlation matrix with $r > 0.2$. Each community is drawn in a different color.

fluctuations, bootstrapping with 1000 replications was performed. Bootstrapping is a statistical technique that forms a distribution of a statistic of interest by subsampling the dataset with replacement [93]. The R "boot" package was used to perform the bootstrapping [94]. Part of the goal of this work was to compare the community structure of men and women. As one part of the sparsification process, infrequently selected nodes and insignificantly weighted edges were removed. Both the threshold for an infrequently selected node and edge significance depend on sample size. For a fair comparison, a sample balanced between men and women was required. The full dataset was very unbalanced. To correct for this, when the male sample was bootstrapped, 1000 samples were drawn, each of the same size as the overall female sample. For the female sample, bootstrapping was performed by subsampling with replacement which preserved the overall size of the sample.

The number of times each pair of responses was found in the same community was recorded for each of the 1000 samples forming a "community matrix." The community matrix was nearly completely disconnected into small clusters. The clusters are shown in Table I. Responses were identified in the same communities at different rates. We define the community fraction C as the fraction of the bootstrap subsamples in which the pair of items were found in the same community. The community matrix was filtered to show items that were identified in $C > 60\%$ and $C > 80\%$ of the communities in the 1000 bootstrap replications in Table I. The majority of the communities extracted from the community matrix were fully connected; each node was connected to every other node in the community. Some, however, were not. The intracommunity density γ is defined as the ratio of the number of edges in the community to the maximum number of edges possible [69]. For communities with $\gamma < 1$, γ is presented as a percentage in parenthesis in Table I. For example, if a community contains four nodes then there are a maximum of six distinct edges between the nodes. If the community only possesses five of those edges, then $\gamma = 5/6$.

TABLE I. Communities identified in the pretest and post-test incorrect answers at $r > 0.2$ and differing levels of the community fraction C. The number in parenthesis is the intracommunity density γ for communities where the intracommunity density is not one.

	Pretest		Post-test	
Community	Men	Women	Men	Women
$C > 60\%$				
1A, 2C, 15B, 19B	X(67%)			
1D, 2D	X	X		
3B, 13B			X	
4A, 15C, 28D		X	X	X
4A, 15C, 16C, 28D	X			
5D, 11C, 13C, 18D, 30E			X	
5D, 11C, 13C, 18D, 19D, 30E				X(73%)
5E, 18E	X	X	X	X
6A, 7A	X	X	X	X
8A, 9B	X	X	X	X
8E, 10D				X
11B, 29A	X	X		
15D, 16D	X	X		
17A, 25D			X	X
21B, 23C			X	X
21C, 22A, 23A, 26A				X(83%)
21C, 22A	X		X	
23D, 24C	X	X	X	X
$C > 80\%$				
1A, 2C	X			
1D, 2D	X			
4A, 15C, 28D	X	X	X	X
5D, 11C, 13C, 18D, 30E				X(60%)
5D, 18D			X	
5E, 18E	X	X	X	X
6A, 7A	X	X	X	X
8A, 9B	X	X	X	X
11B, 29A	X	X		
11C, 13C, 30E			X	
17A, 25D			X	X
21B, 23C			X	X
21C, 22A	X		X	X
23D, 24C	X	X	X	X

D. The structure of incorrect FCI responses

Unless otherwise stated, results below are reported for $C > 0.8$ and $r > 0.2$.

1. Types of incorrect communities

Table II classifies the incorrect reasoning for each community of incorrect answers in Table I. These can be divided into two general classes: communities resulting from blocking and communities resulting for consistently applied incorrect reasoning (misconceptions). Communities $\{8A, 9B\}$, $\{21B, 23C\}$, and $\{21C, 22A\}$ are answers within blocked problems where the second answer in the pair would be correct if the first answer was correct. The other communities apply either the same incorrect reasoning or related incorrect reasoning.

Hestenes and Jackson produced a detailed taxonomy of the naive conceptions (their terminology) tested by the FCI [40]. Table II shows a mapping of this taxonomy onto the incorrect answer communities identified in the current work. The taxonomy divides the naive conceptions into a general category and a number of subcategories. The number in parenthesis in Table II is the subcategory label [40]. Items marked with an asterisk in Table II are part of item blocks. Because the relation between the items seems to be largely generated by the interdependencies resulting from blocking rather than consistently applied misconceptions, the blocked items will not be discussed further.

