Exploring the structure of misconceptions in the Force and Motion Conceptual Evaluation with modified module analysis

James Wells,[1] Rachel Henderson,[2] Adrienne Traxler,[3] Paul Miller,[4] and John Stewart[4,*]

[1]*College of the Sequoias, Science Division, Visalia, California 93277, USA*
[2]*Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan 48824, USA*
[3]*Wright State University, Department of Physics, Dayton, Ohio 45435, USA*
[4]*West Virginia University, Department of Physics and Astronomy, Morgantown, West Virginia 26506, USA*

(Received 7 January 2020; accepted 13 April 2020; published 22 April 2020)

Investigating student learning and understanding of conceptual physics is a primary research area within physics education research. Multiple quantitative methods have been employed to analyze commonly used mechanics conceptual inventories: the Force Concept Inventory (FCI) and the Force and Motion Conceptual Evaluation (FMCE). Recently, researchers have applied network analytic techniques to explore the structure of the incorrect responses to the FCI identifying communities of incorrect responses which could be mapped on to common misconceptions. In this study, the method used to analyze the FCI, modified module analysis was applied to a large sample of FMCE pretest and post-test responses ($N_{pre} = 3956$, $N_{post} = 3719$). The communities of incorrect responses identified were consistent with the item groups described in previous works. As in the work with the FCI, the network was simplified by only retaining nodes selected by a substantial number of students. Retaining as nodes only those incorrect answer choices selected by at least 20% of the students produced communities associated with only four misconceptions. The incorrect response communities identified for men and women were substantially different, as was the change in these communities from pretest to post-test. The 20% threshold was far more restrictive than the 4% threshold applied to the FCI in the prior work that generated similar structures. Retaining nodes selected by 5% or 10% of students generated a large number of complex communities. The communities identified at the 10% threshold were generally associated with common misconceptions producing a far richer set of incorrect communities than the FCI; this may indicate that the FMCE is a superior instrument for characterizing the breadth of student misconceptions about Newtonian mechanics.

DOI: 10.1103/PhysRevPhysEducRes.16.010121

I. INTRODUCTION

Understanding common difficulties students exhibit in learning conceptual physics has been an important research strand in physics education research (PER) since its inception. This work was greatly advanced by the introduction of multiple-choice conceptual instruments measuring students' understanding of mechanics and electricity and magnetism: the Force Concept Inventory (FCI) [1], the Force and Motion Conceptual Evaluation (FMCE) [2], the Conceptual Survey of Electricity and Magnetism (CSEM) [3], and the Brief Electricity and Magnetism Assessment (BEMA) [4]. Studies involving these instruments continue to be of central importance in PER. For an overview of the history of these instruments and their use in PER, see Docktor and Mestre's extensive synthesis of the field [5].

Recently, substantial efforts have been made to apply quantitative techniques to further understand these instruments including factor analysis [6–8], cluster analysis [9], and item response theory [10–13]. In 2016, Brewe, Bruun, and Bearden [14] introduced a new class of quantitative algorithms to analyze the incorrect answers, network analytic methods [15,16]. Network analysis is a broad, flexible, and extremely productive field of quantitative analysis that has been used to analyze systems as diverse as the functional networks in the brain [17] and passing patterns of soccer teams [18].

A network is formed of nodes that are connected by edges. Network analysis seeks to identify structure within the network; one important class of structure is subsets of the network that are more interconnected within themselves than they are connected to the rest of the network. These subsets are called "modules" or "communities"

*jcstewart1@mail.wvu.edu

interchangeably. In anticipation of the "igraph" package [19] in the "R" software system [20] becoming the primary tool used within PER for network analysis, we will call the subgroups communities.

Wells *et al.* [21] attempted to replicate the analysis for the FCI of Brewe *et al.* [14] and found that it did not scale to large datasets. They suggested a modified algorithm called modified module analysis (MMA); the details are discussed below as Study 1. In the current study, the MMA algorithm was applied to explore the community structure of the FMCE; the results are then compared to the results of Study 1.

This study sought to answer the following research questions:

RQ1 What incorrect answer communities are identified by modified module analysis in the FMCE?

RQ2 How are these communities different pre- and postinstruction? How is the community structure different for men and women?

RQ3 How do the communities change as the parameters of the MMA algorithm are modified?

RQ4 How do the communities detected compare to those detected in the FCI in Study 1?

A. The FMCE instrument

The FMCE is a widely used mechanics conceptual inventory that measures students' understanding of force and motion. The instrument consists of 43 items examining student understanding of Newton's laws of motion. The items are presented in groups with each item having at least 6 possible responses, some of which represent common misconceptions. Most items include a "none of the above" response which is not the correct response to any item; none of the above responses have been shown to cause psychometric problems [22]. The FMCE is available at PhysPort [23].

For this work, one of the 43 questions in the FMCE is called an "item." A "response" to an item is one of the possible answers to the item. For example, the first question on the FMCE is item 1 and has 7 answer choices, responses 1A, 1B, 1C, 1D, 1E, 1F, and 1G.

The FMCE uses the practice of "blocking" or "chaining" items where multiple items refer to a common stem. In an item block, a physical system is introduced, then multiple items refer to that system. Of the 43 items in the FMCE, all but one (item 39) are included in item blocks. The FCI also employs item blocks with 13 of the 30 items included in blocks. Multiple studies have suggested that blocking items introduces spurious correlations that can make the instrument difficult to interpret statistically [12,21].

Since its introduction, the blocked structure of the FMCE has been used to provide a compact description of the instrument in terms of the qualitative features of the item blocks. This description has been refined since the introduction of the instrument as will be discussed in Sec. II A. The descriptive terms provide an overview of the instrument. "Force sled" items (items 1–7) ask about the force that an individual would need to exert on a sled on a low-friction surface to produce a set of accelerations; students select for a number of textual responses. "Cart on a ramp" items (items 8–10) ask students to select the force on a cart as it moves up and down an incline. "Coin toss–force" items (items 11–13) ask students to select the force on a coin tossed in the air. "Force graph" items (items 14–21) ask students about the force on a toy car as it moves across a low-friction surface; students select from a number of graphs. "Acceleration graph" items (items 22–26) ask students to select the graph that correctly represents the acceleration of a toy car moving on a horizontal surface. "Coin toss–acceleration" items (items 27–29) ask students to select the acceleration of a coin tossed in the air. "Newton III" items (items 30–39) ask students about the forces during a variety of interactions between cars and trucks. "Velocity graph" items (items 40–43) ask students to select the graph that correctly represents the velocity of a toy car moving on a horizontal surface. The current version of the FMCE has four multiple choice "energy" items (items 44–47) and one free-response item (46a). These items were not present in the original FMCE and will not be analyzed in this study.

B. Prior studies

As this analysis was motivated by prior works, this research will draw heavily from two previous studies that will be referenced as Study 1 and Study 2 throughout the manuscript.

1. Study 1: Modified module analysis

In Study 1, Wells *et al.* [21] introduced modified module analysis (MMA), a network analytic method to explore the structure of the incorrect answers of a multiple-choice instrument. Modified module analysis was introduced to allow the application of the module analysis of multiple-choice responses (MAMCR) method of Brewe *et al.* [14] to large datasets. In both MMA and MAMCR, the incorrect responses to a conceptual inventory are used to define a network with weighted edges. The responses are the nodes of the network. In MAMCR, the number of times two responses are selected by the same student defines the edge weight of the network. For example, if FCI response 1D and 2B were selected together by 40 students, the network would contain 1D and 2B as nodes and have an edge between the nodes with weight 40. The notation 1D represents response "D" to item 1. In MMA, the edge weight is the correlation coefficient between the two responses.

