

Beyond named methods: A typology of active learning based on classroom observation networks

Meagan Sundstrom^{1,*}, Justin Gambrell², Colin Green³,
Adrienne L. Traxler⁴ and Eric Brewe^{1,†}

¹*Department of Physics, Drexel University, Philadelphia, Pennsylvania 19104, USA*

²*Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA*

³*Department of Physics, Bryn Mawr College, Bryn Mawr, Pennsylvania 19010, USA*

⁴*Department of Science Education, University of Copenhagen, Copenhagen, Denmark*



(Received 1 October 2025; accepted 13 April 2026; published 21 May 2026)

A growing number of introductory physics instructors are implementing active learning methods in their classrooms, and they are modifying these methods to fit their local instructional contexts. However, we lack a detailed framework for describing the range of what these instructor adaptations of active learning methods look like in practice. Existing studies apply structured protocols to classroom observations and report descriptive statistics (e.g., the fraction of class time spent on each activity), but this approach overlooks the complex nature of instruction. In this study, we apply network analysis techniques to classroom observations to define a typology of active learning that considers the temporal and interactional nature of instructional practices. We analyze video data from 30 instructors at 27 institutions who implemented one of the following named active learning methods in their introductory physics or astronomy courses: Investigative Science Learning Environment (ISLE), Peer Instruction, Tutorials, and Student-Centered Active Learning Environment with Upside-down Pedagogies (SCALE-UP). We first create one classroom observation network per instructor that captures temporal sequences of activities measured using the Classroom Observation Protocol for Undergraduate STEM (COPUS). We then calculate the cosine similarity between all pairs of observation networks, create a similarity network where each instructor's classroom observation network is a node and the edges represent cosine similarity, and apply the Infomap clustering algorithm to the similarity network to identify types of active learning instruction. We find five types of instruction: clicker lecture, dialogic clicker lecture, dialogic lecture with short groupwork activities, short groupwork activities, and long groupwork activities. We find no significant relationship between these instruction types and the named active learning methods; instead, implementations of each of the four methods are spread across different instruction types. This result prompts a shift in the way we discuss and study active learning: the names of developed active learning methods may not actually reflect classroom instruction. We also find that student conceptual learning does not vary across the identified instruction types, suggesting that instructors may be flexible in the style in which they implement these methods without sacrificing effectiveness. We discuss the implications of these results for both research and the professional development of college physics instructors.

DOI: [10.1103/physrevphyseducres.22.010143](https://doi.org/10.1103/physrevphyseducres.22.010143)

I. INTRODUCTION

An abundance of research demonstrates that active learning methods, where students engage in interactions

with their peers and co-construct their understanding [1], promote conceptual development and improve the odds of student success more than traditional, lecture-based teaching methods in introductory level science courses [2–5]. Active learning encompasses a broad range of research-based instructional strategies, from specific activity types (e.g., computer simulations and small group work) to named methods (e.g., Peer Instruction [6,7]) and fully developed course plans (e.g., Investigative Science Learning Environment (ISLE) [8,9]).

Despite documented barriers to adoption, such as limited time to learn about active learning methods, rigid classroom

*Contact author: sundstrommeagan@gmail.com

†Contact author: eb573@drexel.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

spaces, and lack of institutional support [10–13], there is evidence that a growing number of physics instructors use active learning methods [14–16]. As these methods become more widely implemented, instructors are continually modifying them to suit their specific contexts (e.g., student populations, institutional norms, and physical classroom spaces) [17–20]. These adaptations mean that active learning methods are rarely implemented “as is,” but rather they take on different forms in practice.

Existing research, however, does not adequately capture the extent of or complexity within the variation in these different adaptations of active learning. Many studies rely on self-reported data from surveys and interviews of instructors about what they do during class [15–22], which may not be accurate [18,23]. Other studies use direct classroom observation to examine instructional practices [24–28], but this work is largely descriptive (e.g., documenting the time spent on different activities) and overlooks nuances in instructional practices (e.g., activities happening simultaneously and the chronological order of activities, described further in the next section). We lack a detailed framework for describing the range of instructional practices that occur in active learning classrooms. Yet such a framework would enable researchers to move beyond coarse characterizations of active learning (such as named active learning methods) and make fine-grained comparisons of instructional behaviors across instructors, courses, and institutions. This framework would also directly support investigations into how specific classroom activities relate to student outcomes and guide professional development efforts by helping instructors identify concrete opportunities for change.

The main goal of this article, therefore, is to characterize different implementations of well-established active learning methods in introductory physics. We analyze 114 h of classroom video recordings from 30 instructors at 27 different institutions using classroom observation networks, which capture the temporal and interactional nature of classroom activities, to define a typology of active learning instruction.

We also relate different types of active learning instruction to student conceptual learning, as prior work has shown that the specific ways that active learning methods are implemented may relate to their effectiveness in supporting student outcomes [28–33]. Weir *et al.* [29], for example, find that in introductory biology, certain types of activities (e.g., worksheets) are more strongly associated with gains in student conceptual understanding than others (e.g., clicker questions). Our previous work suggests that different implementations of Peer Instruction in introductory physics (specifically, the extent to which activities are fully interactive, such as clicker questions, versus vicariously interactive, such as listening to a peer respond to an instructor question) may be differentially associated with student learning [28].

II. BACKGROUND AND CURRENT STUDY

Here we summarize how researchers have conducted and analyzed classroom observations in prior work and outline the approaches used in the current study.

A. Conducting classroom observations

Systematic classroom observations are typically conducted using established protocols—tools that allow researchers to reliably document student and instructor activities during instruction. In their literature review [34], Anwar and Menekse identify eight classroom observation protocols that have been used to characterize instructional practices in college-level science courses: Reformed Teaching Observation Protocol [35], Oregon Teacher Observation Protocol [36], VaNTH Observation System [37], Cooperative Learning Observation Protocol [38], Teaching Dimension Observation Protocol (TDOP) [39,40], Classroom Observation Protocol for Undergraduate STEM (COPUS) [41], Classroom Interactive Engagement Observation Protocol [42], and Student Resistance and Instructional Practices [43]. These protocols span both open-ended (e.g., writing field notes) and structured (e.g., checking off when predefined activities occur) approaches, use Likert-scale ratings of activities and/or record whether or not predefined activities occur (without rating the activities), focus on student and/or instructor behaviors, and can be conducted in real time (i.e., with a live observer in the classroom) and/or with classroom video recordings.

In this study, we use the COPUS, which records the presence of predefined classroom activities at 2-min time intervals (described further in Sec. III B 1). We chose this protocol because it is both structured (i.e., not open-ended or ratings-based, which may be biased by an observer’s subjective viewpoints) and standardized (i.e., it can be compared across different courses). The COPUS also distinguishes between student and instructor activities, allowing our analysis to only focus on the instructor portion of the protocol (i.e., to characterize instructional practices). Finally, the COPUS is suitable for either real-time observations or video-recorded observations. We collected and analyzed video recordings of classroom instruction so that we could include data from many institutions and iteratively establish reliability among multiple observers.

B. Analyzing classroom observations

Once a structured, standardized classroom observation protocol is applied in real time or to a video recording, researchers typically conduct quantitative analyses to understand the types of activities that occurred during instruction. The simplest form of analysis relies on descriptive statistics, such as calculating the fraction of class time spent on each activity in the protocol. A handful of studies have used descriptive statistics to characterize

instruction [24–28]; for example, Commeford *et al.* [27] compare the fractions of class time spent on each activity in the COPUS across six named active learning methods in introductory physics. Such descriptive studies are useful for documenting prevalent classroom activities and establishing baselines of instructional practices across a full dataset, but they are not suitable for identifying groups of similar observations at scale (i.e., a researcher would need to manually compare and contrast every observation’s descriptive statistics to make groupings), as we aim to do in the current study.

Another body of work applies clustering algorithms, such as *k*-means clustering, latent class analysis, and latent profile analysis (LPA) to the descriptive statistics of many observations in order to identify groups of observations with similar features [30,44–50]. Across many COPUS observations of introductory-level science courses, for example, Stains *et al.* use LPA to identify three types of instruction: didactic, interactive, and student-centered [48]. Our recent work applies LPA to COPUS observations of four named active learning methods in introductory physics [30]. We find four types or “profiles” of instruction, each of which is characterized by one predominant activity: lecture, clickers (i.e., clicker questions), worksheets, and other groupwork (which encompasses, for example, laboratory activities and solving problems at whiteboards).

These clustering methods, however, still make use of descriptive statistics related to classroom instruction, and, as Erdmann and Stains note, “the use of summary statistics alone has the potential to obscure interesting trends and meaningful differences” in instructional practices [51] (p. 2). Descriptive statistics ignore the interactional and chronological nature of classroom activities: multiple activities may occur simultaneously, and activities precede and follow one another over time rather than occur in isolation. Aggregating the total time spent on an activity also masks any information about the duration of each instance of the activity. As an example, consider two instructors of an hour-long class whose descriptive statistics indicate that they both guided students through a worksheet activity for 40 min and led a follow-up discussion of that activity for 20 min. These two instructors’ classes may have looked very different in practice: one may have proctored the worksheet activity for 40 min and then followed up on that worksheet for 20 min, while the other may have cycled through proctoring worksheets for 10 min and then following up for 5 min four times. There remains a need for an analysis tool that captures these subtleties in instructional practices in order to improve our characterizations of active learning instruction and, subsequently, our understanding of how these specific types of instruction impact student outcomes.

Erdmann and Stains outline one possible method to advance our analysis of classroom observations [51]. The

authors apply methods from bioinformatics, treating classrooms as genomes. Essentially, they represent COPUS observations as strings of activities, where activities and their durations are represented along a timeline (indicating their order), and activities can overlap with one another. This method overcomes some of the weaknesses of descriptive statistics mentioned above: it captures the sequential nature of instructional events as well as co-occurrence. The approach can also be scaled to look at patterns across many observations, such as identifying the fraction of all observations in which different activities precede and follow clicker questions to get a better sense of how clicker questions are implemented. One limitation of this method, however, is that it does not facilitate quantitative comparisons between observations at the grain size of an entire observation (i.e., the example in the previous sentence only focuses on one activity, clicker questions). To do so, a researcher would need to qualitatively examine the complete strings of activities across observations and note similarities and differences.