Some issues arise in comparing the proposed FCI taxonomy with communities identified by MMA and the similar item blocks identified in study 3. First, for some of items in the incorrect communities, no misconception was identified (items 1D, 2C, and 2D). Students are answering these items in a correlated manner which implies the possibility of consistent reasoning patterns; for these items a possible misconception was suggested. The new misconception was labeled "(Add)." Table II shows that often the items in the incorrect communities identified by MMA belong to multiple naive misconception categories and have different subcategories. This would seem to imply that the naive conception taxonomy is more detailed than the actual application of misconceptions by students as measured by the FCI. For example, in the Newton's 3rd law community, $\{4A, 15C, 28D\}$, different items involve objects of different activity from one student pushing on another student (item 28, one active object), to a car pushing a truck (item 15, one active object), to a head-on collision (item 4, two active objects). Further, the three items involve objects of different mass with the more active object having less mass in item 15 and more mass in item 28. It is unclear why these items involving multiple different misconceptions are answered consistently incorrectly. This may result from all of the responses representing a failure to understand Newton's 3rd law or from students reasoning in a manner inconsistent with the misconception view.

Table II includes a column which proposes a title for the dominant misconception. In many cases, the dominant misconception was identified as the misconception shared by the majority of the items. In some cases, a dominant misconception was proposed. For the Newton's 3rd law community, $\{4A, 15C, 28D\}$, multiple misconceptions were shared equally and no dominant misconception was identified. In the following, the combination of the greater mass implies greater force and most active agent produces greater force misconceptions are called "Newton's 3rd law misconceptions."

TABLE II. Misconceptions represented by incorrect answer communities. Communities marked with a ∗ result from blocked problems. Proposed additions are marked (Add). Proposed items to be removed are marked (Remove). If Add or Remove is placed before all items, it applies to all items. If Add or Remove is placed before only one of many items, it applies to that item.

	Naive conceptions		
Community	Category	Subcatagory	Dominant misconception
1A, 2C	Gravity	1A (G3): Heavier objects fall faster (Add) 2C: Heavier objects travel farther	Heavier objects fall faster
1D, 2D	Unknown	(Add) 1D: Lighter objects fall faster (Add) 2D: Lighter objects travel farther	Lighter objects fall faster
4A, 15C, 28D	Action-reaction pairs	4A, 28D (AR1): Greater mass implies greater force 15C, 28D (AR2): Most active agent produces greatest force	Greater mass implies greater force Most active agent produces greatest force
5D, 18D	Impetus Active forces	(Remove) 5D, 18D (I5): Circular impetus 5D (I1): Impetus supplied by "hit" 5D, (Add) 18D (AF2): Motion implies active forces	Motion implies active forces
5E, 18E	Impetus Active forces	5E (I1): Impetus supplied by "hit" (Remove) 5E (I5): Circular impetus 5E (I1): Motion implies active forces 5E, 18E (CF): Centrifugal force	Motion implies active forces Centrifugal force
6A,7A	Impetus	6A, 7A (I5): Circular impetus	Circular impetus
8A, 9B*	Concatenation of influences	8A, 9B (CI3): Last force to act determines motion	
11B, 29A	Other influences on motion Impetus	29A (Ob): Obstacles exert no force (Remove) 11B (I1): Impetus supplied by "hit" (Add) 11B (AF2): Motion implies active forces	Motion implies active forces
11C, 13C, 30E	Impetus	11C, 30E (I1): Impetus supplied by "hit" (Add) 11C, 13C (AF2): Motion implies active force 13C (I3): Impetus dissipation	Motion implies active forces
17A, 25D	Concatenation of influences Resistance	17A, (Add) 25D (CI1): Largest force determines motion 25D (R2): Motion when force overcomes resistance	Largest force determines motion
21B, 23C*	Concatenation of influences	21B, 23C (CI3): Last force to act determines motion	
21C, 22A*	Concatenation of influences Active forces	21C (CI2): Force compromise determines motion 22A (AF4): Velocity proportional to applied force	
23D, 24C*	Impetus	23D, 24C (I3): Impetus dissipation 23D (I2): Loss or recovery of original impetus	Impetus dissipation

The communities $\{1A, 2C\}$ and $\{1D, 2D\}$ are difficult to resolve with the "gravity" misconception described in Sec. I C where heavier objects tend to move nearer the center of Earth. Response 1A is a clear application of this misconception, while item 1D applies the opposite of the misconception where lighter objects tend to move closer to the center of Earth. This misconception has been added and the category labeled as "Unknown." More research would be needed to determine if the students were applying a misconception about gravity or a more general misconception that lighter objects travel faster. Neither responses 2C or 2D fit within the gravity misconception; both reason that the object that falls faster travels farthest. This misconception has been added for both objects, but it is equally possible the students are applying less coherent reasoning not well represented by the misconception view. The students could be applying the *p* prim "larger implies larger" to the result that the object travels faster. Again, more research is needed to resolve the ambiguity.