To analyze this network, the correlation matrix was calculated and a threshold applied. In Study 1, only edges that were correlated at the $r > 0.2$ level were retained where r is the correlation coefficient. The remaining

correlated items define a network with edge weight equal to the correlation. A community detection algorithm was then applied to detect substructure in the network. A community represents a set of nodes that are preferentially selected together by many students. The MMA algorithm detects incorrect answer communities, subsets of the network formed of incorrect answers which are preferentially selected together. Modified module analysis identified 9 pretest communities and 11 post-test communities on the FCI. Three of the communities were the result of blocked items. For these blocked items, the later response was the correct response if an earlier response had been correct. In most cases, the remaining communities could be related to the misconceptions associated with the items in the original paper introducing the FCI [1] and in the more detailed taxonomy provided by Hestenes and Jackson [24]. For eight of the communities, a dominant misconception was identified and for two of the communities, two common misconceptions were identified. For example, one FCI community included responses $\{4A, 15C, 28D\}$, common incorrect answers to the Newton's 3rd law items. Students were applying both the greater mass implies greater force and the most active agent produces greater force misconceptions for these items.

Study 1 found the communities identified for men and women on both the pretest and post-test, while not identical, were very similar.

2. Study 2: Multidimensional item response theory and the FMCE

Study 1 made extensive use of a prior study of the FCI applying constrained multidimensional item response theory (MIRT) to produce a detailed model of the physical reasoning required to correctly solve the items in the instrument [12]. The incorrect communities not related to the blocking of items often required similar physical reasoning for their solution. This methodology has recently been extended to the FMCE and will be referenced as Study 2. In Study 2, Yang *et al.* performed a detailed analysis of the correct answers to the FMCE using constrained MIRT [25]. This technique produced a detailed model of the instrument in terms of the fundamental reasoning steps (principles) required for its solution. Results of factor analysis and correlation analysis were also presented. All analyses suggested the existence of subsets of items within the instrument that shared a common solution structure. These item groups included items 40–43 (definition of velocity), 22–26 (definition of acceleration), 30–39 (Newton's 3rd law), and 8–13 and 27–29 (motion under gravity). A fifth group of items, items 1–7 and 14–20, measured a combination of Newton's 1st and 2nd law and corollaries of motion derived from these laws. These item groups presented responses to students using different representations with items 1–7 asking students to select textual responses and items 14–20 asking students to choose between two-dimensional graphs. The constrained MIRT analysis demonstrated that the student skill in interpreting a graph was an important factor in understanding student reasoning on the instrument.

The groups identified as requiring a common solution structure are well aligned with the item groups identified by previous research and described in Sec. I A supporting the identification of these groups as measuring distinct elements of Newtonian thinking. Some of the groups suggested by MIRT combine groups suggested by previous authors. For example, cart on a ramp, coin toss–force, and coin toss–acceleration items all require an understanding of the force or acceleration due to gravity for their solution. Item groups with similar correct solution structure will often also have responses that represent consistently applied misconceptions in the analysis that follows.

In general, the FMCE had many more items requiring similar reasoning for their solution than the FCI; this may make it a productive instrument for the exploration of structure of misconceptions about mechanics using MMA.

II. PREVIOUS STUDIES OF THE FMCE

A. General analyses

Multiple subdivisions of the FMCE have been suggested. Thornton and Sokoloff introduced four subgroups of items with the original publication of the instrument: force sled items, cart on a ramp items, coin toss items, and force graph items [2] as described above. Items 5, 6, and 15 were identified as potentially problematic, leading to modified subgroups: force sled items (items 1–4 and 7) and force graph items (items 14 and 16–21).

Using data collected after the instrument's publication, Thornton *et al.* proposed an alternate scoring scheme that eliminated some items and scored some groups of items (clusters) together [26]. The alternate scoring scheme for the clusters suggested item groups 8–10, 11–13, and 27–29 be scored together because students had not mastered the concept tested by the group unless they answered each item in the group correctly. Each cluster received two points if all items were answered correctly, zero points if not. They also suggested the elimination of items 5, 15, 33, 35, 37, and 39 because students without an understanding of Newtonian mechanics often answered them correctly. They also suggested the elimination of item 6 because content experts often answered it incorrectly.

Multiple authors proposed other revisions to the subgroups of items initially introduced by Thornton and Sokoloff. Wittmann identified five subgroups: force (Newton I and II) (items 1–4, 7–14, 16–21), acceleration (items 22–29), Newton III (items 30–32, 34, 36, 38), velocity (items 40–43), and energy (items 44–47) [27]. These subgroups were further refined using a resource framework by Smith and Wittmann who proposed a set of seven subgroups: force sled (items 1–4, 7), "reversing

direction" (items 8–13, 27–29), force graphs (items 14, 16–21), acceleration graphs (items 22–26), Newton III (items 30–32, 34, 36, 38), velocity graphs (items 40–43), and energy (items 44–47) [27]. The problematic items identified by Thornton *et al.* were eliminated from all subgroups in these two studies. More recently, Smith, Wittmann, and Carter applied the revised subgroup structure to understand of the effect of instruction [28].

Study 2 provided partial support for the identification of problematic items of Thornton *et al.* [26], with items 5, 6, 33, 35, and 37 having relatively small discriminations and item 15 having negative discrimination. The models in Study 2 also suggest items 20 and 21 may not be appropriately grouped with the other items probing graphical interpretation of forces.

B. Exploratory analyses

While multiple studies have presented exploratory analyses of the FCI, only two studies have performed factor analysis on the FMCE. Ramlo examined the reliability of the FMCE using a sample of 146 students [29] finding adequate reliability on the pretest (Cronbach's $\alpha = 0.742$) and excellent reliability on the post-test (Cronbach's $\alpha = 0.907$). While the pretest factor structure was undefined, three conceptually coherent factors were identified on the post-test.

In Study 2, exploratory factor analysis found 5, 6, 9, and 10 factor models optimized some fit statistics. Overall, the model fit of the five-factor model was superior. The factor loadings in this model were very consistent with the groups of conceptually similar items identified by the confirmatory MIRT analysis. These groups also had adequate to excellent internal consistency measured by Cronbach's α ranging from $\alpha = 0.66$ to $\alpha = 0.93$. There is also strong theoretical support for the selection of either a 5 or 10 factor model as discussed in Study 2. Study 2 concluded that the three-factor structure identified by Ramlo probably resulted from the low sample size.

C. Gender and the FMCE

On mechanics conceptual inventories (the FCI and the FMCE), men, on average, outperform women by 13% on the pretest and 12% on the post-test [30]. Researchers have explored various factors that could explain the differences between men and women on the FMCE. For example, differences in academic backgrounds and preparation, measured by FMCE pretest and math placement exam scores, have been shown to explain much of the gender gap on the FMCE post-test [31,32]. Studies have also investigated the impact of interactive engagement on the overall gender gap [31,33–35].

While many studies have focused on the overall average gender differences on the FMCE, recently, researchers have explored the fairness in the individual items on the FMCE [36]. An item is fair if men and women of equal overall ability with the material score equally on the item. Applying the modified scoring method proposed by Thornton *et al.* [26], only item cluster 27–29 scored as a single item consistently showed substantial unfairness in multiple samples; this item was unfair to women. In one of the two samples, item 40 demonstrated substantial gender unfairness; this item was also unfair to women.