Hora *et al.* introduce yet another approach to capture complexities in instructional practices: classroom observation networks [39,40,52]. In their analysis of TDOP observations, network nodes represent activities from the protocol and network edges represent the relative amount of time that pairs of activities co-occur. The networks contain information about both the time duration of activities and co-occurrence, while also affording direct comparisons between entire observations (either between different observations from a single instructor or between observations from different instructors) because every network has the same set of nodes, and edges can be normalized by class duration. For instance, in one of their studies [52], the researchers compared the classroom observation networks of two different biology instructors whose instructional styles appeared similar according to the descriptive statistics alone: they both lectured for most of the class time. The networks, however, distinguished specific features of each instructor’s lecturing style. One instructor typically used PowerPoint slides and integrated anecdotes and humor into their lectures, while the other instructor integrated various student-centered activities and instructional technologies into their lectures.

Classroom observation networks, therefore, offer a useful tool for both capturing nuance in instructional practices *and* directly comparing observations. We leverage this network approach in the current study to categorize the teaching practices of the 30 instructors in our dataset.

C. Current study

Classroom observation networks have the potential to improve our characterization of active learning methods, but to our knowledge, they have not been applied in other contexts beyond the few studies in biology mentioned above [39,40,52]. In the current study, we use such

networks to explore variation in introductory physics instructors' implementations of named active learning methods. To define a typology of active learning instruction (i.e., identify and interpret groups of instructors with similar teaching practices), we move beyond Hora *et al.*'s qualitative comparison of network structures and quantify the extent to which instructors' observation networks are similar using cosine similarity (described further in the next section). We then follow the "network-of-networks" approach established by Bruun *et al.*, where the grouping of similar classroom observation networks is performed by applying a clustering algorithm to a similarity network in which the nodes represent the classroom observation networks themselves and the edges are the cosine similarities between the networks [53]. This approach allows us to not only distinguish and characterize the instruction types, but also to visualize the relationships *between* the types (which is not possible with other clustering methods, such as *k*-means clustering, latent class analysis, and latent profile analysis). We also examine the relationships between instruction type, course and institutional attributes, and student conceptual learning. The following research questions guided our study:

1. What typology of instructional practices used in named active learning methods may be inferred from the resultant clusters of classroom observation networks?
2. How are the instruction types related to the profiles found in our previous analysis of these data [30], the named active learning method, and other course and institutional attributes (e.g., number of enrolled students)?
3. How are the instruction types related to student conceptual learning?

III. METHODS

In this section, we first describe our procedures for collecting classroom observation and student conceptual learning data and then provide step-by-step details on how we analyzed these data. All analyses were conducted in R (Version 4.3.2). Deidentified data and analysis scripts are available at Ref. [54].

A. Data collection

As part of an ongoing national research project aimed at characterizing different active learning methods used to teach introductory physics, we recruited instructors who self-reported using one of the following active learning methods in their introductory physics or astronomy course:

1. ISLE [8,9]: In all course components ("whole-class" implementation), or only in laboratories ("lab-only" implementation), students take on the role of novice scientists and engage in a lab-based learning

cycle. In small groups, students observe a physics experiment, explain their observations, make predictions about new experiments, design and conduct these experiments, and revise their explanations.

2. Peer Instruction [6,7]: During lectures, students work in small groups of nearby peers to answer multiple-choice questions using electronic clickers or another form of voting system. Typically, class sessions start with a short lecture and then cycle through several multiple-choice questions where the instructor poses a question, students answer the question individually, students discuss the question in small groups, students reanswer the question, and then the instructor explains the solution.
3. Tutorials in Introductory Physics [55] and Astronomy [56]: In all course components (whole-class implementation), or only in recitation sections ("recitation-only" implementation), students complete worksheets in small groups. These worksheets are highly scaffolded and are intended to elicit, confront, and resolve common misconceptions.
4. Student-Centered Active Learning Environment with Upside-down Pedagogies (SCALE-UP) [57]: The course components (i.e., lectures, labs, and recitations) are fully integrated and take place in a classroom designed to promote student interaction, often containing large tables that seat nine students and whiteboards along the classroom perimeter. During class sessions, students solve conceptual and quantitative physics problems and complete lab activities in small groups.

We note that several other named active learning methods for introductory physics exist and were part of the preliminary study for this project [27,49] (e.g., Modeling Instruction [58] and Context-Rich Problems [59,60]). We limited our current study to the above four methods because they are both well-known and widely implemented [16], and we aimed to collect data from multiple implementations of each method.

We recruited instructors in several ways. First, we contacted a grant advisory board member affiliated with each active learning method and asked for names of instructors who were known to be using the method. Beyond these lists, we used additional approaches such as reaching out to instructors in our personal networks, posting advertisements for our study within the American Physical Society, and snowball sampling (i.e., asking instructors who participated in our study to identify other instructors they knew who were implementing the same method). Data collection took place in the Fall 2023, Spring 2024, and Fall 2024 semesters. In total, 31 instructors participated in the study, each with one course in one semester (i.e., no instructors participated in the study

TABLE I. Number of courses included in our analysis by data source and by active learning method. All courses included in the concept inventory analysis are also included in the classroom observation analysis.

Method	Observations	Concept inventory
ISLE	6	5
Peer Instruction	9	9
Tutorials	8	7
SCALE-UP	7	5

with multiple courses). Each instructor received \$1000 for participating.

In this paper, we focus on two data sources that we asked participating instructors to collect: classroom observations and concept inventories (i.e., research-developed and validated assessments used to measure student understanding of physics concepts). 30 out of the 31 instructors collected classroom observations, and 26 out of these 30 instructors collected concept inventory data. All 30 of these instructors are included in the analysis presented here (Table I). These instructors come from 27 unique institutions in the United States, including both public and private institutions, a range of research-intensive and undergraduate-focused institutions, and several minority-serving institutions (Table II).

For the classroom observations, we asked instructors to use Zoom, their own camera, or a camera that we provided to them via mail to record videos of three consecutive class sessions, capturing a “typical week” in the course as recommended by prior studies [29,48]. All courses were conducted in person, and videos included both students and

the instructor(s) in the camera’s view. All 30 instructors recorded three full class sessions, totaling 114 h of classroom video across 90 observations. In all Peer Instruction courses, instructors only recorded lecture sections (and not any lab or recitation sections student may have also been attending). All SCALE-UP instructors recorded the entirety of three class sessions (i.e., possibly including lab activities). In all ISLE courses, the instructors only recorded the lecture sections. This means that the video recordings may not capture all aspects of ISLE for the lab-only implementations, as the main portion of the method was implemented in the lab sections. However, the recorded lecture sections contained elements of ISLE and active learning, and we aim to examine how instructional practices may vary by *reported* active learning method, so it is appropriate to still refer to the lab-only implementations as ISLE courses. Similarly, in one recitation-only Tutorials course, the instructor only recorded lecture sections (and not recitation sections, where the Tutorials worksheets were implemented). The lecture section of this course contained elements of active learning (e.g., clicker questions), and we still refer to this course as Tutorials. In the other Tutorials courses, instructors recorded the class section where Tutorials were being implemented (i.e., lecture sections for whole-class implementations and recitation sections for recitation-only implementations). We denote all of these different cases when we present the results.

For the concept inventories, we asked instructors to choose an existing concept inventory to give to their students, such that the concept inventory aligned with the topics taught in the course (e.g., because instructors taught a range of student populations and courses spanned both physics and astronomy). Instructors administered the assessment at both the beginning (before any exposure to the material) and the end of the semester. Most instructors used the Force Concept Inventory [61], Force and Motion Concept Evaluation [62], or Mechanics Baseline Test [63] (Table IX in Sec. A 6 in the Appendix). All 26 instructors had more than 40% of and/or at least 30 enrolled students with matched responses (i.e., completing both the pre-semester and postsemester concept inventory; response rates ranged from 41% to 98%, with a mean of 72%). We did not impute any concept inventory data because we did not have any information about individual students (e.g., demographics) besides their scores [64].

B. Data analysis

1. Video coding

We used the COPUS [41] to characterize instruction from the classroom video recordings. The COPUS records whether or not certain student and instructor activities, or “codes,” occur in each 2-min time interval.

The first three authors individually applied the COPUS to one video observation (i.e., one of the three class sessions recorded by one instructor) from each of the four

TABLE II. Characteristics of the 27 unique institutions included in our study. Carnegie classifications of research activity are from 2025. R1, Very High Research Spending and Doctorate Production; R2, High Research Spending and Doctorate Production; and RCU, Research Colleges and Universities.

Type of institution	N (%)
Public/private	16 (59%)/11 (41%)
Highest degree awarded	
Associate’s	1 (4%)
Bachelor’s	2 (7%)
Master’s	9 (33%)
Ph.D.	15 (56%)
Carnegie classification of research activity	
R1	10 (37%)
R2	3 (11%)
RCU	5 (19%)
Minority-serving status	
Hispanic-serving institution	3 (11%)
Women’s college	1 (4%)

TABLE III. Abbreviations and definitions of the nine instructor codes in the COPUS [41] that we included in this study.

Code	Abbreviation	Definition
Lecturing	Lec	Presenting content to the whole class
Real-time writing	RtW	Writing in real time on a projector or board while presenting to the whole class
Clicker question	CQ	Asking a clicker (or other voting system) question to the students
Posing question	PQ	Asking a nonclicker (and non-rhetorical) question to the students
Moving and guiding	MG	Moving through class and guiding ongoing student work during an active learning task
Following up	FUp	Providing feedback on clicker question or another activity for the whole class
Answering question	AnQ	Listening to and answering student questions in front of the whole class
Demonstrations/videos	D/V	Showing a demonstration, video, simulation, or animation
Administration	Adm	Presenting administrative information (e.g., related to exams) to the whole class

active learning methods. We then met to discuss disagreements and clarify inclusion and exclusion criteria for each COPUS code. The three coders iteratively recoded the videos and met for discussions until we reached greater than 80% agreement (as measured with Cohen's kappa [65]) for each of the four observations for each pair of coders. After reaching this level of agreement, we randomly assigned one coder to each of the remaining observations such that any resulting patterns among the observations are not attributable to systematic differences between coders. The first author coded 46 video observations, the second author coded 32 video observations, and the third author coded 12 video observations.

In our coding, we used the full COPUS, which includes 13 student codes and 12 instructor codes. While student behaviors are certainly relevant to characterizing active learning instruction, we only included instructor codes (Table III) in our analysis for two reasons. First, including fewer codes (i.e., variables) ensures that analytically determined clusters of similar observations are meaningfully distinct from one another and interpretable. Clusters defined by over 20 variables may be difficult to characterize, especially if only a few observations are assigned to each cluster (as will be the case with our sample size of 30 instructors). Second, as we will discuss in the next section, we used network analysis to capture transitions between COPUS codes. These sequences of activities are only meaningful for one group of people at a time; that is, instructor codes transition between one another separately from student codes transitioning between one another (though they may be correlated). We would need to make separate networks for instructor and student codes for each course, complicating our ability to calculate and interpret one similarity metric between each pair of courses. We acknowledge that this approach limits our characterizations (e.g., we cannot distinguish whether students are working in groups and/or individually when the instructor poses a clicker question); however, focusing on instructor behaviors still serves as a useful launching point for establishing a nuanced framework of active learning.