Multiple items were identified as involving the misconception of circular impetus. Circular impetus is used in two alternate ways. In responses 6A and 7A, circular impetus involves an object continuing to move in a circle after a constraint is removed. Item 5 represents a ball shot into a circular channel and item 18 represents a boy swinging on a rope. In responses 5D, 5E, 18D, and 18E, the constraint is still in place. Both items 5 and 18 also include a force of the channel or rope in the list of forces; these forces are unnecessary if the object moves in a circle of its own accord. As such, we propose removing the circular impetus misconceptions from 5D, 18D, 5E, and 18E; these items have been labeled (Remove) in Table II. This suggestion is supported by the failure to find $\{5D, 5E, 6A, 7A, 18D, 18E\}$ as a single community.

The coding of the misconceptions represented by the 5D and 18D community seems problematic. Answer D on both items includes a force in the direction of motion, and therefore, it is unclear why item 18 is not included in the motion implies active forces misconception. This has been added to Table II. This is supported by the identification of the $\{5D, 18D\}$ community. Item 18 also does not provide a response that includes both the centrifugal force and the motion implies active forces misconceptions; as such, students may still be applying the motion implies active force misconception, it is just not tested by item 18E.

Items 17 and 25 also require some additional analysis. Both items involve objects moving at a constant speed under the influence of multiple forces. In both 17A and 25D, the greater force is in the direction of motion. It seems that 25D should also test the largest force determines motion misconception. This has been added to Table II and is supported by the identification of the $\{17A, 25D\}$ community.

The pretest community $\{11B, 29A\}$ is also curious. In item 11, a hockey puck is struck activating the impetus supplied by the "hit" misconception, but response 11B explicitly asks about a force in the direction of motion. As such, we propose that this item also tests the motion implies active forces misconception. Item 29 involves a chair sitting on a floor; response 29A identifies only the force of gravity on the object and ignores the normal force. It seems difficult to claim this community probes a common misconception. Item 29 was also demonstrated to have poor psychometric properties in study 2; the correlation between 11B and 29A may have resulted from 29A not functioning as intended.

The community $\{11C, 13C, 30E\}$ continues to convolve the motion involves active forces misconception with the impetus supplied by the hit misconception. Response 30E explicitly discusses the force of the hit while items 11C and 13C discuss a force in the direction of motion. We propose that items 11C and 13C also test this misconception. This addition is supported by the identification of $\{11C, 13C, 30E\}$ as a community. Further, only item 13C involves the idea of a dissipation of impetus. For this community, while multiple misconceptions are tested, one seems to dominate student responses, motion implies active forces.

Finally, the blocked item responses 23D, 24C differ from the other blocked responses. Rather than the second response being the correct answer if the first response was correct, both appear to be applications of the dissipation of impetus misconception.

2. Reducing sparsification

The $r > 0.2$ and $C > 0.8$ thresholds generated a fairly disconnected network. This network was productive in identifying incorrect answers that were frequently selected at the same time by the same student. As these thresholds are relaxed, the network becomes more connected as shown in Fig. 4. As the network becomes more connected, related misconceptions may merge showing the students have a coherent non-Newtonian force concept. The communities identified at $r > 0.15$ and for $C > 0.6$ and $C > 0.8$ are shown in the Supplemental Material [86]. While relaxing the thresholds did allow the community $\{5D, 11C, 13C, 18D, 30E\}$ to be detected for both men and women, most other new communities identified did not result from the merger of communities identified at more restrictive thresholds. Particularly on the pretest, the larger communities do not make much sense in terms of the framework of study 3. This is particularly evident in the mixing of the Newton's 3rd law items $\{4, 15, 16, \text{and } 28\}$ with other items. As such, it appears that student misconceptions exist relatively independently as small groups of consistent answers, not as a part of a larger coherent framework.

3. The strength of common misconceptions

One motivation of study 1 was to provide instructors with a mechanism for identifying common misconceptions so that specific interventions could be targeted to address those misconceptions. The communities of incorrect answers remaining on the post-test as shown in Table I could be used to provide a measure of the prevalence of the misconception in the classes studied. Table III presents an overall average for each incorrect community in Table I on

TABLE III. Percentage of students selecting each incorrect community for the FCI post-test ($C > 80\%$). A t test was performed to determine if the differences between men and women were significant, the p value is presented. Cohen's d for the difference is also presented.