III. THE STRUCTURE OF KNOWLEDGE

The MMA algorithm detects sets of incorrect answers that are commonly selected together by multiple students. Study 1 showed that, for the FCI, these incorrect answer communities were related to either misconceptions proposed by the authors of the FCI or to the practice of blocking items. The reason students answer physics questions incorrectly is a broad area of research and multiple frameworks have been developed to explain incorrect answering.

A. Knowledge frameworks

Much of the early work in PER conceptualized patterns of incorrect answers as "misconceptions," coherently applied incorrect reasoning often related to Aristotelian or medieval theories of nature. Early research investigated common student difficulties in applying Newtonian mechanics [37–43]. As the field evolved, systematic studies were developed to explore student understanding and epistemology [2,44–47].

Eventually, alternate frameworks not involving misconceptions were proposed. One of the most prominent frameworks is knowledge in pieces [48,49]. Knowledge in pieces models student thinking as resulting from the activation of granular pieces of reasoning (resources) which are used independently or collectively to solve problems. Multiple authors have investigated this model and these reasoning pieces have been called phenomenological primitives (*p* prims) [48,49], resources [50–52], and facets of knowledge [53]. In the knowledge-in-pieces model, misconceptions represent consistently activated *p* prims. Unlike the misconception view, the knowledge-in-pieces model views *p* prims as potentially positive resources that can be activated as part of the process of constructing knowledge.

For a careful and accessible exploration of the relation of and differences between the misconception view and the knowledge-in-pieces framework, see Scherr [54]; the current study applies the definitions from this work. The misconception model is defined as "a model of student thinking in which student ideas are imagined to be determinant, coherent, context independent, stable, and rigid" [54]. The knowledge-in-pieces framework models student ideas "as being at least potentially truth indeterminate, independent of one another, context dependent, fluctuating, and pliable" [54].

The network analysis presented in this work is a quantitative exploratory technique that does not require the adoption of a theoretical framework. Once communities of incorrect responses are identified, one can examine the structure of the communities for evidence of either reasoning better described by the misconception or the knowledge-in-pieces view by applying Scherr's definitions [54].

B. Misconceptions

The FCI was developed using the misconceptions model; Hestenes, Wells, and Swackhamer proposed a detailed taxonomy of the misconceptions measured by the instrument [1]. The taxonomy was developed from qualitative studies investigating students' "alternate view of the relationship between force and acceleration" where researchers interviewed students about their difficulties while solving conceptual physics problems [55–57]. The authors of the FCI provided a detailed description of the misconceptions measured by the instrument [1]; this taxonomy was later refined by Hestenes and Jackson [24]. The analysis in the current work demonstrates that the FMCE probes a limited number of the misconceptions that were originally outlined by the authors of the FCI; only these misconceptions are described below. For more information about the other misconceptions probed by the FCI, see Study 1.

Velocity-acceleration undiscriminated.—The misconception of velocity-acceleration undiscriminated stems from the concept of "motion is vague" [1]. This misconception is characterized by the inability to differentiate the concepts of position, velocity, and acceleration within kinematics. For example, items 22–26 on the FMCE refer to a car moving on a horizontal surface and ask for the acceleration as a function of time. The velocity-acceleration undiscriminated misconception would suggest that when the car is speeding up or slowing down at a constant rate, the graph would show a linear trend of acceleration with respect to time and when the car is traveling at a constant velocity, the graph would show a nonzero constant acceleration.

Velocity proportional to applied force.—The velocity proportional to applied force misconception is one of the subcategories outlined under the "active forces" category of misconceptions describe by the authors of the FCI [1]. This misconception asserts that the force and velocity of an object are proportional; it suggests that Newton's 2nd law is not well understood. For example, items 1–7 on the FMCE probe this misconception; a sled is being pushed along the ice and students are asked to describe the force which would keep the sled moving with a given velocity which changes with time. The velocity proportional to applied force misconception would predict that force is proportional to velocity. For example, in FMCE item 1 the sled is being pushed to the right and speeding up; students applying velocity proportional to applied force misconception would predict the force is to the right and increasing.

Action-reaction pairs.—The misconceptions of greater mass implies greater force and the most active agent produces the greatest force are the two subcategories within the "action-reaction pairs" group of student difficulties. This group of misconceptions suggests that Newton's 3rd law is not well understood. For example, FMCE items 30–32 probe these misconceptions by describing collisions between a heavy truck and a small car. The greater mass implies a greater force misconception would predict that the heavy truck would exert a greater force on the small car than the small car would on the heavy truck. The most active agent produces the greatest force would predict that the object that is moving the fastest would produce the greatest force.

IV. METHODS

A. Sample

The sample was collected at a large eastern land-grant university serving approximately 30 000 students. The demographics of the undergraduate population at the university were 80% White, 6% International, 4% African-American, 4% Hispanic, 2% Asian, 4% two or more races, and other groups less than 1% [58]. The general undergraduate population had a range of ACT scores from 21–26 (25th to 75th percentile).

The data were collected in the introductory calculus-based mechanics course from Spring 2011 to Spring 2017. The majority of the students enrolled in this course were physical science and engineering majors. This sample was previously analyzed in Henderson *et al.* (Sample 3A [36]) where the instructional environment is described in detail. The course was taught by multiple instructors and generally featured an interactive pedagogy in lecture and laboratory.

Over the period studied, the FMCE was given at the beginning and at the end of the class in each semester. The sample contains 3956 FMCE pretest records and 3719 FMCE post-test records (each with 80% men); only the students who completed the course for a grade were included in the study. The overall pretest to post-test gains for men and women were 28% and 21%, respectively. The descriptive statistics for the FMCE pretest and the FMCE post-test are presented in Table II in Henderson *et al.* (Sample 3A) [36].

B. Analysis methods

This work applies MMA described in Study 1 to the FMCE. Although the method is described in detail in Study 1 [21], we provide an overview of the method here.

All responses to the FMCE were dichotomously coded where response $1D_i$ would be coded as one if student i selected response D to item 1 and zero otherwise. The correct responses were eliminated; network analysis is unproductive if the correct responses are included because

they form a single tightly connected community that hides the structure of the incorrect answers. Responses that were selected by fewer than 5% of the students were eliminated as statistically unreliable.

The correlation matrix was calculated for the remaining incorrect responses. This correlation matrix defines a network with nodes representing the incorrect responses and weighted edges between the nodes representing the strength of the correlation between the two responses. Edges that represent correlations that were not significant at the $\alpha = 0.05$ level with a Bonferroni correction applied were eliminated. The network was further simplified by eliminating any correlation where $r < 0.2$; this was the threshold applied in Study 1. This also served to remove the large negative correlations between two responses to the same item. Network analysis often uses methods to simplify the network while retaining important structure; this process is called "sparsification."