We also note that we excluded three of the instructor codes (one-on-one extended discussion with one or a few

students while ignoring the rest of the class, waiting around, and other) because they were rarely applied to the observations in our dataset (each of these three codes was observed in less than 2% of all 2-min time intervals across the entire dataset).

2. Constructing classroom observation networks

We used network analysis to represent classroom observations as complex systems. Networks consist of *nodes*, the actors or entities relevant to the system, depicted as circles or rectangles, and *edges*, the connections or relationships between the actors, depicted as lines connecting the circles or rectangles. Edges may have arrows to indicate the direction of a relationship and/or weights to indicate the strength of a relationship.

We constructed one classroom observation network for each of the 30 instructors (i.e., we aggregated the three video observations together for each instructor). All networks included nine nodes, one per COPUS code (Table III). Directed edges represented the chronological order of codes, pointing from an initial code to another code that occurred subsequently in the observation. The edges were also weighted as the number of transitions that occurred from one code to another divided by the total number of 2-min time intervals in the instructor's three recordings (to normalize for different class lengths across instructors). We focused on transitions between codes rather than the co-occurrence of codes, as in Hora *et al.*'s work [39,40,52], because instructor codes do not frequently co-occur (i.e., the instructor predominantly does one activity at a time), and we wanted to capture the flow of classroom activities over time (as in Ref. [51]).

The chronological order of codes is not obvious in COPUS observations because multiple codes can be recorded in the same time interval. If an observer records that two codes occur in the same 2-min time interval, this could indicate either that the two codes occurred simultaneously or that one code occurred and then the other code occurred afterward (and we cannot tell which one occurred first). As our analysis aimed to characterize transitions between instructional activities and not the co-occurrence

TABLE IV. Toy example of COPUS coding for 10 min of class time. Each row represents a 2-min time interval. One (zero) indicates the presence (absence) of the code within the 2-min time interval.

Lec	RtW	CQ	PQ	MG	FUp	AnQ	D/V	Adm
1	0	0	0	0	0	0	0	1
1	0	1	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0
1	0	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0

of activities, we only considered a transition between codes to be when a code that was not present in the previous 2-min time interval became present in the following 2-min time interval. We created an edge from all codes in the previous time interval to the newly appearing code(s) in the following interval. As an example, consider the 10 min of COPUS coding in Table IV. This observation would yield the following edges in an observation network (shown in Fig. 1):

- Lec $\rightarrow$ CQ (weight = $2/5 = 0.4$), from the first to second interval and from the fourth to fifth interval,
- Adm $\rightarrow$ CQ (weight = $1/5 = 0.2$), from the first to second interval,
- Lec $\rightarrow$ FUp (weight = $1/5 = 0.2$), from the second to third interval,

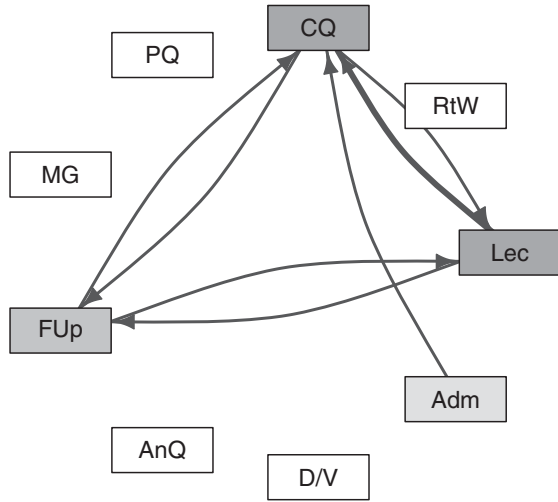


FIG. 1. Toy classroom observation network for the example COPUS coding in Table IV. Nodes (rectangles) represent COPUS codes. Though not directly part of our analysis, node color indicates the fraction of all 2-min time intervals that the code was present: darker shades indicate larger fractions (i.e., longer time durations). Edges (lines with arrows) point from an initial code to the code that occurs in the following 2-min time interval of the COPUS observation. Edge weight (line thickness) is proportional to the number of transitions occurring between those two codes normalized by the total number of 2-min time intervals observed.

- CQ $\rightarrow$ FUp (weight = $1/5 = 0.2$), from the second to third interval,
- CQ $\rightarrow$ Lec (weight = $1/5 = 0.2$), from the third to fourth interval,
- FUp $\rightarrow$ Lec (weight = $1/5 = 0.2$), from the third to fourth interval, and
- FUp $\rightarrow$ CQ (weight = $1/5 = 0.2$), from the fourth to fifth interval.

The first two intervals in Table IV, for example, do not produce an edge from Lec to Lec or from Adm to Lec because Lec occurs in the first interval and so may have already been happening when the second time interval started, which would not be a transition to a new code. Our definition of transitions between codes also prevents any self-loops (i.e., edges from a node to itself, indicating the instructor is doing the same activity in multiple back-to-back time intervals) from appearing in the network. In cases where the instructor is doing the same activity for long periods of time, these self-loops would dominate the network, and information about transitions between activities would be suppressed. We note, though, that our observation networks still reflect whether individual codes occurred for long, continuous time periods versus short, interspersed time periods. If a network only has a few edges (i.e., few transitions between codes), for example, this would imply that the instructor was doing only a few activities for extended time periods.

3. Constructing the similarity network

To quantify the extent to which each pair of classroom observation networks was similar, we considered two measures: the Jaccard index (as in Ref. [66]) and cosine similarity (as in Refs. [67–69]). Both measures have been used to compare networks with the same set of nodes (as we have in our study) [70], and they have been shown to produce similar rankings of similarity between pairs of networks [71]. We calculated both of these measures for every pair of classroom observation networks in our analysis (see Sec. A 1 in the Appendix). Cosine similarity produced many more intermediate values in its possible range, whereas the Jaccard index was suppressed to low values for almost every pair of classroom observation networks (previous work has also shown that this measure is volatile when networks are sparse [72,73]). We used cosine similarity in the remainder of our analysis because the wider range of values allows for more distinct clusters of observation networks to be created.

Cosine similarity measures the structural similarity between two networks. For a pair of directed, weighted networks, G_1 and G_2 , with the same set of nodes V and respective vectors of edge weights for each pair of nodes $\vec{v}_1 = [W_1(1,2), W_1(1,3), \dots, W_1(V, V-1)]$ and $\vec{v}_2 = [W_2(1,2), W_2(1,3), \dots, W_2(V, V-1)]$, cosine similarity is calculated as

$$\text{cosine similarity} = \frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\| \|\vec{v}_2\|}.$$

Values range from 0 to 1 (assuming there are no negative edge weights, as is the case in our analysis), with values closer to 1 indicating that the two networks have more similar sets of edges and patterns in the relative distribution of edge weights.

The matrix of cosine similarity values between every pair of classroom observation networks can itself be represented as a network, and we will refer this network as the “similarity network.” In the full similarity network, there are 30 nodes representing the classroom observation networks (one per instructor) and 435 undirected edges (there is an edge connecting every possible pair of instructors) with edge weights equal to the cosine similarity between the two connected instructors.

4. Sparsifying the similarity network

The fully connected similarity network is difficult to analyze; therefore, we followed the approach in Ref. [53], which uses the locally adaptive network sparsification (LANS) algorithm [74] to remove insignificant edges. Rather than setting a universal, fixed threshold where we remove edges with weights lower than the threshold (which may remove edges with low weights compared to the rest of the network but that are locally important to the nodes to which they are connected), the LANS algorithm considers the distribution of edge weights for each individual node and preserves edges that are statistically significant for that node. Mathematically, the LANS algorithm makes use of fractional edge weights between nodes i and j , defined as

$$p_{ij} = \frac{W_{ij}}{\sum_{k=1}^{N_i} W_{ik}},$$

where N_i is the number of nodes connected to node i . For each node in the network, the fraction of edges with weights less than or equal to p_{ij} is

$$\hat{F}(p_{ij}) = \frac{1}{N_i} \sum_{k=1}^{N_i} \mathbf{1}\{p_{ik} \leq p_{ij}\},$$

where $\mathbf{1}$ is the indicator function (equal to 1 if the statement in brackets is true and 0 otherwise). For a given node, $\hat{F}(p_{ij})$ is effectively the probability of choosing an adjacent edge at random that has a weight less than or equal to p_{ij} . Therefore, edges are considered to be locally significant if $1 - \hat{F}(p_{ij})$ is less than a predetermined significance level, α . All nonsignificant edges are removed from the network, retaining only those edges that are locally important to the network structure. In this analysis, we set $\alpha = 0.05$, which retained 46 edges in our similarity

network and did not leave any nodes isolated (i.e., with no adjacent edges). A detailed description of the methods we used to choose this significance level is provided in Sec. A 2 in the Appendix (Fig. 7).

5. Identifying clusters in the sparsified similarity network

Following Ref. [53], we applied the Infomap community detection algorithm [75,76] to the sparsified similarity network to find clusters of nodes (i.e., classroom observation networks) that are more similar to each other than they are to other nodes. The Infomap algorithm identifies clusters within a network by modeling the flow of information as a random walk and compressing the description of this flow, grouping nodes into clusters that minimize the expected description length of a random walker’s path. The expected description length uses a two-level codebook: one for clusters and one for nodes within the clusters. Mathematically, the algorithm aims to minimize the map equation:

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^m p_{\circlearrowleft}^i H(\mathcal{P}^i),$$

where $q_{\curvearrowright}$ is the probability of the random walker exiting a cluster, $H(\mathcal{Q})$ is the entropy of the cluster-level codebook, $p_{\circlearrowleft}^i$ is the probability of the random walker staying within cluster i , $H(\mathcal{P}^i)$ is the entropy of the codebook within cluster i , and m is the number of clusters. Further mathematical details can be found in Ref. [75].

The output of the algorithm is a partition of the similarity network into clusters of nodes with similar characteristics, with each node assigned to one cluster. In this study, the clusters represent instruction types, and cluster assignments indicate the instruction type used by each instructor.