	Male	Female			
Community	Ave. (%)	Ave. (%)	p	d	Misconception
4A, 15C, 28D	32 ± 47	33 ± 47	0.27	0.02	Newton's 3rd law misconceptions
5D, 11C, 13C, 18D, 30E	22 ± 42	20 ± 40	<0.001	0.06	Motion implies active forces
5E, 18E	7 ± 25	7 ± 25	0.69	0.01	Motion implies active forces, centrifugal force
6A, 7A	14 ± 35	5 ± 21	<0.001	0.39	Circular impetus
17A, 25D	42 ± 49	37 ± 48	<0.001	0.11	Largest force determines motion

the post-test. Only communities that did not result from problem blocking are presented. Averages were calculated by assigning a score of 1 if the response was selected and 0 if it was not, then averaging over each item in the group. Results are disaggregated by gender and the p value for a t test to determine if differences by gender are significant is also presented; Cohen's d provides a measure of effect size. Cohen suggests $d = 0.2$ as a small effect, $d = 0.5$ as a medium effect, and $d = 0.8$ as a large effect [95].

The overall difference in post-test percentage score between men and women was 8%. The percentage of students who answer an item correctly directly influences the percentage of students who answer an item incorrectly; therefore, only differences in Table III greater than 8% represent unexpected differences between men and women. Only items $\{6A, 7A\}$ exceed this difference, but then only slightly with a difference of 9%. Items $\{6A, 7A\}$ are also the only community with differences of at least a small effect size; however, the effect size is likely inflated by the small standard deviation of women because of a floor effect. In general, the rate of selecting one of the communities of common incorrect answers was very similar for men and women.

For the class studied, the results of Table III suggest that additional effort be directed to addressing the largest force determines motion misconception measured by $\{17A, 25D\}$ and Newton's 3rd law misconceptions measured by $\{4A, 15C, 28D\}$.

IV. DISCUSSION

A. Research questions

This study sought to answer four research questions; they will be addressed in the order proposed.

RQ1: Are the results of module analysis for multiple-choice responses replicable for large FCI datasets? If not, what changes to the algorithm are required to detect meaningful communities of incorrect answers? The MAMCR process described in study 1 identified only one or two communities in our data whether using LANS or an edge weight threshold to sparsify the network. This result held for Infomap and for other CDAs. Reducing the data to a comparable size by subsampling generated more communities, but still fewer than identified in study 1; however, the communities identified did not make conceptual sense. We concluded that the community structure identified in study 1 was the result of the low sample size and the LANS algorithm and that modifications to MAMCR were needed to productively identify incorrect answer communities.

The failure of MAMCR for large samples led us to propose a variant of the algorithm using the correlation matrix instead of the adjacency matrix to build the network. This matrix was sparsified by removing statistically insignificant correlations and correlations below a threshold ($r < 0.2$ for most of our analyses). Using the fast-greedy CDA on this new network produced a rich set of incorrect communities (Table I). These communities were often related to items with related correct answers as identified by study 3. As such, the communities represent consistent application of incorrect reasoning to contextually different items that would be viewed as isomorphic by experts meeting our definition of a misconception. These communities fall into two broad categories: those employing similar incorrect reasoning and those resulting from "blocked" items where an incorrect choice later in the block is the correct answer given an incorrect choice earlier in the block.

RQ2: How do the communities detected change as network-building parameters are modified? Do these changes support the existence of a coherent non-Newtonian conceptual model? A more permissive threshold for the correlation matrix ($r > 0.15$) yielded larger communities as shown in the Supplemental Material [86]. These larger communities were not formed by the joining of smaller communities related to the same misconception; in fact, many of the communities contained items that had little conceptual relation. As such, it appears that the best model of student misconceptions are as isolated pieces of reasoning associated with items with a similar correct solution structure.

RQ3: How is the incorrect answer community structure different between the pretest and the post-test? For $C > 0.8$ and $r > 0.2$, a total of 14 incorrect answer communities were identified for either men or women pre- and post-instruction; 5 of the communities were consistently identified for both genders pre- and postinstruction. Three of these five represent consistently applied misconceptions: $\{4A, 15C, 28D\}$, Newton's 3rd law misconceptions; $\{5E, 18E\}$, motion implies active forces and the existence of a centrifugal force; and $\{6A, 7A\}$, circular impetus. The items from which the incorrect answers in these communities were drawn were all identified as having very similar correct solution structure in study 3. The other two communities were drawn from problem blocks: $\{8A, 9B\}$ and $\{23D, 24C\}$. Three incorrect communities disappeared with instruction: for all students $\{11B, 29A\}$, motion implies active forces; for men only, $\{1A, 2C\}$ heavier objects fall faster and $\{1D, 2C\}$ lighter objects fall faster. Many incorrect communities were only identified post-instruction including $\{17A, 25D\}$ involving items with similar solution structure as identified in study 3.