Consider the FCI as an example (the FCI is easier to explain because all items have the same number of responses). The FCI has 30 items each with 5 responses. Removing the correct responses leaves 4 responses per item, or 120 total incorrect responses. The answers for each student are used to produce a vector of 120 zeros or ones indicating whether the student selected a particular incorrect response. The correlation between each entry in this vector is then calculated forming a 120×120 matrix where each entry represents the correlation between two responses. Correlations between different responses to the same item and correlations that are not statistically significant are set to zero. A threshold is then applied to sparsify the matrix by setting elements of the matrix smaller than $r = 0.2$ to zero. A row or column of this matrix represents a particular response to an item in the FMCE; this response becomes a node in the network. For example, response G to item 1 becomes node 1G. Two nodes are connected by an edge if the modified correlation matrix has a nonzero value for the two nodes. The weight of the edge between the two nodes is the value of the correlation.

A community detection algorithm was then applied to detect structure in the network. Study 1 applied the "fast-greedy" algorithm [59] included in the igraph package [19] for R. Many community detection algorithms exist; Study 1 reported that most produced similar results for the correlation network. The fast-greedy algorithm is designed to maximize the modularity of the division of the network into unified subnetworks. Modularity measures the number of intracommunity edges in a particular division of the network compared to the number expected in a random division.

To account for randomness in both the sample and the algorithm, 1000 bootstrapped replications were carried out. As a result, 1000 divisions of the network into communities were calculated sampling the data with replacement. For each pair of incorrect responses, the number of times the two responses appeared in the same community was calculated. This number is divided by the number of bootstrap replications to form the community fraction C. In this study, we analyzed communities that were identified in 80% of the 1000 bootstrapped samples.

Because the incorrect answer communities of men and women are compared and the number of men in the sample is significantly larger than the number of women, care was taken to produce a balanced sample. For men, the data were downsampled to the size of the female dataset. For women, the dataset was sampled with replacement preserving the size of the dataset.

V. RESULTS

Modified module analysis was applied to the FMCE; the communities identified are shown in the first table in the Supplemental Material [60]. Retaining nodes where at least 5% of the students selected the response (approximately the threshold used in Study 1) produced 35 communities. These communities were often formed of small subsets of item groups identified in previous studies. This was dramatically different than the small number of communities identified in the FCI by Study 1. The complex nature of the communities identified made understanding their structure difficult.

To produce a simpler structure more open to interpretation, the network was further sparsified retaining as nodes only those incorrect answer choices selected by at least 20% of the students. The community structure of this network is shown in Table I. In nearly every case, the communities form completely disconnected, complete graphs. The intracommunity density measures the connectivity of a community and is defined as $\gamma = 2m/n(n-1)$, where n is the number of nodes and m is the number of realized edges. A fully connected community has an intracommunity density of 1.

Table I offers partial support for the identification of items 5, 6, 15, 33, 35, 37, and 39 as problematic in Thornton *et al.* [26]. Items 20 and 21 were modeled as having a different solution structure to other items in the force graph group in Study 2; these items are inconsistently connected to the other items in this group in Table I. Incorrect answers to items 15, 33, and 37 were never identified as part of a community. Incorrect answers to items 20, 21, 35, and 39 were inconsistently identified as parts of the communities associated with the items in the group. As such, some of the complexity in Table I results from these items. If items 5, 6, 15, 20, 21, 33, 35, 37, and 39 are eliminated from the analysis, the structure of Table I simplifies substantially to produce Table II. The communities in Table II are shown graphically in Fig. 1.

The sets of items in Table I and II generally conform to the item groups identified in previous works and discussed in Sec. I A. Table II suggests items 27–29 should be treated as an independent group; we propose this group be called coin toss–acceleration to distinguish it from items 11–13

TABLE I. Communities identified in the pretest and post-test incorrect answers at $r > 0.2$ and community fraction, $C > 0.8$. Only nodes selected by 20% of the students are included. The number in parentheses is the intracommunity density γ for communities where the intracommunity density is not 1. Newton III* denotes that this community does not contain 31F.

Community	Pretest		Post-test		Item group
	Men	Women	Men	Women	
1A, 2B, 3C, 4G, 5B, 6C, 7E	X	X		X	Force sled
1A, 2B, 3C, 4G, 5B, 6C, 7E, 14A, 16C, 17B, 18H, 19D, 20F			X ($\gamma = 0.88$)		Force sled Force graph
8G, 9D, 10B, 11G, 12D, 13B	X		X		Cart on a ramp Coin toss–force
8G, 9D, 10B, 11G, 12D, 13B, 27G, 28D, 29B		X		X	Cart on a ramp Coin toss–force Coin toss–acceleration
14A, 16C, 17B, 18H, 19D, 20F, 21H	X				Force graph
14A, 16C, 17B, 18H, 19D		X		X	Force graph
22E, 23G, 24B, 25F, 26A, 27G, 28D, 29B	X				Acceleration graphs Coin toss–acceleration
22E, 23G, 24B, 25F, 26A		X	X	X	Acceleration graphs
27G, 28D, 29B			X		Coin toss–acceleration
30A, 31F, 32B, 34B, 36C, 38B, 39D	X				Newton III
30A, 31F, 32B, 34B, 36C, 38B		X			Newton III
30A, 32B, 34B, 35B, 36C, 38B, 39D			X	X	Newton III*

TABLE II. Communities identified in the pretest and post-test incorrect answers at $r > 0.2$ and community fraction, $C > 0.8$. Only nodes selected by 20% of the students are included. Problematic items identified in Study 1 and Study 2 have been eliminated. The number in parentheses is the intracommunity density γ for communities where the intracommunity density is not one.

Community	Pretest		Post-test		Item group
	Men	Women	Men	Women	
1A, 2B, 3C, 4G, 7E	X	X		X	Force sled
1A, 2B, 3C, 4G, 7E, 14A, 16C, 17B, 18H, 19D			X ($\gamma = 0.88$)		Force sled Force graph
8G, 9D, 10B, 11G, 12D, 13B	X		X		Cart on a ramp Coin toss–force
8G, 9D, 10B, 11G, 12D, 13B, 27G, 28D, 29B		X		X	Cart on a ramp Coin toss–force Coin toss–acceleration
14A, 16C, 17B, 18H, 19D	X	X		X	Force graph
22E, 23G, 24B, 25F, 26A, 27G, 28D, 29B	X				Acceleration graphs Coin toss–acceleration
22E, 23G, 24B, 25F, 26A		X	X	X	Acceleration graphs
27G, 28D, 29B			X		Coin toss–acceleration
30A, 31F, 32B, 34B, 36C, 38B	X	X	X	X	Newton III

which become coin toss–force. Both sets of items ask about a coin tossed in the air; items 11–13 ask about the force on the coin, items 27–29 about the acceleration. Smith and Wittmann combined these items into a reversing direction (items 8–13, 27–29) group; MMA suggests this grouping may not be appropriate for all students. We also note that Smith and Wittmann's velocity graphs (items 40–43) group does not appear. This group had relatively poor Cronbach α when used as a subscale in Study 2.

At this level of sparsification, for each item only a single response appeared in each community, indicating that there is a single, dominant incorrect answer that students tend to select. This was consistent between the pretest and the post-test and by gender.

A. The structure of incorrect FMCE responses

Study 2 allows the description of the physical principles tested by each item group. Both force sled and force graph

FIG. 1. Communities identified in the FMCE pretest and post-test for men and women.

test a combination of Newton's 1st and 2nd law and the definition of acceleration. The force graph items also require the use of graphical reasoning. The cart on a ramp, coin toss–force, and coin toss–acceleration groups each require the law of gravitation, that the gravitational force is downward and constant. The acceleration graphs group requires the definition of acceleration and reading a graph. The Newton III group requires Newton's 3rd law.