6. Interpreting the clusters

To define a typology of active learning instruction (i.e., the first research question), we characterized the classroom observation networks assigned to each emergent cluster using several node-level and network-level measures (summarized in Table V), similar to prior work [77]. For the classroom observation networks assigned to each cluster, we calculated the average value of each of these measures (i.e., these measures were calculated using the classroom observation networks, not the similarity network). Node-level measures were calculated for all nine nodes (i.e., COPUS codes) in the classroom observation networks and provided information about the role and relative importance of each COPUS code to each instructor’s style (e.g., whether clickers are central to the way they teach). Network-level measures provided information about overall instructional style (e.g., distinguishing instructors who frequently transition between activities

TABLE V. Definitions and interpretations of the node-level and network-level measures calculated for the classroom observation networks assigned to each cluster.

Measure	Definition	Interpretation in this study
Node-level		
Indegree	The number of adjacent, incoming edges to a node (ignoring edge weight).	The number of unique other COPUS codes that precede a given code (e.g., if posing a clicker question always follows lecturing, CQ would have an indegree of 1).
Outdegree	The number of adjacent, outgoing edges to a node (ignoring edge weight).	The number of unique other COPUS codes that follow a given code (e.g., if lecturing always precedes posing a clicker question, Lec would have an outdegree of 1).
Betweenness	The fraction of all shortest paths between every pair of nodes in the network (where path length is the sum of edge weights) that pass through a given node.	The extent to which a code serves as a “bridge” between other codes (e.g., if lecturing often occurs in between other activities, Lec will have a betweenness closer to 1).
Network-level		
Edges	The total number of edges in the network (considering direction but ignoring edge weight).	The number of unique pairs of COPUS codes with transitions occurring between them.
Total strength	The sum of all edge weights in the network.	The total number of transitions between COPUS codes (normalized by class length).
Mutuality	The fraction of edges that are reciprocal (i.e., there are edges from code A to code B and from code B to code A).	The extent to which transitions between codes occur in both orders (e.g., lecturing both precedes and follows posing a clicker question) rather than one (e.g., lecturing always precedes posing a clicker question).
Transitivity	The proportion of two-paths (e.g., code A follows code B and code B follows code C) that close to form triangles (e.g., code C also follows code A).	The extent to which codes occur in short, cyclical sequences (e.g., iteratively lecturing, posing a clicker question, and following up on that clicker question).

from those who do a few activities for extended periods of time). We triangulated these average measures with the classroom observation networks themselves and our qualitative knowledge of the classroom video recordings to inform our labels and interpretations of the clusters. These labels and interpretations were discussed and negotiated among the full research team until we reached consensus.

7. Segregation analysis

With the typology of active learning instruction defined, we aimed to identify the extent to which instructors were assigned to clusters (i.e., instruction types) with other instructors having similar course or institution attributes (i.e., the second research question). We examined the following attributes:

- profile from our previous LPA (lecture, clickers, worksheets, or other groupwork) [30] of these data to which the majority of the instructor’s three classroom observations were assigned (the previous analysis did not aggregate each instructor’s three class sessions; we excluded one course whose three observations were each assigned to different profiles),
- active learning method (ISLE, Peer Instruction, Tutorials, or SCALE-UP; we did not separate different implementations of ISLE and Tutorials, given the small sample size),

- class size (smaller or larger than 45 enrolled students, see Sec. A 3 in the Appendix for details on how we chose this threshold),
- discipline (physics or astronomy),
- public or private institution,
- research designation (R1/R2 or other), and
- Ph.D.-granting or non-Ph.D.-granting institution (see Table IX in Sec. A 6 in the Appendix).

We looked at the alignment of instruction types with our previous latent profile analysis to compare the grouping of classroom observations based on a more descriptive approach to that based on classroom observation networks. We investigated active learning method to understand the extent to which instructional practices vary within and across the named methods. We examined class size because previous research indicates that there are challenges to incorporating active learning strategies in larger classes; therefore, instructional practices may vary between small and large classes [78]. We included the last three variables because it is plausible that instructors at institutions that devote more resources to teaching than to research may be supported in using different types of instruction than those at research-focused institutions.

For each of these attributes, we calculated the segregation Z score, established in Ref. [76], of the clustered similarity network. The authors of Ref. [76] define a network-level segregation measure as follows:

$$D_{\text{seg}}(M) = \frac{1}{N} \sum_{k \in M} N_k \left(\sum_{t=1}^s p_{kt} \log_2 \left(\frac{p_{kt}}{q_t} \right) \right),$$

where M is the clustering of the network (in our case, from the Infomap algorithm), N is the number of nodes in the network, s is the number of different categories or levels of the attribute of interest, $p_{kt} = N_{kt}/N_k$ is the probability of choosing a node at random from cluster k that has attribute t , and $q_t = N_t/N$ is the probability of choosing a node at random from the entire network with attribute t . The segregation Z score compares the above segregation measure for our similarity network to that for a large sample of networks r (we used 10 000 networks as in Ref. [76]) where the attributes are randomly distributed across the nodes while keeping the network structure and clustering the same:

$$Z = \frac{D_{\text{seg}}(M) - \langle D_{\text{seg}}^r(M) \rangle}{\sigma_r}.$$

If $|Z| > 1.96$, then the segregation of our network is significantly different from random. For the variables for which we found significant segregation, we examined the distributions of that attribute across the clusters to better understand the nature of the segregation (i.e., which categories of the attribute were more likely to be assigned to certain clusters).

8. Comparing student learning across instruction types

To address the third research question, we first calculated an individual effect size—a standardized mean difference between students' presemester and postsemester concept inventory scores—for each individual course using Hedges' g [79]. Hedges' g is a form of Cohen's d [80] that accounts for possible biases due to small sample sizes (e.g., small-enrollment physics courses). Prior studies recommend using Hedges' g for all courses in a study, even if only a few of them have small sample sizes [81].

As in meta-analytic studies [3], we then combined the individual effect sizes to calculate a net effect size for each cluster (i.e., instruction type). We used a random effects model, which assumes the presence of both within-cluster and between-cluster variation. The weight of each individual effect size, i , is calculated as

$$w_i = \frac{1}{v_i + \tau^2},$$

where v_i is the within-cluster variance and τ^2 is the between-cluster variance. The net effect size for each cluster is calculated by taking the sum of each individual effect size multiplied by its weight and dividing that sum by the sum of all the weights:

$$g_{\text{net}} = \frac{\sum_{i=1}^k w_i g_i}{\sum_{i=1}^k w_i},$$

where k is the number of courses in the cluster and g_i is the individual effect size for course i . The 95% confidence interval of each net effect size is given by $g_{\text{net}} \pm (1.96 \times s_{g_{\text{net}}})$, where $s_{g_{\text{net}}}$ is the standard deviation of the net effect size:

$$s_{g_{\text{net}}} = \sqrt{\frac{1}{\sum_{i=1}^k w_i}}.$$

We determined whether the net effect sizes significantly differ between the clusters by conducting a test of heterogeneity. This test quantifies whether the variation between the net effect sizes is greater than chance (i.e., random noise). We calculated the Q statistic:

$$Q = \sum_{j=1}^m w_j \times (g_{\text{net},j} - \bar{g}_{\text{net}})^2,$$

where m is the number of clusters, w_j is the sum of weights within cluster j , $g_{\text{net},j}$ is the net effect size for cluster j , and $\bar{g}_{\text{net}}$ is the mean net effect size across all clusters. The degrees of freedom of Q is equal to one less than the number of clusters. The statistical significance of Q , when compared to a chi-squared distribution, indicates whether variation in net effect sizes is due to random chance alone (if Q is not significant) or instead due to some other source of variation (e.g., different types of active learning instruction, if Q is significant).

IV. RESULTS

In this section, we present the results by research question: the typology of instructional practices used in named active learning methods, the relationships between instruction type and course and institution characteristics, and the relationship between instruction type and student conceptual learning.

A. Typology of active learning

The Infomap community detection algorithm identifies five clusters in the sparsified similarity network, each containing five to seven instructors [i.e., classroom observation networks; Fig. 2(a)]. Below, we use the average node-level and network-level measures [Fig. 2(b) and Table VI] for classroom observation networks assigned to each cluster to characterize the type of instruction the cluster represents. These five instruction types comprise our typology of active learning.

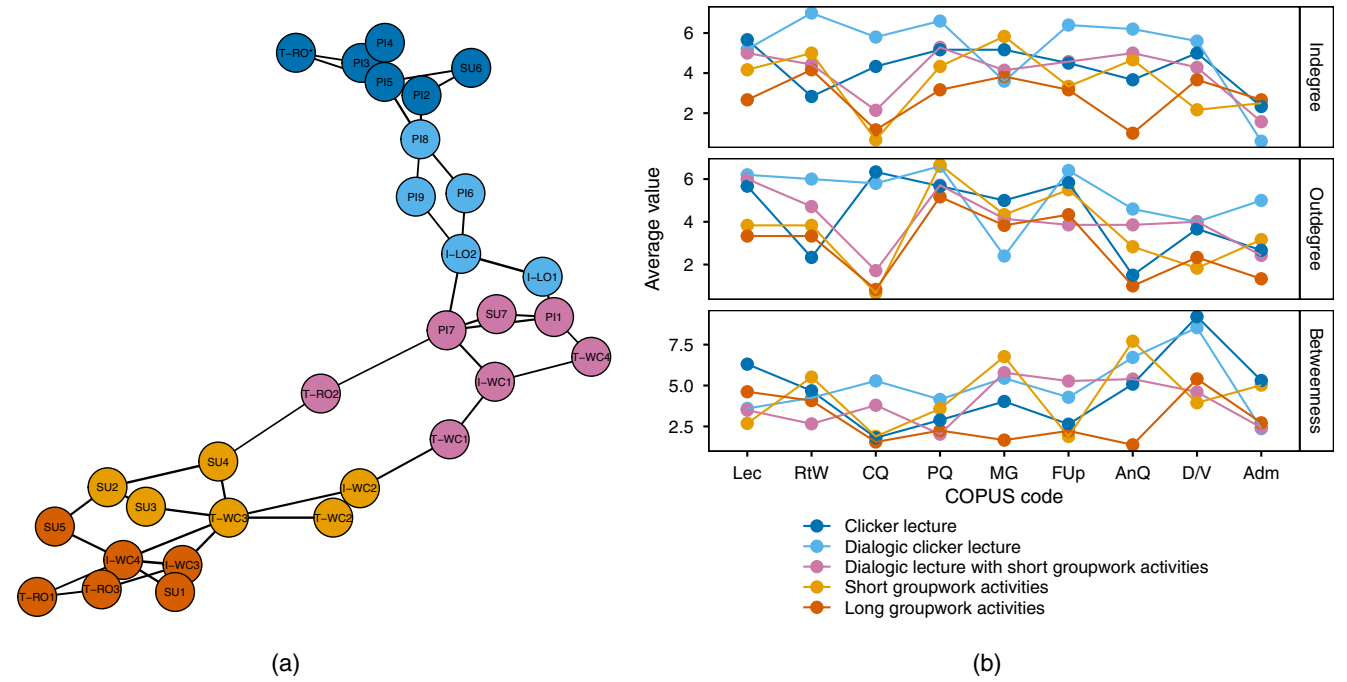


FIG. 2. (a) Sparsified similarity network with nodes (circles) representing classroom observation networks for each instructor, colored by cluster (i.e., instruction type). Edges (lines connecting the nodes) represent similarity, with edge weights (line thicknesses) proportional to cosine similarity. I-WC indicates whole-class implementation of ISLE, I-LO indicates lab-only implementation of ISLE (i.e., the observation networks represent lecture sections), PI indicates Peer Instruction, T-WC indicates whole-class implementation of Tutorials, T-RO indicates recitation-only implementation of Tutorials (T-RO* indicates recitation-only implementation of Tutorials but with the observation network representing the lecture section of the course), and SU indicates SCALE-UP. The attributes of each course are provided in Table IX in Sec. A 6 in the Appendix. (b) Average node-level measures for observation networks assigned to each instruction type (see definitions in Table V and values in Table VII in Sec. A 4 in the Appendix).