RQ4: How is the incorrect answer community structure different for men and women? Do the differences explain the gender unfairness identified in the instrument? Postinstruction, using $C > 0.8$ and $r > 0.2$, 11 communities were identified for either men or women; 8 of these communities were identified for both men and women. One of the other three communities was only identified for women, $\{5D, 11C, 13C, 18D, 30E\}$, and represents the motion implies active forces misconception. This

community was the merger of the two communities only identified for men {5D, and 18D} and {11C, 13C, 30E}; the female community was also not completely connected, $\gamma = 0.6$. As such, women may have a slightly more integrated motion implies active forces misconception postinstruction, but, in general, the misconception structure of men and women is strikingly similar postinstruction. The differences between men and women postinstruction did not involve the unfair items identified in study 2 and cannot explain the unfairness of these items.

Pre-instruction, nine communities were identified for either men or women; six were identified for both men and women. All of the communities not shared by men and women were only identified for men. One of these communities, {21C, 22A}, was the result of blocking and was identified for both men and women postinstruction. The other two communities unique to men, {1A, 2C} and {1D, 2D}, involve the heavier objects fall faster and the lighter objects fall faster misconceptions. The misconception structure of men and women was quite similar pre-instruction, with men holding more consistent misconceptions.

B. Additional observations

The misconception communities identified by MMA were not completely consistent with the naive conception taxonomy provided by Hestenes and Jackson for the FCI [40]. Often multiple naive conceptions were associated with the same community. This may indicate that student reasoning is better modeled by a more general framework such as knowledge-in-pieces or ontological categories. It may also indicate that the FCI cannot fully resolve the detailed set of misconceptions identified in the taxonomy.

The results of this work were not consistent with recent exploratory analyses of the FCI [41–43] which identified a few large factors; these factors mixed very different correct and incorrect responses. The small communities identified in the current work, which are partially supported by the taxonomy of Hestenes and Jackson, seem to indicate the MMA may be a more productive quantitative method to explore misconceptions.

V. IMPLICATIONS

Not all of the communities identified in Table I represent misconceptions. Some represent combinations of dependent answers. For these combinations, the second answer is correct if the first answer were the correct answer. This suggests that, because of the blocking of items in the FCI, a simple scoring of the instrument with each item as correct or incorrect may understate a student's knowledge of the material. Previous authors have called for reevaluating the scoring of the FCI [96–99], but not because of problem blocking.

The identification of three communities of incorrect answers that were the result of item blocking further supports the conclusions of study 3 that item blocking should be discontinued in future PER instruments because it may make the instruments difficult to interpret statistically.

The misconception communities identified in Table II allow instructors to determine the strength of students' misconceptions as they enter a physics class and the remaining strength after instruction, as shown in Table III. While it is unlikely many instructors will replicate the MMA analysis for their class, the misconceptions identified in Table II are generally consistent with both the taxonomy of Hestenes and Jackson [40] and the theoretical model of study 3. As such, it is likely these misconception communities are present in the thinking of students in many classes. An instructor not wishing to apply MMA could use the communities identified in this work and the scoring method used to produce Table III to measure the strength of these misconceptions in their students as a useful approximation. This should allow instructors to adjust their classes to address misconceptions remaining after instruction and to direct fewer resources to addressing misconceptions that are not present pre-instruction.

VI. FUTURE WORK

MMA was productive in extending the understanding of the incorrect answer structure of the FCI; it will be extended to other conceptual instruments including the Force and Motion Conceptual Inventory [13] and the Conceptual Survey of Electricity and Magnetism [100].

This work showed that a number of incorrect communities were only identified postinstruction. The reason for this is unclear and additional research is needed to understand this effect.

Network analysis encompasses a broad collection of powerful analysis techniques. The analysis in this work represents the barest beginnings of the possibilities of these techniques. Future research may consider networks with multiple types of nodes (possibly correct and incorrect answers or pretest and post-test answers) or multiple types of edges (possibly negative and positive correlations).