In addition to the communities being strongly related to the item groups, often multiple item groups testing the same physical principles were part of the same community. Much of the complexity of Table II results from the inconsistent joining of incorrect answers to items testing the same concept. Table III summarizes the item groups, the physical principle tested by the group, and the common misconception characterizing the group.

TABLE III. Item groups, the physical principle tested by the group, and the common misconception applied by the students.

Item group	Community	Physical principle	Misconception
Force sled	1A, 2B, 3C, 4G, 7E	Newton's 1st and 2nd law	Velocity proportional to applied force
Cart on a ramp	8G, 9D, 10B	Motion under gravity	Velocity proportional to applied force
Coin toss–force	11G, 12D, 13B	Motion under gravity	Velocity proportional to applied force
Force graph	14A, 16C, 17B, 18H, 19D	Newton's 1st and 2nd law	Velocity proportional to applied force
Acceleration graphs	22E, 23G, 24B, 25F, 26A	Definition of acceleration	Velocity-acceleration undiscriminated
Coin toss–acceleration	27G, 28D, 29B	Motion under gravity	Velocity-acceleration undiscriminated
Newton III	30A, 31F, 32B, 34B, 36C, 38B	Newton's 3rd law	Greater mass implies greater force Most active agent produces greatest force

The misconceptions represented by the items in the incorrect communities are quite consistent. As in Study 1, we use Hestenes and Jackson's extensive taxonomy of misconceptions measured by the FCI to classify the misconceptions [24]. The force sled, force graph, coin toss–force, and cart on a ramp responses all represent the velocity proportional to applied force misconception; all select a force proportional to the velocity. The acceleration graphs and coin toss–acceleration groups both represent the velocity-acceleration undiscriminated misconception; all select an acceleration proportional to velocity.

Study 1 found that the FCI presented the students with two misconceptions related to Newton's 3rd law: greater mass implies greater force and most active agent produces greatest force. MMA was unable to disentangle the application of these two misconceptions for the FCI. Both misconceptions are also in the same community for the FMCE. Item 30A represents the greater mass implies greater force misconception. Items 32B, 34B, 36C, 38B apply the most active agent produces greatest force misconception. Interestingly, item 31 gives the student a situation where both misconceptions apply, a head-on collision between a large truck and a faster moving car. Response 31F indicates the student does not believe they have enough information to solve the item suggesting they are indeed trying to apply both misconceptions simultaneously.

B. Gender differences in community structure

Both men and women consistently answer the force sled and force graph items incorrectly on the pretest; however, these item groups are identified as different communities. The physical principles needed to solve these items are very similar, but the responses to the force sled items are textual whereas the responses to the force graph items are graphical. This seems to indicate that the representation chosen for the answer affects the application of the misconception on the pretest for both men and women. These item groups continue to be different communities for women on the post-test; for men, they have generally merged ($\gamma = 0.88$) into a single community on the post-test.

Men and women also differ in their application of misconceptions to items involving motion under gravity: Cart on a ramp items, coin toss–force items, and coin toss–acceleration items. Responses to these items form a single community on both the pretest and post-test for women. For men, the coin toss–acceleration responses are in a different community on both the pretest and post-test. These three groups represent different misconceptions with cart on a ramp and coin toss–force responses applying a force proportional to velocity misconception while the coin toss–acceleration responses apply an acceleration proportional to velocity misconception. If a student understands that force and acceleration are proportional, then these two misconceptions should produce the same results. Women answer consistently to all three item groups, while men do not, which seems to indicate women apply both misconceptions consistently, while men do not.

While most communities make theoretical sense, both in terms of the item group suggested for the instrument and the physical principles required to solve items in the group identified in Study 2, one does not. For men, one pretest community combines acceleration graphs with coin toss–acceleration. These items require very different physical reasoning for their correct solution, but apply the same misconception, velocity-acceleration undiscriminated. For these items, the misconception is more important in determining the community than the correct answer structure.

C. The strength of common misconceptions

One potential application of these results is to provide classroom instructors with a measurement of how strongly a misconception is held by their students. The instructor could then tailor his or her instruction to emphasize material on those subjects. The strength of a misconception community, called the "misconception score," is defined as the fraction of the responses within the community that are selected by the student. For example, if a community contains $\{22E, 23G, 24B, 25F, 26A\}$, a student who selected 22E, 24B, and 26A would have a misconception score of sixty percent, while a student who selected all five answer choices would have a score of 100%. A higher score indicates a more strongly held misconception. A student who answered items 22, 23, 24, 25, and 26 correctly would have a misconception score of 0%.

The Mann-Whitney U test [61] was used to determine if the misconception scores were significantly different for men and women on the post-test because the data were highly non-normal and discontinuous. The Mann-Whitney U test is a nonparametric test that may be used instead of the unpaired t test. In this sample, the overall post-test score was higher for men than women: the median number of incorrect responses was 20 for men and 26 for women. The effect size of this difference, measured using Vargha and Delaney's A statistic [62], was small: 0.63. This indicates that a randomly selected female student will have more incorrect answers than a randomly selected male student 63% of the time. If there were no effect, A would be 0.50, reflecting a 50-50 chance of a score from either group being higher. The small, medium, and large effect sizes for Cohen's d correspond to values of Vargha and Delaney's A greater than 0.56, greater than 0.64, and greater than 0.71, respectively.

Table IV presents the A statistic, the mean, 1st quartile (1Q), median (Med.), and third quartile (3Q) for men and women for the misconception scores for each incorrect answer community. While the Mann-Whitney U test found a significant difference in each case, all of the A values were in the small or negligible effect size range. Furthermore, all of the A values were lower than the overall chance of selecting a female student at random with more incorrect

TABLE IV. Percentage of students selecting each incorrect community for the FMCE post-test; mean, 1st quartile (1Q), median (med), and 3rd quartile (3Q). A Mann-Whitney U test was performed to determine if the differences between men and women were significant, the p value is presented. The effect size is given as Vargha and Delaney's A [62], the probability that a randomly selected woman will score higher than a randomly selected man.

	Men		Women				
Community	Mean	1Q, Med, 3Q	Mean	1Q, Med, 3Q (%)	p	$A(\%)$	Misconception
Force sled, Force graph	48	10, 50, 80	59	40, 70, 80	<0.001	59	Velocity proportional to applied force
Cart on a ramp	48	0, 50, 83	59	33, 67, 83	<0.001	61	Velocity proportional to applied force
Coin toss–force							
Acceleration graphs	27	0, 0, 60	35	0, 20, 60	<0.001	56	Velocity-acceleration undiscriminated
Coin toss–acceleration	30	0, 0, 67	44	0, 33, 67	<0.001	62	Velocity-acceleration undiscriminated
Newton III	43	0, 40, 80	46	0, 40, 80	0.07	52	Greater mass implies greater force Most active agent produces largest force

answers than a random male student. This is consistent with the finding in Study 1 showing while significant differences exist between the misconception scores of men and women, that these differences are largely explained by overall differences in the post-test scores of men and women.

For the class studied, students hold the velocity proportional to applied force and the Newton's 3rd law misconceptions more strongly than the velocity-acceleration undiscriminated misconception.