1. Clicker lecture

The first type of active learning instruction—clicker lecture—consists of four Peer Instruction courses, one SCALE-UP course, and one Tutorials course [recitation-only implementation, but with the classroom observation coming from the lecture section; see dark blue nodes at the top of Fig. 2(a)]. Instructors whose classroom observation networks are assigned to this instruction type frequently cycle through lecturing, asking and guiding groups through a clicker question, and following up on that clicker question with a demonstration or video (relatively high indegree and outdegree for all of these codes and high betweenness for lecturing and demonstrations/videos, see Fig. 2(b); relatively high edges, total strength, mutuality, and transitivity,

signifying a more “complex” instructional style that involves many transitions between different activities, see Table VI; see Fig. 3(a) for an example classroom observation network of this instruction type). In this instruction type, follow-up discussions of clicker questions tend to involve quick shout-out responses from all students to questions posed by the instructor [e.g., “So the answer is?” “C!”; relatively high indegree and outdegree for posing questions, see Fig. 2(b)].

2. Dialogic clicker lecture

Three Peer Instruction courses and two ISLE courses (both lab-only implementations, recall that these instructors recorded the lecture sections) are assigned to the second

TABLE VI. Average network-level measures for classroom observation networks assigned to each instruction type (see definitions in Table V).

Instruction type	Edges	Total strength	Mutuality	Transitivity
Clicker lecture	38.7	1.02	0.72	0.83
Dialogic clicker lecture	47.0	1.24	0.75	0.85
Dialogic lecture with short groupwork activities	36.4	0.71	0.72	0.82
Short groupwork activities	32.7	0.59	0.62	0.82
Long groupwork activities	25.5	0.47	0.52	0.64

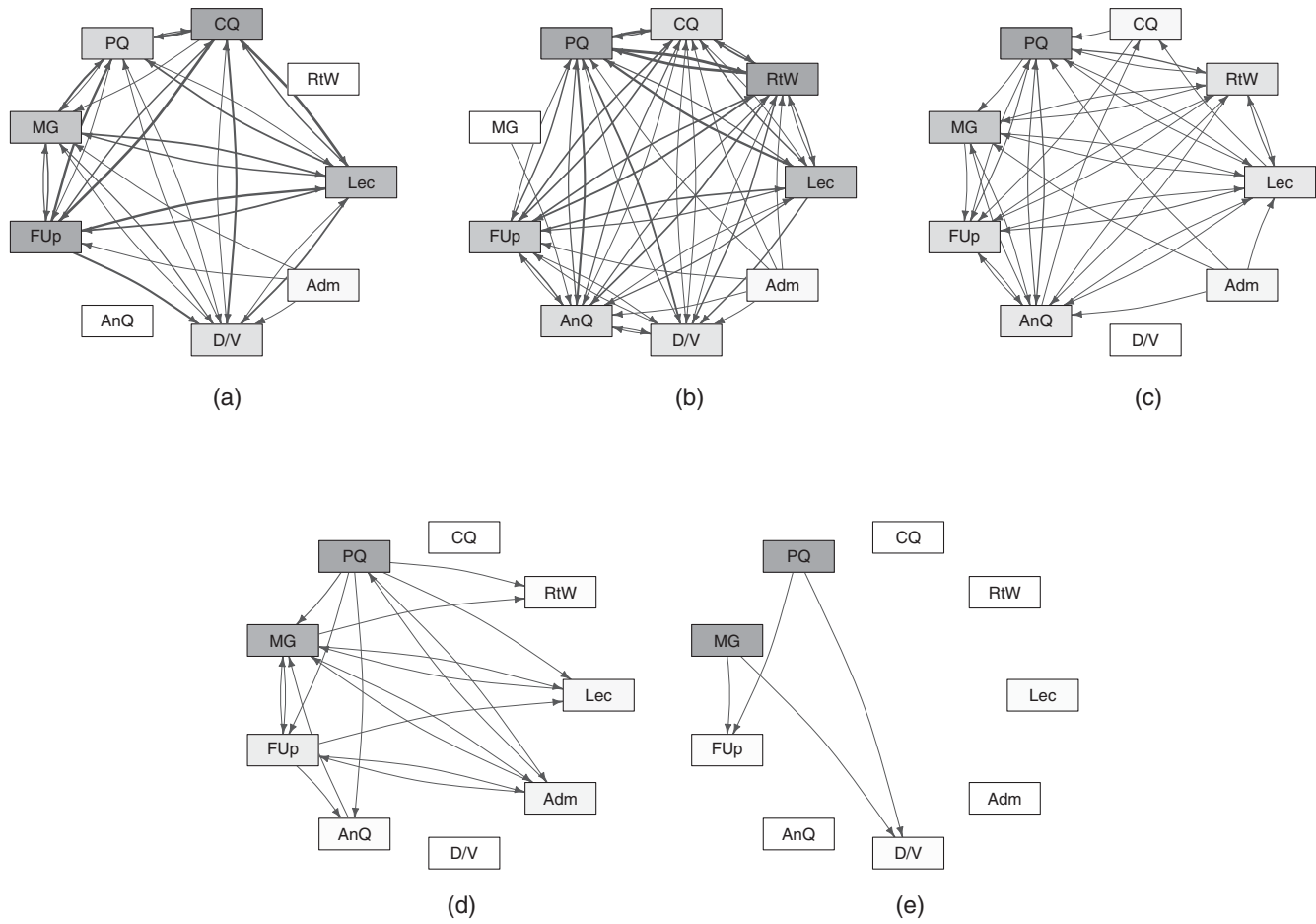


FIG. 3. Example classroom observation networks for each instruction type. The active learning method and course label [as in Fig. 2(a) and Table IX in Sec. A 6 in the Appendix] for each network are as follows: (a) Clicker lecture (Peer Instruction, PI3), (b) Dialogic clicker lecture (ISLE lab-only implementation, I-LO2), (c) Dialogic lecture with short groupwork activities (Tutorials whole-class implementation, T-WC1), (d) Short groupwork activities (SCALE-UP, SU3), and (e) Long groupwork activities (Tutorials recitation-only implementation, T-RO1). Though not directly part of our analysis, node color indicates the fraction of all 2-min time intervals that the code was present; darker shades indicate larger fractions (i.e., longer time durations). Edges (lines with arrows) point from an initial code to the code that occurs in the following 2-min time interval of the COPUS observation. Edge weight (line thickness) is proportional to the number of transitions occurring between those two codes normalized by the total number of 2-min time intervals observed. All 30 classroom observation networks are available at Ref. [54].

instruction type: dialogic clicker lecture [see light blue nodes toward the top of Fig. 2(a)]. Similar to the first instruction type, instructors whose classroom observation networks are assigned to this instruction type frequently cycle through lecturing, asking a clicker question, and following up on that clicker question with a demonstration or video and/or real-time writing (relatively high indegree and outdegree for all of these codes and high betweenness for clicker questions, posing questions, following up, and demonstrations/videos; see Fig. 2(b); highest edges, total strength, mutuality, and transitivity of all five instruction types, indicating the most transitions between activities, see Table VI; see Fig. 3(b) for an example classroom observation network of this instruction type). Unlike the first instruction type, the lecture and follow-up discussions involve dialogue between the instructor and individual

students: the instructor poses questions to which individual students raise their hands and respond and/or individual students ask questions to which the instructor responds [high indegree, outdegree, and betweenness for posing questions and answering questions; see Fig. 2(b)].

3. Dialogic lecture with short groupwork activities

The third instruction type—dialogic lecture with short groupwork activities—is the only type that contains classroom observation networks from all four active learning methods: two Peer Instruction courses, one SCALE-UP course, one ISLE course (whole-class implementation), and three Tutorials courses [two whole-class implementations and one recitation-only implementation, see pink nodes in the middle of Fig. 2(a)]. Instructors of this type cycle

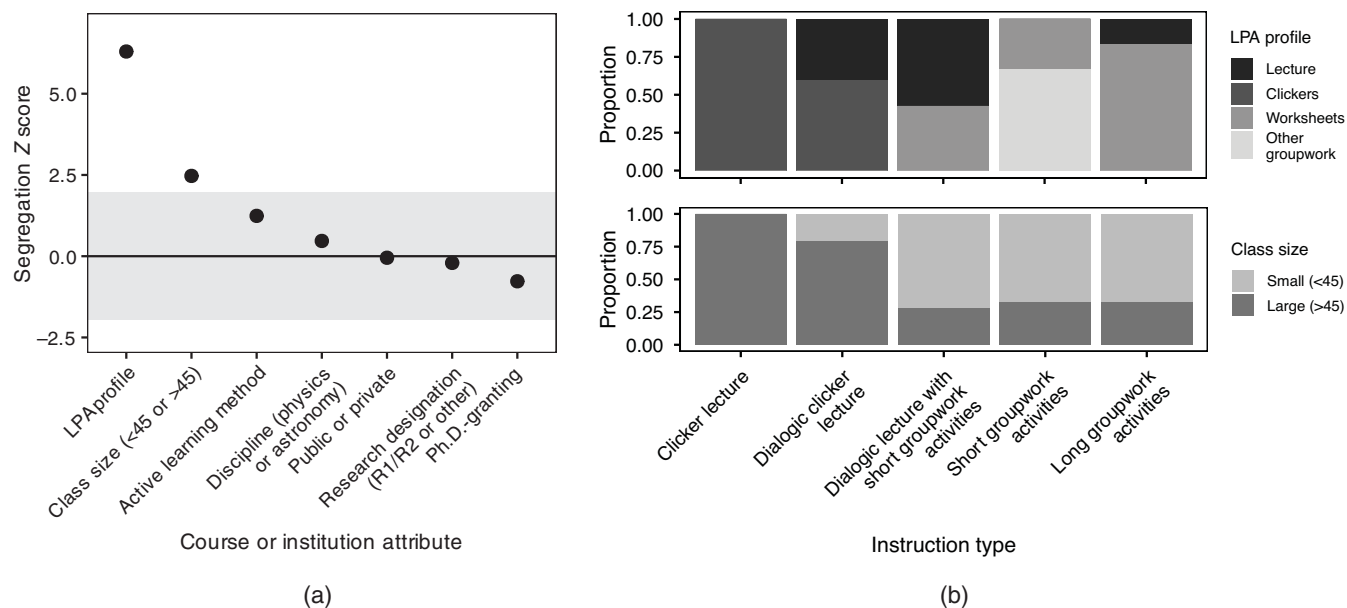


FIG. 4. (a) Segregation Z scores for various course and institution attributes. Points indicate the Z scores and the gray shaded area indicates where the Z score is not statistically significant, $|Z| < 1.96$. Points falling outside of this region indicate that there is a significant relationship between the corresponding attribute and instruction type. (b) For the two attributes with significant segregation Z scores, LPA profile and class size, the distributions of courses with these attributes across instruction types.