VII. CONCLUSION

Previous results reported for module analysis for multiple-choice responses could not be replicated for a large sample. The failure of the algorithm at large sample sizes likely results from a combination of unpurposeful edges in the adjacency matrix at large sample sizes and properties of the LANS sparsification algorithm. A modification of the algorithm, modified module analysis (MMA), based on the correlation matrix was productive in identifying useful community structure. MMA identified 11 communities on the post-test and 9 on the pretest. Most of these

communities were identified both for men and women: 8 on the post-test, 6 on the pretest. In general, the incorrect answer community structure identified for men and women was very similar and could not explain the gender differences previously identified in a subset of items in the instrument. The communities identified at high sparsification failed to merge into larger communities addressing similar misconceptions as sparsification was reduced. This suggests that students do not have an integrated non-Newtonian conceptual framework, but rather isolated incorrect beliefs strongly tied to the type of question asked.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, PHY-0108787.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[2] A. Madsen, S. B. McKagan, and E. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. Phys. Educ. Res. **9**, 020121 (2013).
[3] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).
[4] E. Brewe, J. Bruun, and I. G. Bearden, Using module analysis for multiple choice responses: A new method applied to Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **12**, 020131 (2016).
[5] G. Csardi and T. Nepusz, The igraph software package for complex network research, InterJournal, Complex Systems **1695**, 1 (2006).
[6] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2017).
[7] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure? Phys. Teach. **33**, 138 (1995).
[8] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. Phys. Educ. Res. **8**, 020105 (2012).
[9] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).
[10] T. F. Scott and D. Schumayer, Students' proficiency scores within multitrait item response theory, Phys. Rev. Phys. Educ. Res. **11**, 020134 (2015).
[11] 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).
[12] 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).
[13] 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).
[14] C. Nord, S. Roey, S. Perkins, M. Lyons, N. Lemanski, J. Schuknecht, and J. Brown, *American High School Graduates: Results of the 2009 NAEP High School Transcript Study* (U.S. Department of Education, National Center for Education Statistics, Washington, DC 2011).
[15] B. C. Cunningham, K. M. Hoyer, and D. Sparks, *Gender Differences in Science, Technology, Engineering, and Mathematics (STEM) Interest, Credits Earned, and NAEP Performance in the 12th Grade* (National Center for Education Statistics, Washington, DC, 2015).
[16] B. C. Cunningham, K. M. Hoyer, and D. Sparks, *The Condition of STEM 2016*, (ACT Inc., Iowa City, IA, 2016).
[17] P. M. Sadler and R. H. Tai, Success in introductory college physics: The role of high school preparation, Sci. Educ. **85**, 111 (2001).
[18] Z. Hazari, R. H. Tai, and P. M. Sadler, Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors, Sci. Educ. **91**, 847 (2007).
[19] Y. Maeda and S. Y. Yoon, A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT: R), Educ. Psychol. Rev. **25**, 69 (2013).
[20] D. F. Halpern, *Sex Differences in Cognitive Abilities*, 4th ed. (Psychology Press, Francis & Tayler Group, New York, NY, 2012).
[21] J. S. Hyde and M. C. Linn, Gender differences in verbal ability: A meta-analysis., Psychol. Bull. **104**, 53 (1988).
[22] J. S. Hyde, E. Fennema, and S. J. Lamon, Gender differences in mathematics performance: A meta-analysis., Psychol. Bull. **107**, 139 (1990).
[23] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, Cross-national patterns of gender differences in mathematics: A meta-analysis., Psychol. Bull. **136**, 103 (2010).
[24] X. Ma, A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics, J. Res. Math. Educ. **30**, 520 (1999).
[25] J. V. Mallow and S. L. Greenburg, Science anxiety: Causes and remedies, J. Coll. Sci. Teach. **11**, 356 (1982).
[26] M. K. Udo, G. P. Ramsey, and J. V. Mallow, Science anxiety and gender in students taking general education science courses, J. Sci. Educ. Technol. **13**, 435 (2004).