D. Reducing sparsification

Sparsification is a network analytic term for removing edges from a network to reduce its density. In MMA, sparsification is accomplished by removing nodes selected by a small number of students and edges correlated below some threshold ($r < 0.2$ in this study). Sparsification allows important structure to be identified in the network. Table II presents the community structure identified after sparsifying the network by retaining as nodes only those incorrect answer choices selected by at least 20% of the students. This sparsification results in a community structure very similar to that identified in Study 1 with a small number of communities each associated with a misconception discussed in Hestenes and Jackson's [24] taxonomy.

This sparsification threshold is far more strict than that applied in Study 1, which only removed nodes not selected by 30 students (about 4% of the sample). When a similar threshold was applied to the FMCE, 5%, 35 communities were found in either the pretests or post-tests of men and women. These results are presented in the Supplemental Material [60]. Most of these communities were very similar to one another, differing by only a single response in some cases. These differences may have resulted from the very different manner in which the two instruments treat incorrect responses. The FCI presents the student with a number of responses developed from student interviews, most designed to test a specific misconception. Most students select only one or two of the available incorrect answers. The FMCE presents the students with many possible options that come close to exhausting the available responses.

This greater scope of possible answers produces a more complex community structure that offers the possibility of identifying misconceptions not explicitly used to construct the instrument. The communities identified for men and women on the pretest and post-test for responses selected by a minimum of 10% of the students are also presented in the Supplemental Material [60]. The misconceptions represented by communities not identified at 20% sparsification are shown in Table V. While some responses do not have an obvious relation to the general misconception tested by the community (marked with an *), most responses in the communities can be associated with a single misconception. Often these misconceptions are outside the taxonomy [24] developed for the FCI, suggesting students have a much richer set of misconceptions than is measured by the FCI. In Table V, misconceptions identified by Hestenes and Jackson [24] are bolded. Many of the items represent combinations of misconceptions in this taxonomy involving the failure to discriminate force, acceleration, velocity, and position in varying combinations. The items mix the position-velocity undiscriminated, the velocity-acceleration undiscriminated, and the velocity proportional to applied force misconceptions identified by Hestenes and Jackson [24].

VI. DISCUSSION

A. Research questions

This study sought to answer four research questions; the first three will be addressed in the order proposed. The fourth research question compares the results of Study 1 for the FCI to the results of this study. The differences of the FCI and FMCE will be discussed as part of the answer to each of the first three research questions.

RQ1: What incorrect answer communities are identified by modified module analysis in the FMCE? The communities of incorrect responses identified on the FMCE

TABLE V. Misconceptions represented by communities identified in items selected by at least 10% of the students which were not identified in items selected by at least 20% of the students. Items marked * do not have an obvious relation to the misconception. Misconceptions identified by Hestenes and Jackson [24] are bolded.

Community	Misconception
3D, 7D	No force is required to slow an object.
3E, 7C	To slow an object at a constant rate, a decreasing force opposite motion must be applied.
3G, 7A	To slow an object at a constant rate, an increasing force opposite motion must be applied.
8E, 11E, 27E	Gravity exerts a constant force in the direction of motion.
8F, 11F, 27F	Gravity exerts an increasing force in the direction of motion.
8F, 10C, 11F, 13C, 27F, 29C	Gravity exerts an increasing force as an object travels upward and a decreasing force as it travels downward.
8F, 10C, 11F	Gravity exerts an increasing force as an object travels upward and a decreasing force as it travels downward.
11E, 27E	Gravity exerts a constant force in the direction of motion.
14C, 17H, 24G, 26E, 40D, 42C, 43A*	Force-acceleration-velocity undiscriminated from position.
14C, 17D, 17H, 23D*, 24G, 26E, 40D, 42C, 43A*	Force-acceleration-velocity undiscriminated from position.
14C, 17D, 40D, 42C	Force-velocity undiscriminated from position.
14C, 17D, 17H, 40D, 42C, 42H*	Force-velocity undiscriminated from position.
17A, 18D, 19C, 19H, 23F, 24A, 25E, 25G	**Velocity proportional to applied force.**
17A, 19C, 24A, 25E, 42A*	**Velocity proportional to applied force.**
18D, 19H, 23F, 25G	**Velocity proportional to applied force.**
19C, 25E	**Velocity proportional to applied force.**
24F, 26E	**Velocity-acceleration undiscriminated.**
27B, 27C, 29F	Gravitational acceleration not constant and in the opposite direction of motion.
27C, 29F	Gravitational acceleration proportional to velocity and in the opposite direction of motion.

generally conformed to the block structure of the instrument and were associated with item groups identified in previous work. This discussion will focus on the analysis retaining nodes selected by 20% of the students; results retaining nodes selected by 5% and 10% of the students are discussed in RQ3. Modified module analysis showed the responses to item groups proposed by Smith and Wittman were being consistently answered using a common misconception: the force sled (items 1–4, 7), the force graph (items 14, 16–19), acceleration graphs (items 22–26) and Newton III (items 30–32, 34, 36, 38) [27]. The responses to the reversing direction subgroup of items (items 8–10, 11–13, 27–29) [27] were not consistently identified as an incorrect answer community. The responses to the subgroup of items 27–29 sometimes formed their own community and were sometimes grouped with the other responses. We proposed renaming the subgroups: cart on a ramp (items 8–10), coin toss–force (items 11–13), and coin toss–acceleration (items 27–29). Cart on a ramp and coin toss–force responses were identified in the same community both pre- and postinstruction and for men and women; coin toss–acceleration responses were inconsistently identified as part of this community.

As an example of an incorrect answer community, consider $\{8G, 9D, 10B, 11G, 12D, 13B\}$ identified for men on both the pretest and post-test. Items 8, 9, and 10 involve a toy car on an inclined plane. The car is pushed up the plane. The item asks about the forces on the car at different points in its motion. Item 8 asks about the force as the cart moves up the ramp; response 8G asserts the force is up the ramp and decreasing as the cart moves up the ramp. Item 9 asks about the force at the highest point of its motion; response 9D asserts the force is zero. Item 10 asks about the force as the cart moves back down the ramp; response 10B asserts the force is down the ramp and increasing. Each response in the community represents the velocity proportional to applied force misconception. Items 11, 12, and 13 ask about the forces on a coin tossed in the air at different points in its motion: item 11 as it moves upward, item 12 at the top of its motion, and item 13 as it moves downward. Responses 11G, 12D, and 13B apply the same velocity proportional to applied force misconception as responses 8G, 9D, and 10B.

Only four misconceptions were identified retaining responses selected by the 20% of the students: velocity proportional to applied force, velocity-acceleration undiscriminated, and two Newton's 3rd law misconceptions. The Newton's 3rd law misconceptions, greater mass implies greater force and most active agent produces largest force, were not identified as independent incorrect answer communities. This is consistent with Study 1 which also found the two Newton's 3rd law misconceptions in the same

community in the FCI. Also consistent with Study 1, the incorrect answer communities contained items testing the same physical principles as identified in Study 2. The physical principle tested by the item, rather than the misconception, was the most important factor in determining the incorrect answer community. In this study, four separate item groups were associated with the velocity proportional to applied force misconception (Table III): Force sled, force graph, cart on a ramp, and coin toss–force. Study 2 showed that the first two groups required Newton's 1st and 2nd law for their solution while the last two required the law of gravitation. While testing the same misconception, the first two groups were never detected in the same community as the last two groups. This is consistent with Study 1 which identified multiple incorrect answer communities in the FCI measuring the related motion implies active forces misconception; these communities also had similar correct solution structure [12].