through lecturing, assigning and guiding various short groupwork activities (some of which are clicker questions as in the first two instruction types, and some of which are other activities such as worksheets or solving problems on whiteboards), and following up on those activities with demonstrations or videos (relatively high indegree and outdegree for lecturing, posing questions, and demonstrations/videos and relatively high betweenness for clicker questions, moving and guiding groupwork, and following up, see Fig. 2(b); moderate number of edges, total strength, mutuality, and transitivity, indicating a moderate amount of transitions between activities, see Table VI; see Fig. 3(c) for an example observation network of this instruction type). Similar to the previous instruction type, the lecture and follow-up discussions to groupwork activities are dialogic, with the instructor posing questions to which individual students respond and responding to questions asked by individual students [high indegree and outdegree for posing questions and answering questions, see Fig. 2(b)].

4. Short groupwork activities

The next instruction type—short groupwork activities—consists of three SCALE-UP courses, two Tutorials courses (both whole-class implementations), and one ISLE course [whole-class implementation, see light orange nodes toward the bottom of Fig. 2(a)]. Instructors of this type do not lecture as much as the previous three instruction types; instead, they cycle through assigning and guiding short groupwork activities (not including clicker questions) and following up on those activities with real-time writing

(low node-level measures for lecturing and clicker questions, high node-level measures for moving and guiding groupwork, high outdegree for following up, and high indegree and betweenness for real-time writing, see Fig. 2(b); lower network-level measures than the previous instruction types, see Table VI; see Fig. 3(d) for an example observation network of this instruction type). Similar to the previous two instruction types, the follow-up discussions to groupwork activities look like class-wide discussions, with the instructor posing and responding to questions [highest outdegree for posing questions and highest betweenness for answering questions, see Fig. 2(b)]. Unlike the first three instruction types, this type does not use many demonstrations or videos [relatively low node-level measures for demonstrations/videos, see Fig. 2(b)].

5. Long groupwork activities

The final instruction type—long groupwork activities—contains two SCALE-UP courses, two ISLE courses (both whole-class implementations), and two Tutorials courses [both recitation-only implementations, see dark orange nodes at the bottom of Fig. 2(a)]. Instructors whose classroom observations are assigned to this instruction type predominantly guide one or a few long groupwork activities (e.g., worksheets) and then follow up on them using real-time writing and/or demonstrations or videos (node-level measures are low compared to the other instruction types for most COPUS codes, but comparable indegree and outdegree to other instruction types for moving and guiding groupwork and following up and

moderately high betweenness for real-time writing and demonstrations/videos, see Fig. 2(b); lowest number of edges, total strength, mutuality, and transitivity of all five instruction types, indicating fewer transitions between activities and longer activities, see Table VI; see Fig. 3(e) for an example classroom observation network of this instruction type). Follow-ups to the groupwork activities are more didactic than dialogic [low node-level measures for posing questions and answering questions, see Fig. 2(b)].

B. Relating instruction types to course and institution attributes

There is a significant relationship between instruction type and latent profile identified in our prior work ($Z = 6.35$) [30] and between instruction type and class size [whether smaller or larger than 45 enrolled students, $Z = 2.44$; see points falling above the gray highlighted area in Fig. 4(a)]. For the latent profiles, the first two instruction types (both clicker lectures) include all of the courses that were assigned to the clickers profile in our previous analysis [Fig. 4(b)]. The dialogic clicker lecture type also includes several of the courses assigned to the lecture profile in our LPA. Both the dialogic lecture with short groupwork activities and long groupwork activities types include courses that were assigned to the lecture and worksheets profile in our previous work. The short groupwork activities type includes several courses assigned to the worksheets profile and all of the courses assigned to the other groupwork profile. For class size, the first two instruction types, clicker lecture and dialogic clicker lecture, include most of the large classes while the remaining three instruction types include most of the small classes [Fig. 4(b)].

Instruction type is not significantly related to active learning method (whether ISLE, Peer Instruction, Tutorials, or SCALE-UP; $Z = 1.23$), discipline (whether physics or astronomy; $Z = 0.47$), whether the institution is public or private ($Z = -0.04$), institution research designation (whether R1/R2 or not according to the 2025 Carnegie Research Classification; $Z = -0.23$), or Ph.D.-granting status [$Z = -0.75$, see points falling within the gray shaded area, $|Z| < 1.96$, in Fig. 4(a)]. In other words, none of these course or institution attributes correlate with the instruction type used by instructors; instead, courses with the same attribute are spread across the instruction types fairly randomly. With regard to the active learning method, this suggests that there is a large amount of variation in the style of implementation of the four named active learning methods we study here, as courses from each method are distributed across different instruction types. Named active learning methods, therefore, may not necessarily map onto specific instructional practices: instructors likely adapt the methods to their own style of teaching and local context (e.g., student population and classroom constraints).

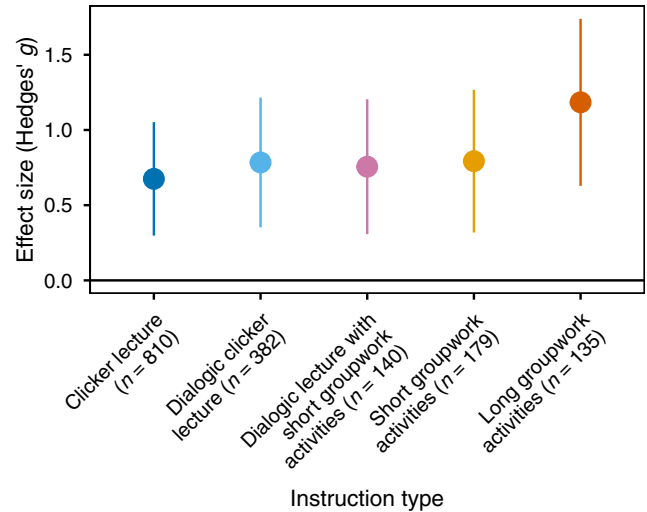


FIG. 5. Effect sizes for concept inventory scores by instruction type. Points represent Hedges' g values and error bars indicate 95% confidence intervals (see values in Table VIII in Sec. A 5 in the Appendix). Positive (negative) values indicate increases (decreases) in student scores from presemester to postsemester. N values indicate the number of students included in the analysis for each instruction type (i.e., the number of students with matched presemester and postsemester concept inventory scores).

C. Relating instruction types to student conceptual learning

A heterogeneity analysis indicates no statistically significant variation in student conceptual learning gains based on instruction type ($Q = 2.32$, $df = 4$, $p = 0.68$). Qualitatively, the effect size for the last instruction type—long groupwork activities—is slightly larger than the other four instruction types; however, the confidence intervals of Hedges' g for all instruction types overlap with one another (Fig. 5). This means that not only are established active learning methods implemented in different ways (i.e., we saw no significant segregation of active learning methods across instruction types), but also that certain implementations (as measured by classroom behaviors alone) may not necessarily be more effective than others for student conceptual learning.

V. DISCUSSION

We have employed a novel method for characterizing and comparing instructional practices based on classroom observation networks. Such networks allow researchers to visualize chronological transitions between classroom activities and directly measure similarity between pairs of observations. Applying this method to a dataset of 30 instructors implementing one of four named active learning methods (ISLE, Peer Instruction, Tutorials, and SCALE-UP) in their introductory physics or astronomy course, we identified five instruction types: clicker lecture, dialogic clicker lecture, dialogic lecture with short

groupwork activities, short groupwork activities, and long groupwork activities.

This typology of active learning instruction not only corroborates our previous latent profile analysis of these data [30], providing some validity to both groupings of observations, but also provides a far more detailed framework for describing instructional practices than existing literature. Smith *et al.*, for example, found that “most faculty members fall somewhere in the continuum between pure lecturing and and primarily active-engagement instruction” [25] (p. 634) based on descriptive statistics of classroom observations. The current study provides a more nuanced characterization of this continuum that includes information about specific sequences of activities used by instructors (e.g., short cycles of clicker questions, following up, and lecturing).

The analysis presented here also substantiates prior work highlighting that instructors do not adopt existing active learning methods as is; rather, they adapt them to their own teaching style and instructional context (e.g., student population and physical classroom space) [17–20]. We found that instruction type is not significantly related to named active learning method. Instead, instructors using each method were distributed across the identified instruction types. An alternative interpretation of this result is that named methods only appear to be unrelated to instruction types because some of the video observations did not capture the parts of the class in which the named methods were being implemented. Therefore, we suggest that future research further test this relationship between named methods and instruction types, such as by collecting video recordings of one full week of course contact hours, capturing all course components that a student would attend in a given week, or ensuring that recordings only capture the course components where the active learning method is being used. If our result holds in these studies, it would point to a shift in the way we discuss and study active learning: the names of developed active learning methods may not actually reflect the underlying activities (and the sequences of these activities) involved in the instruction. In this case, researchers should not rely on method labels as a basis for comparison (e.g., comparing student outcomes across methods); rather, they should empirically categorize instruction using in-class behaviors and/or other features.