[27] J. Mallow, H. Kastrup, F. B. Bryant, N. Hislop, R. Shefner, and M. Udo, Science anxiety, science attitudes, and gender: Interviews from a binational study, J. Sci. Educ. Technol. **19**, 356 (2010).
[28] J. R. Shapiro and A. M. Williams, The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields, Sex Roles **66**, 175 (2012).
[29] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).
[30] L. McCullough and D. E. Meltzer, Differences in male/female response patterns on alternative-format versions of FCI items, in *Proceedings of the 2001 Physics Education Research Conference*, edited by K. Cummings, S. Franklin, and J. Marx (AIP, New York, 2001), pp. 103–106.
[31] S. Osborn Popp, D. Meltzer, and M. C. Megowan-Romanowicz, Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, in *2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).
[32] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Proc. **1413**, 171 (2012).
[33] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **14**, 020103 (2018).
[34] J. Clement, Students' preconceptions in introductory mechanics, Am. J. Phys. **50**, 66 (1982).
[35] J. Clement, D. E. Brown, and A. Zietsman, Not all preconceptions are misconceptions: Finding anchoring conceptions for grounding instruction on students intuitions, Int. J. Sci. Educ. **11**, 554 (1989).
[36] J. Clement, Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics, J. Res. Sci. Teach. **30**, 1241 (1993).
[37] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).
[38] I. A. Halloun and D. Hestenes, Common sense concepts about motion, Am. J. Phys. **53**, 1056 (1985).
[39] 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).
[40] Table II for the Force Concept Inventory (revised from 081695r), http://modeling.asu.edu/R&E/FCI-RevisedTable-II_2010.pdf. Accessed 3/17/2019.
[41] 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).
[42] 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).
[43] T. F. Scott and D. Schumayer, Central distractors in Force Concept Inventory data, Phys. Rev. Phys. Educ. Res. **14**, 010106 (2018).
[44] L. Viennot, Spontaneous reasoning in elementary dynamics, Eur. J. Sci. Educ. **1**, 205 (1979).
[45] D. E. Trowbridge and L. C. McDermott, Investigation of student understanding of the concept of acceleration in one dimension, Am. J. Phys. **49**, 242 (1981).
[46] A. Caramazza, M. McCloskey, and B. Green, Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects, Cognit. **9**, 117 (1981).
[47] P. C. Peters, Even honors students have conceptual difficulties with physics, Am. J. Phys. **50**, 501 (1982).
[48] M. McCloskey, Intuitive physics, Sci. Am. **248**, 122 (1983).
[49] R. F. Gunstone, Student understanding in mechanics: A large population survey, Am. J. Phys. **55**, 691 (1987).
[50] C. W. Camp and J. J. Clement, *Preconceptions in Mechanics: Lessons dealing with Students' Conceptual Difficulties* (Kendall/Hunt, Dubuque, IA, 1994).
[51] L. C. McDermott, Students' conceptions and problem solving in mechanics, in *Connecting Research in Physics Education with Teacher Education*, edited by A. Tiberghien, E. Leonard Jossem, and J. Barojas (International Commission on Physics Education, Singapore, 1997), pp. 42–47.
[52] R. Rosenblatt and A. F. Heckler, Systematic study of student understanding of the relationships between the directions of force, velocity, and acceleration in one dimension, Phys. Rev. Phys. Educ. Res. **7**, 020112 (2011).
[53] N. Erceg and I. Aviani, Students' understanding of velocity-time graphs and the sources of conceptual difficulties, Croat. J. Educ. **16**, 43 (2014).
[54] B. Waldrip, Impact of a representational approach on students' reasoning and conceptual understanding in learning mechanics, Int. J. Sci. Math. Educ. **12**, 741 (2014).
[55] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[56] A. A. diSessa and B. L. Sherin, What changes in conceptual change? Int. J. Sci. Educ. **20**, 1155 (1998).
[57] M. T. H. Chi and J. D. Slotta, The ontological coherence of intuitive physics, Cognit. Instr. **10**, 249 (1993).
[58] M. T. H. Chi, J. D. Slotta, and N. De Leeuw, From things to processes: A theory of conceptual change for learning science concepts, Learn. Instr. **4**, 27 (1994).
[59] J. D. Slotta, M. T. H. Chi, and E. Joram, Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change, Cognit. Instr. **13**, 373 (1995).
[60] D. Hammer, Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions, J. Learn. Sci. **5**, 97 (1996).
[61] D. Hammer, More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research, Am. J. Phys. **64**, 1316 (1996).
[62] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).
[63] 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, Germany, 1992), pp. 110–128.
[64] R. E. Scherr, Modeling student thinking: An example from special relativity, Am. J. Phys. **75**, 272 (2007).
[65] R. Duit and D. F. Treagust, Conceptual change: A powerful framework for improving science teaching and learning, Int. J. Sci. Educ. **25**, 671 (2003).