Study 2 demonstrated that the FMCE has substantially less complete coverage of mechanics than the FCI which was consistent with previous work by Thornton *et al.* [26]. The FCI also measures a broader set of misconceptions than the FMCE. Communities associated with 9 different misconceptions were identified in the FCI, while only 4 were identified in the FMCE. While covering fewer misconceptions, the FMCE does measure the critical velocity-acceleration undiscriminated and velocity proportional to applied force misconceptions more thoroughly than the FCI. Responses 19A, 20B, and 20C in the FCI are reported to measure the velocity-acceleration undiscriminated misconception in Hestenes and Jackson [24], but were not detected as an incorrect answer community in Study 1. Responses 22A and 26A measure the velocity proportional to applied force misconception in the FCI; these responses were also not detected in the same community in Study 1.

Study 1 also identified 3 communities in the FCI that directly resulted from the blocked structure of the instrument. In these communities, the second item in an item block was the correct answer if the first answer had been the correct answer. No such communities were identified in the FMCE. While extensively blocked, the items in the FMCE do not directly refer to the results of previous items.

The communities identified in the FMCE were generally substantially larger than those identified in the FCI. The FCI contained 13 distinct communities for a 30-item instrument while the FMCE contained 9 communities for a 43-item instrument. In the FMCE, some of the distinct communities resulted from joining other communities. All communities in the FMCE can be formed of 6 groups of items: Force sled, force graph, acceleration graphs, coin toss–acceleration, Newton III, and a community that combines cart on a ramp and coin toss–force. As such, substantially fewer distinct groups of misconceptions are identified in the FMCE; however, the groups were often substantially larger in the FMCE than the FCI. For the FMCE, the fundamental groups have sizes ranging from 3 to 6 with all but one group containing at least 5 responses. Only 2 of the 13 groups in the FCI contain as many as 3 responses with 11 groups containing only two responses. Because the incorrect answer communities contain more responses, the FMCE may provide a substantially more accurate characterization of the strength of the misconception (Table IV) than the FCI.

The MMA method also provided support for eliminating the problematic items which were identified by Thornton *et al.* [26]. With items 5, 6, 15, 20, 21, 33, 35, 37, and 39 included in the analysis, the community structure was complex which made it rather difficult to interpret because some of these items were inconsistently associated with a misconception community.

RQ2: How are these communities different pre- and postinstruction? How is the community structure different for men and women? The pre- and postinstruction differences of the community structure were very different for men and women, and as such, these two questions will be addressed together. The communities identified for men and women were often different; on the FMCE pretest, only three out of the nine communities were the same, while on the FMCE post-test, two out of the nine were the same. The differences were generally the result of joining two communities with similar correct solution structure as identified in Study 2. Men integrated the force sled and force graph item groups on the post-test while women did not; however, women integrated the coin toss–acceleration item group with the cart on a ramp and coin toss–force item groups on the post-test while men did not. As such, neither men nor women were more likely to form more integrated misconceptions with instruction. The same physical reasoning is required to solve the items in the larger integrated misconception groups and, therefore, more consistency in selecting a misconception may represent progress in recognizing the same reasoning is required by the items.

The difference between men and women both pre- and postinstruction was dramatically different than the results of Study 1 for the FCI. Generally, the incorrect answer community structure was very similar for men and women on both the pretest and the post-test for the FCI.

The change in misconception structure between the pretest and the post-test was dramatically different for men and women. For women, the misconception communities identified were completely consistent from the pretest to the post-test. For men, of the five communities identified pre-instruction, only two were identified postinstruction. The differences resulted from the force graph and force sled communities merging postinstruction, possibly indicating that men developed more facility with working with the same type of problem in multiple representations with instruction. Pre-instruction, the acceleration graphs and coin toss–acceleration item groups were combined; these were

separate postinstruction. These groups require different physical principles for their solution; however, both apply the same misconception. This may possibly indicate that men differentiate the ideas of force and acceleration in an inconsistent manner pre-instruction.

These results also help to explain the unfairness that was identified in items 27–29 by Henderson *et al.* [36]. Women consistently integrated this item group (coin toss–acceleration) with the other item groups measuring motion under gravity (cart on a ramp and coin toss–force); men did not. Coin toss–force and coin toss–acceleration items differ only by asking about the force and acceleration on a coin moving under the force of gravity; integrating these misconceptions about force and acceleration may indicate the student holds the misconceptions more strongly.

The strength of the misconception, measured by the misconception score in Table IV, shows how strongly students hold a particular misconception. The misconception score was smaller than the overall difference in FMCE score between men and women showing there are not particular misconceptions more strongly held by men or women. No gender difference in misconception score was larger than a small effect.

RQ3: How do the communities change as the parameters of the MMA algorithm are modified?

Study 1 investigated variations in two network building parameters: the correlation threshold r and the community fraction C. These parameters were adjusted to produce productive community structure using the model of the correct solution structure provided in Study 2 and the taxonomy of misconceptions provided by Hestenes and Jackson [24]. The threshold of the minimum number of students who could select a response was not investigated because productive structure was identified retaining only responses selected by at least 30 students, the minimum statistically viable threshold. The FMCE behaved differently; the misconception structure changed dramatically as the threshold for the minimum percentage of students selecting a response was modified.

Retaining nodes selected by at least 5% of the students, MMA identified 35 incorrect response communities; many of these communities were similar, with some differing by only a single response. Retaining responses selected by at least 10% of the students, the structure of the communities was still complex (Table V) but, in general, a single coherent misconception could be identified for each community. Some, but not all, of these misconceptions were described in the taxonomy proposed by Hestenes, Wells, and Swackhamer [1,63] and refined by Hestenes and Jackson [24].

If responses selected by a minimum of 20% of the students were retained, the community structure simplified substantially (Table I). Examination of the community structure showed that much of the remaining complexity involved the sporadic inclusion of items identified as problematic by Thornton *et al.* [26]. Removal of these items produced the relatively simple community structure in Table II. With the exception of one male pretest community, these communities all measured misconceptions described in Hestenes and Jackson's taxonomy [24] as well as requiring the same physical reasoning described in Study 2. The male pretest community applied the same misconception, but required different physical reasoning for its correct solution.

The FCI and the FMCE community structures were dramatically different if responses selected by 5% of the students were retained. At this threshold, the FCI had only 13 small communities and the FMCE 35 often fairly large communities even though the coverage of the FCI is substantially more broad than the FMCE. These differences likely resulted from two sources: students in the FCI sample scored substantially higher on the instrument than the students in the FMCE sample and the unusual distractor structure of the FMCE. The FCI uses only 5 responses for each question and the incorrect responses were developed from student interviews and include common student incorrect views. The FMCE uses items with more than 5 responses that often generally exhaust the possible responses. This offers far greater latitude for students to express uncommon misconceptions and, therefore, are only selected by a small fraction of the students.

The broad set of misconception communities identified retaining nodes selected by 10% of the students suggest that the state of student incorrect reasoning may be substantially more complex than the structure measured by the FCI.

VII. IMPLICATIONS

The responses to the FCI were constructed to measure common misconceptions allowing Jackson and Hestenes to provide a detailed taxonomy of the misconceptions measured by each item [24]. While common misconceptions were certainly considered in the construction of the instrument, the FMCE presents students with many possible incorrect answers. These answers largely exhaust the possible responses. As such, the FMCE may be a much better instrument for a purely exploratory analysis of student incorrect thinking less tied to the misconception view.