Variables related to institution type (i.e., public or private, research designation, and Ph.D.-granting or not) are also not significantly related to instruction type despite existing research indicating that level of institutional support is often an important factor for instructors deciding whether or not to adopt active learning strategies [10–13]. Our measures of institutional support, however, were likely too broad (e.g., representative of the entire institution rather than the physics or astronomy department) to capture any meaningful relationships (if they exist). We suggest that

future work collect more detailed information from the instructor participants about support from their department to make more rigorous claims about the role of institutional support in instructor adaptations of active learning methods.

At the same time, we identified that one of the measured course and institution attributes—class size—does relate to instruction type. Large classes (containing more than 45 students) were more often assigned to one of the two clicker lecture types and small classes were more often assigned to one of the remaining three instruction types. A few of the large classes, however, were assigned to dialogic lecture with short groupwork activities, short groupwork activities, or long groupwork activities. This means that despite documented challenges to incorporating active learning strategies into larger courses [78], it is possible to do so *and* in a way that involves class-wide dialogue and/or a range of groupwork activities (i.e., not only clickers). Researchers should conduct controlled experiments to compare the implementations and impacts (e.g., on conceptual learning) of active learning methods in different class sizes.

We also noticed substantial variation *within* each instruction type, signaling that a broader typology of active learning instruction (e.g., with more than five types and/or a different set of types) likely exists. We recommend for future research studies to conduct a similar analysis with a broader range of active learning courses (i.e., not only those using named active learning methods) to build on our typology. Future work may also map instructors’ perceived barriers to adopting active learning strategies [10–13] and/or received training (e.g., whether they have previously attended the Physics and Astronomy Faculty Teaching Institute) to their instruction types (e.g., by also surveying or interviewing participating instructors). This line of work would offer two important implications. First, researchers could start to unpack some of the mechanisms (i.e., beyond class size) underlying different implementations of the same named active learning method, as observed here. It is plausible, for example, that instructors who received general training about active learning implement a given method differently from instructors who received training about the particular method (e.g., by attending a workshop). Second, in professional development settings, we could provide physics instructors with examples of what different implementations of each method look like and a guide to choosing a particular implementation based on their personal teaching philosophies, perceived barriers at their institutions, and other constraints (e.g., similar to Ref. [82]).

Additionally, we expand upon the few existing studies that relate specific instructional practices to student conceptual learning [28–33]. In our previous analysis of this dataset, we identified that while all four active learning methods exhibited conceptual learning gains larger than zero, SCALE-UP exhibited significantly larger conceptual

learning gains than either ISLE or Peer Instruction [30]. In the current study, however, we found no significant differences in student conceptual learning across the instruction types and that each method gets implemented in a variety of ways (e.g., SCALE-UP courses span four out of the five instruction types). Taken together, this suggests that instructors may tailor active learning methods to their own style without reducing the effectiveness of the methods for student conceptual learning. In other words, effective active learning likely does not require strict adherence to the method developers' intentions, at least with regard to classroom behaviors: instructors can be flexible in adapting the structures of the methods.

We suggest that future research investigate additional explanations for why SCALE-UP may be more effective for student learning than other methods [30], as this study suggests that specific classroom activities likely do not provide a complete explanation. Other possible explanations include the quality or fidelity of classroom activities (e.g., the SCALE-UP courses in our dataset may have been higher fidelity implementations of the method than the courses using the other three methods) and the nature of the methods in and of themselves (e.g., SCALE-UP provides one cohesive picture of the course material by blending lecture, problem solving, and lab activities).

A. Limitations and future work

We acknowledge several limitations to this work that prompt further research. Related to instructor recruitment, most instructors included in this study are physics education researchers themselves or fairly involved in the physics education research community. Yet research shows that science education researchers employing active learning methods in their classroom teach differently and exhibit more significant improvements to student outcomes than a random sample of instructors employing active learning methods [21]. Science education researchers likely have more pedagogical content knowledge and feel stronger self-efficacy with regard to teaching than other researchers, which may facilitate their skills and willingness to use active learning methods and adapt them to their local conditions. Future research, therefore, should aim to collect similar data from a broader sample of physics instructors, including information about the instructors' familiarity with physics education research.

We chose to use the COPUS for our analysis of the classroom observations. The instructor codes in the COPUS distinguish clicker questions from other forms of groupwork, but they do not tease apart the instructor assigning or proctoring other types of groupwork such as worksheets versus laboratory activities. The COPUS was also designed for descriptive, frequency-based analyses and not the temporal analysis conducted here [41]. Future research, therefore, may add more refined groupwork codes to the COPUS

and/or conduct a similar analysis using a different observation protocol that uses continuous (rather than interval-based) coding (e.g., the Framework for Interactive Learning in Lectures, FILL+ [26,83]). Researchers may also explore ways to overcome the methodological challenges we faced in including both instructor and student COPUS codes in a network-based analysis to gain a more holistic picture of interactive instruction.

In addition, our typology of active learning is defined by instructional behaviors seen in classroom video recordings. The typology does not capture specific content within the curriculum, such as the theory of learning underlying course content (e.g., elicit, confront, and resolve in Tutorials) or the extent to which the classroom activities (e.g., worksheet questions) are research-based, which may also impact student learning gains [84]. Future studies should aim to characterize *both* classroom behaviors and the nature of classroom activities and relate them to student learning.

Finally, we measured student conceptual learning as one outcome that may be impacted by different types of active learning instruction. We recommend that future work measure additional outcomes of interest, such as attitudes toward science, critical thinking skills, and sense of belonging, to understand the full scope of effects of instruction type on students.

VI. CONCLUSION

We have advanced previous descriptive studies of active learning instruction by capturing the temporal and interactional nature of instructional practices. We hope that the typology of active learning instruction defined here serves as a starting set of instruction types for future studies to modify and build upon. We encourage these future studies to incorporate data from a broader range of active learning strategies (i.e., not only named methods), instructors (i.e., not only those who are involved in science education research), courses (i.e., not only introductory level), scientific disciplines (e.g., biology, chemistry, and engineering), and institutions (e.g., outside of the United States). Further iterations of our typology have the potential to improve the professional development of college physics instructors by providing them with different ways of implementing existing active learning methods that have been used in a variety of instructional contexts.

ACKNOWLEDGMENTS

We thank all of the instructors and students who participated in our study. We also thank Ibukunoluwa Bukola for providing meaningful feedback on this manuscript. This material is based upon work supported by the National Science Foundation under Grants No. 2111128 and No. 2111275. M. S. is partly funded by the Cotswold Foundation Postdoctoral Fellowship at Drexel University.

DATA AVAILABILITY

The data that support the findings of this article are openly available [54].

APPENDIX

1. Comparing Jaccard index and cosine similarity

Figure 6 shows the Jaccard index and cosine similarity values for each pair of classroom observation networks. The Jaccard index measures specific edge agreement, or the amount of overlap between the edges present in two networks. For a pair of directed, weighted networks G_1 and G_2 with the same set of nodes V and respective sets of edges E_1 and E_2 and edge weights W_1 and W_2 , the Jaccard index is calculated as:

$$J(G_1, G_2) = \frac{\sum_{e \in E_1 \cap E_2} \min[W_1(e), W_2(e)]}{\sum_{e \in E_1 \cup E_2} \max[W_1(e), W_2(e)]}.$$

Values range from 0 to 1, with values closer to 1 indicating that the two networks have more similar sets of edges and corresponding edge weights.

As mentioned in the main text, the Jaccard index values are all relatively low (mostly light shades in the lower triangle). The cosine similarity values, however, span many of the intermediate values in the range of 0–1 (mix of light and dark shades in the upper triangle).

2. Choosing significance level for sparsifying the similarity network

Similar to Ref. [85], we identified an appropriate significance level, α , by applying the LANS algorithm at α values between 0 and 0.2 in steps of 0.0001 and then

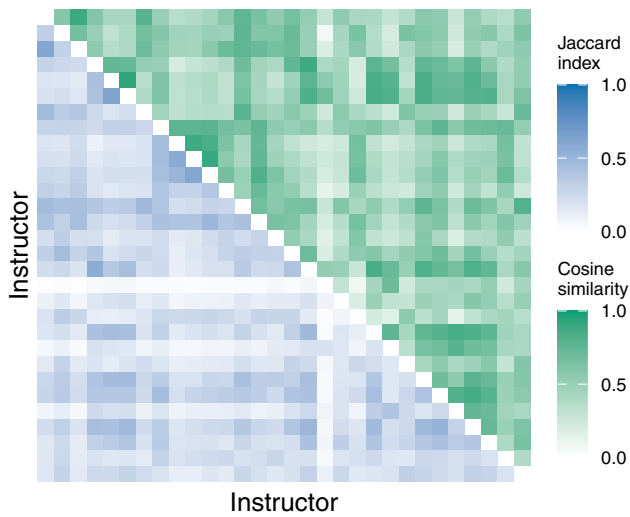


FIG. 6. Jaccard index (lower triangle) and cosine similarity (upper triangle) values for each pair of classroom observation networks.

running the Infomap community detection algorithm on the resulting sparsified similarity network. We calculated four measures for the Infomap clustering at each α value and chose an α value with suitable properties across all measures (Fig. 7).

The first measure, modularity (Q), captures the extent to which the sparsified similarity network can be partitioned into meaningful clusters where nodes share more edges with nodes in their cluster than they do with nodes in other

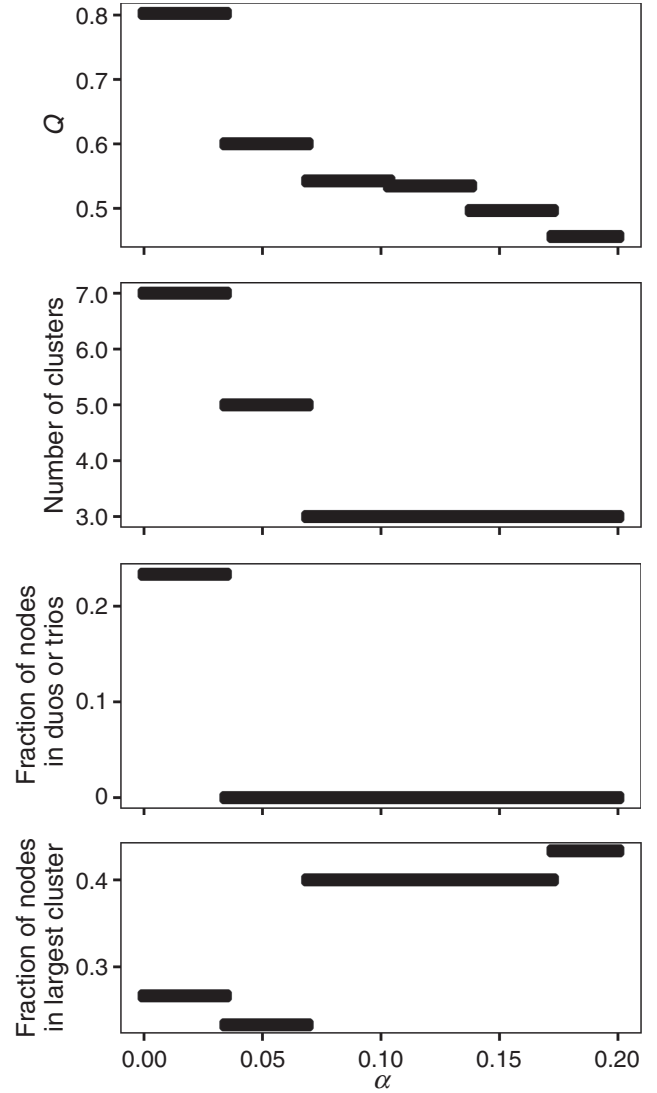


FIG. 7. Finding an appropriate significance level, α , for network sparsification. All measures were calculated using increments of 0.0001 in α . (a) Modularity, Q , of the similarity network when it is sparsified with different α values. (b) The number of clusters identified by the Infomap algorithm when the similarity network is sparsified with different α values. (c) The number of clusters identified by the Infomap algorithm that contain exactly two or three nodes when the similarity network is sparsified with different α values. (d) The fraction of nodes in the largest cluster identified by the Infomap algorithm when the similarity network is sparsified with different α values.