[66] E. Etkina, J. Mestre, and A. O'Donnell, The impact of the cognitive revolution on science learning and teaching, in *The Cognitive Revolution in Educational Psychology*, edited by J. M. Royer (IAP, Charlotte, NC, 2005), pp. 119–164.
[67] J. D. Bransford, A. L. Brown, and R. R. Cocking, *How People Learn: Brain, Mind, Experience, and School* (National Academy Press, Washington, DC, 2000).
[68] M. J. Newman, *Networks*, 2nd ed. (Oxford University Press, New York, NY, 2018).
[69] K. A. Zweig, *Network Analysis Literacy: A Practical Approach to the Analysis of Networks* (Springer-Verlag, Wien, Austria, 2016).
[70] A. V. Papachristos and C. Wildeman, Network exposure and homicide victimization in an African American community, Am. J. Public Health **104**, 143 (2014).
[71] F. De Vico, J. Richiardi, M. Chavez, and S. Achard, Graph analysis of functional brain networks: Practical issues in translational neuroscience, Phil. Trans. R. Soc. B **369**, 1 (2014).
[72] J. Lopéz Peña and H. Touchette, A network theory analysis of football strategies, in *Sports Physics: Proc. 2012 Euromech Physics of Sports Conference*, edited by C. Clanet (Éditions de l'École Polytechnique, Paris, France, 2012), pp. 517–528.
[73] Z. Zheng and Y. Zhao, Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to "CandidatusLiberibacter asiaticus" infection, BMC Genomics **14**, 27 (2013).
[74] S. Fortunato and D. Hric, Community detection in networks: A user guide, Phys. Rep. **659**, 1 (2016).
[75] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).
[76] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, New York, 1986).
[77] R. J. De Ayala, *The Theory and Practice of Item Response Theory* (Guilford Publications, 2013).
[78] P. W. Holland and D. T. Thayer, An alternate definition of the ETS delta scale of item difficulty, ETS Research Report Series Research Report RR-85-43 (1985).
[79] P. W. Holland and D. T. Thayer, Differential item performance and the Mantel-Haenszel procedure, in *Test Validity*, edited by H. Wainer and H. I. Braun (Lawrence Erlbaum, Hillsdale, NJ, 1993), pp. 129–145.
[80] W. M. Yen, Scaling performance assessments: Strategies for managing local item dependence, J. Educ. Measure. **30**, 187 (1993).
[81] I. Halloun, R. R. Hake, E. P. Mosca, and D. Hestenes, Force Concept Inventory (revised 1995) (1995), http://modeling.asu.edu/R&E/Research.html Accessed 7/19/2019.
[82] Physport, https://www.physport.org. Accessed 8/8/2017.
[83] U.S. News & World Report: Education, https://premium.usnews.com/best-colleges. Accessed 4/30/2017.
[84] D. Edler and M. Rosvall, The mapequation software package, Available online at http://www.mapequation.org/. Accessed 2/1/2019.
[85] S. P. Borgatti and D. S. Halgin, Analyzing affiliation networks, in *The Sage Handbook of Social Network Analysis*, edited by J. Scott and P. J. Carrington (Sage Publications, Thousand Oaks, CA, 2011), pp. 417–433.
[86] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.15.020122 for the communities detected at the $r > 0.15$ sparsification threshold.
[87] N. J. Foti, J. M. Hughes, and D. N. Rockmore, Nonparametric sparsification of complex multiscale networks, PLoS One **6**, e16431 (2011).
[88] A. Traxler, A. Gavrin, and R. Lindell, Networks identify productive forum discussions, Phys. Rev. Phys. Educ. Res. **14**, 020107 (2018).
[89] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E **69**, 026113 (2004).
[90] M. E. J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E **69**, 066133 (2004).
[91] A. Clauset, M. E. J. Newman, and C. Moore, Finding community structure in very large networks, Phys. Rev. E **70**, 066111 (2004).
[92] S. Epskamp, A. O. J. Cramer, J. L. Waldorp, V. D. Schmittmann, and D. Borsboom, qgraph: Network visualizations of relationships in psychometric data, J. Stat. Softw. **48**, 1 (2012).
[93] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications* (Cambridge University Press, Cambridge, UK, 1997).
[94] A. Canty and B. D. Ripley, boot: Bootstrap R (S-Plus) Functions (2017), R package version 1.3-20.
[95] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Academic Press, New York, NY, 1977).
[96] R. C. Hudson and F. Munley, Re-score the Force Concept Inventory!, Phys. Teach. **34**, 261 (1996).
[97] J. Yasuda and M. Taniguchi, Validating two questions in the Force Concept Inventory with subquestions, Phys. Rev. Phys. Educ. Res. **9**, 010113 (2013).
[98] S. DeVore, J. Stewart, and G. Stewart, Examining the effects of testwiseness in conceptual physics evaluations, Phys. Rev. Phys. Educ. Res. **12**, 020138 (2016).
[99] J. Yasuda, N. Mae, M. M. Hull, and M. Taniguchi, Analyzing false positives of four questions in the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010112 (2018).
[100] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Huevelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).