The identification of incorrect answer communities representing the same misconception allows the calculation of a misconception score as a quantitative measure of how strongly the misconception is held. This should allow instructors to determine which misconceptions are most prevalent in their classes and to target instruction to eliminate these misconceptions. Because the FMCE includes more items measuring each misconception, it may provide a more accurate characterization of these misconceptions than the FCI.

VIII. LIMITATIONS

The MAMCR and MMA algorithms require a number of choices to be made by the researcher to produce network structure that is productive in furthering the understanding of a conceptual instrument. As the use of network analysis matures in PER, quantitative criteria for optimally selecting network parameters should be developed.

IX. CONCLUSION

Physics conceptual inventories have played an important role in quantitative physics education research and understanding students' difficulties with conceptual physics continues to be a central research area within PER. Network analysis, specifically modified module analysis, has recently been used as a tool to investigate the common misconceptions on the FCI [21]. The current study replicated this work for the FMCE.

In general, retaining responses selected by 20% of the students, the community structure for the FMCE was consistent with the item groups identified in previous studies [2,27]. The misconceptions represented by these communities were limited: velocity proportional to applied force, velocity-acceleration undiscriminated, greater mass implies greater force, and most active agent produces greatest force. Two of these incorrect answer communities were previously identified in the FCI [21]; however, the velocity-acceleration undiscriminated misconception and the velocity proportional to applied force were only detected as incorrect answer communities in the FMCE. The FCI was found to measure nine misconceptions in the previous study.

The FCI and the FMCE behaved dramatically differently as network parameters were adjusted. For the FCI, including responses selected by 4% of the students, only 13 communities were detected, most with only two responses. Retaining responses selected by a similar percentage of students, 35 communities were detected in the FMCE with up to 15 members.

The evolution of the communities identified was dramatically different for men and women. The communities identified for women did not change from pretest to posttest, while only 2 of the 5 communities identified for men remained consistent. Unlike the FCI, there was little consistency in the communities identified for men and women either pre-instruction or postinstruction.

Overall, modified module analysis was productive in understanding the misconception structure of both the FCI and the FMCE and allowing the comparison of the instruments.

ACKNOWLEDGMENTS

Data collection for this work was supported by National Science Foundation Grants No. EPS-1003907 and No. ECR-1561517.

[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] 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).
[4] 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).
[5] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).
[6] 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).
[7] 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).
[8] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010124 (2018).
[9] C. Fazio and O. R. Battaglia, Conceptual understanding of Newtonian mechanics through cluster analysis of FCI student answers, Int. J. Sci. Math. Educ., **17**, 1497 (2019).
[10] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).
[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] 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] C. Zabriskie and J. Stewart, Multidimensional item response theory and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **15**, 020107 (2019).
[14] 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).

[15] M. J. Newman, *Networks*, 2nd ed. (Oxford University Press, New York, NY, 2018).
[16] K. A. Zweig, *Network Analysis Literacy: A Practical Approach to the Analysis of Networks* (Springer-Verlag, Wien, Austria, 2016).
[17] 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**, 20130521 (2014).
[18] J. Lopéz Peña and H. Touchette, A network theory analysis of football strategies, in *Sports Physics: Proceedings of 2012 Euromech Physics of Sports Conference*, edited by C. Clanet (Editions de l'Ecole Polytechnique, Paris, France, 2012), pp. 517–528.
[19] G. Csardi and T. Nepusz, The igraph software package for complex network research, InterJournal, Complex Syst. **1695**, 1 (2006).
[20] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2017).
[21] 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).
[22] S. DeVore, J. Stewart, and G. Stewart, Examining the effects of testwiseness in conceptual physics evaluations, Phys. Rev. Phys. Educ. Res. **12**, 020138 (2016).
[23] Physport, https://www.physport.org. Accessed 8/8/2017.
[24] 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.
[25] 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).
[26] 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).
[27] 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).
[28] T. I. Smith, M. C. Wittmann, and T. Carter, Applying model analysis to a resource-based analysis of the Force and Motion Conceptual Evaluation, Phys. Rev. ST Phys. Educ. Res. **10**, 020102 (2014).
[29] S. Ramlo, Validity and reliability of the Force and Motion Conceptual Evaluation, Am. J. Phys. **76**, 882 (2008).
[30] 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. ST Phys. Educ. Res. **9**, 020121 (2013).
[31] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).
[32] S. Salehi, E. Burkholder, G. P. Lepage, S. Pollock, and C. Wieman, Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics, Phys. Rev. Phys. Educ. Res. **15**, 020114 (2019).
[33] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[34] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre-and post-studio transition, in *2009 Physics Education Research Conference Proceedings*, Vol. 1179, edited by M. Sabella, C. Singh, and C. Henderson (AIP Publishing, New York, 2009), pp. 173–176.
[35] S. J. Pollock, N. D. Finkelstein, and L. E. Kost, Reducing the gender gap in the physics classroom: How sufficient is interactive engagement?, Phys. Rev. ST Phys. Educ. Res. **3**, 010107 (2007).
[36] 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).
[37] L. Viennot, Spontaneous reasoning in elementary dynamics, Eur. J. Sci. Educ. **1**, 205 (1979).
[38] 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).
[39] A. Caramazza, M. McCloskey, and B. Green, Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects, Cognition **9**, 117 (1981).
[40] P. C. Peters, Even honors students have conceptual difficulties with physics, Am. J. Phys. **50**, 501 (1982).
[41] M. McCloskey, Intuitive physics, Sci. Am. **248**, 122 (1983).
[42] R. F. Gunstone, Student understanding in mechanics: A large population survey, Am. J. Phys. **55**, 691 (1987).
[43] C. W. Camp and J. J. Clement, *Preconceptions in mechanics: Lessons dealing with students' conceptual difficulties* (Kendall/Hunt, Dubuque, IA, 1994).
[44] 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, Paris, 1997), pp. 42–47.
[45] 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. ST Phys. Educ. Res. **7**, 020112 (2011).
[46] N. Erceg and I. Aviani, Students' understanding of velocity-time graphs and the sources of conceptual difficulties, Croat. J. Educ. **16**, 43 (2014).
[47] Sutopo and B. Waldrip, Impact of a representational approach on students' reasoning and conceptual understanding in learning mechanics, Int. J. Sci. Math. Educ. **12**, 741 (2014).
[48] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[49] A. A. diSessa and B. L. Sherin, What changes in conceptual change? Int. J. Sci. Educ. **20**, 1155 (1998).
[50] D. Hammer, Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions, J. Learn. Sci. **5**, 97 (1996).
[51] 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).

[52] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).
[53] 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.
[54] R. E. Scherr, Modeling student thinking: An example from special relativity, Am. J. Phys. **75**, 272 (2007).
[55] J. Clement, Students' preconceptions in introductory mechanics, Am. J. Phys. **50**, 66 (1982).
[56] 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).
[57] J. Clement, Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics, J. Res. Sci. Teach. **30**, 1241 (1993).
[58] US News & World Report: Education, https://premium.usnews.com/best-colleges. Accessed 4/30/2017.
[59] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E **69**, 026113 (2004).
[60] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.16.010121 for the communities detected at the 5% and 10% node retention threshold.
[61] H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. **18**, 50 (1947).
[62] A. Vargha and H. D. Delaney, A critique and improvement of the "CL" common language effect size statistics of McGraw and Wong, J. Educ. Behav. Stat. **25**, 101 (2000).
[63] 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.