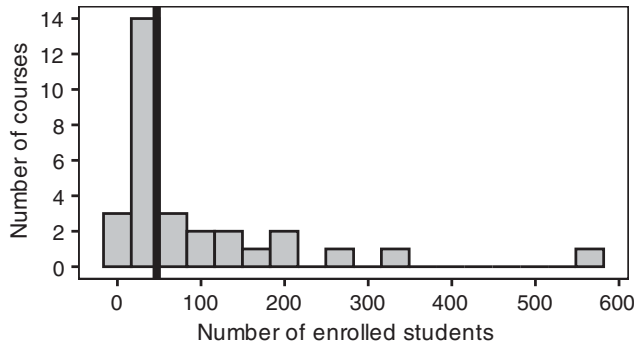


FIG. 8. Histogram of class sizes (i.e., number of enrolled students) for the 30 courses included in our study. The thick black line indicates the median of the distribution (48 students).

clusters. Q ranges from -0.5 to 1 and $Q > 0.3$ indicates significant modular structure [86]. For our similarity network, all α values had $Q > 0.3$ and $\alpha < 0.069$ had $Q > 0.6$ [Fig. 7(a)].

TABLE VII. Average node-level measures for classroom observation networks assigned to each instruction type [as shown in Fig. 2(b); see definitions in Table V].

Instruction type/code	Indegree	Outdegree	Betweenness
Clicker lecture			
Adm	2.33	2.67	5.31
AnQ	3.67	1.50	5.08
CQ	4.33	6.33	1.81
D/V	5.00	3.67	9.21
FUp	4.50	5.83	2.64
Lec	5.67	5.67	6.31
MG	5.17	5.00	4.03
PQ	5.17	5.67	2.89
RtW	2.83	2.33	4.68
Dialogic clicker lecture			
Adm	0.60	5.00	2.37
AnQ	6.20	4.60	6.72
CQ	5.80	5.80	5.28
D/V	5.60	4.00	8.53
FUp	6.40	6.40	4.28
Lec	5.20	6.20	3.60
MG	3.60	2.40	5.45
PQ	6.60	6.60	4.15
RtW	7.00	6.00	4.25
Dialogic lecture with short groupwork activities			
Adm	1.57	2.43	2.43
AnQ	5.00	3.86	5.40
CQ	2.14	1.71	3.80
D/V	4.29	4.00	4.60
FUp	4.57	3.86	5.27
Lec	5.00	6.00	3.50
MG	4.14	4.14	5.78
PQ	5.29	5.71	2.03
RtW	4.43	4.71	2.66

(Table continued)

TABLE VII. (Continued)

Instruction type/code	Indegree	Outdegree	Betweenness
Short groupwork activities			
Adm	2.50	3.17	5.03
AnQ	4.67	2.83	7.71
CQ	0.67	0.67	1.89
D/V	2.17	1.83	3.94
FUp	3.33	5.50	1.89
Lec	4.17	3.83	2.69
MG	5.83	4.33	6.77
PQ	4.33	6.67	3.58
RtW	5.00	3.83	5.51
Long groupwork activities			
Adm	2.67	1.33	2.72
AnQ	1.00	1.00	1.38
CQ	1.17	0.83	1.55
D/V	3.67	2.33	5.40
FUp	3.17	4.33	2.23
Lec	2.67	3.33	4.62
MG	3.83	3.83	1.67
PQ	3.17	5.17	2.25
RtW	4.17	3.33	4.08

The other three measures helped us to identify α values where the Infomap algorithm extracts multiple different clusters (i.e., not a solution that is one single cluster containing every node) and with multiple nodes in each cluster (i.e., not a solution where some clusters only contain two or three nodes and not a solution where one cluster is much larger than the other clusters). We calculated the number of clusters resulting from the Infomap algorithm at each α value [Fig. 7(b)], the fraction of nodes assigned to clusters containing only two or three nodes [Fig. 7(c)], and the fraction of nodes assigned to the largest cluster [Fig. 7(d)]. All α values yielded three or more clusters; however, $\alpha < 0.035$ resulted in some nodes being in clusters of size two or three, and $\alpha > 0.069$ resulted in

TABLE VIII. Effect sizes for concept inventory scores by instruction type (as shown in Fig. 5). N values indicate the number of students included in the analysis for each instruction type (i.e., the number of students with matched presemester and postsemester concept inventory scores).

Instruction type	Hedges' g [95% confidence interval]
Clicker lecture ($n = 810$)	0.68 [0.30, 1.05]
Dialogic clicker lecture ($n = 382$)	0.78 [0.35, 1.22]
Dialogic lecture with short groupwork activities ($n = 140$)	0.76 [0.31, 1.20]
Short groupwork activities ($n = 179$)	0.79 [0.32, 1.27]
Long groupwork activities ($n = 135$)	1.18 [0.63, 1.74]

TABLE IX. Course and institution attributes, using the same course labels as in Fig. 2(a). I-WC indicates whole-class implementation of ISLE, I-LO indicates lab-only implementation of ISLE (i.e., the observation networks represent lecture sections), PI indicates Peer Instruction, T-WC indicates whole-class implementation of Tutorials, T-RO indicates recitation-only implementation of Tutorials (T-RO* indicates recitation-only implementation of Tutorials, but with the observation network representing the lecture section of the course), and SU indicates SCALE-UP. SU6 was excluded from our segregation analysis of LPA profiles because each of the three classroom observations was assigned to different profiles, thus we could not assign the course to a profile where the majority of the observations were assigned. In the table, GW, groupwork; FCI, Force Concept Inventory [61]; Gender FCI, Gender Force Concept Inventory [87]; Half-FCI, Half-Length Force Concept Inventory [88]; EMCS, Energy and Momentum Conceptual Survey [89]; ECA, Energy Concept Assessment [90]; MBT, Mechanics Baseline Test [63]; FMCE, Force and Motion Concept Evaluation [62]; LSCI, Light and Spectroscopy Concept Inventory [91]; TOAST, Test of Astronomy Standards [92]; and C.I., confidence interval.

Course	LPA profile [30]	Instruction type in current study	Class size	Public/private	Ph.D.-granting	Discipline	Research designation	Concept inventory	Hedges' g [95% C.I.]
I-WC1	Lecture	Dialogic lecture with short GW activities	16 (small)	Private	No	Physics	Other	FCI	0.21 [−0.24, 0.65]
I-WC2	Worksheets	Short GW activities	15 (small)	Public	Yes	Physics	R1 or R2	FCI	0.89 [0.50, 1.29]
I-WC3	Worksheets	Long GW activities	32 (small)	Public	No	Physics	Other	FMCE	0.65 [0.29, 1.00]
I-WC4	Worksheets	Long GW activities	33 (small)	Public	No	Physics	Other	...	...
I-LO1	Lecture	Dialogic clicker lecture	150 (large)	Public	Yes	Physics	R1 or R2	FCI	0.85 [0.51, 1.20]
I-LO2	Lecture	Dialogic clicker lecture	118 (large)	Public	Yes	Physics	R1 or R2	FCI	0.40 [0.13, 0.68]
PI1	Lecture	Dialogic lecture with short GW activities	28 (small)	Public	No	Physics	Other	FCI	0.28 [−0.20, 0.76]
PI2	Clickers	Clicker lecture	253 (large)	Public	Yes	Physics	R1 or R2	FMCE	0.47 [0.28, 0.66]
PI3	Clickers	Clicker lecture	576 (large)	Private	Yes	Physics	R1 or R2	MBT	0.24 [0.14, 0.34]
PI4	Clickers	Clicker lecture	201 (large)	Private	Yes	Physics	R1 or R2	MBT	0.19 [0.02, 0.36]
PI5	Clickers	Clicker lecture	81 (large)	Private	Yes	Physics	R1 or R2	Half-FCI	0.41 [0.21, 0.61]
PI6	Clickers	Dialogic clicker lecture	94 (large)	Public	No	Physics	Other	FCI	1.02 [0.74, 1.30]
PI7	Lecture	Dialogic lecture with short GW activities	17 (small)	Private	Yes	Physics	Other	Gender FCI	1.38 [0.38, 2.38]
PI8	Clickers	Dialogic clicker lecture	320 (large)	Public	Yes	Physics	R1 or R2	FMCE	1.06 [0.89, 1.22]
PI9	Clickers	Dialogic clicker lecture	20 (small)	Private	No	Physics	Other	FCI	0.49 [0.10, 0.87]
T-WC1	Worksheets	Dialogic lecture with short GW activities	18 (small)	Public	Yes	Astronomy	R1 or R2	LSCI	0.84 [0.01, 1.66]
T-WC2	Other groupwork	Short GW activities	36 (small)	Private	Yes	Physics	Other	FMCE	0.49 [0.14, 0.83]
T-WC3	Worksheets	Short GW activities	29 (small)	Public	No	Physics	Other	FCI	0.28 [0.14, 0.42]
T-WC4	Lecture	Dialogic lecture with short GW activities	70 (large)	Private	Yes	Astronomy	R1 or R2	...	...
T-RO1	Worksheets	Long GW activities	32 (small)	Private	Yes	Physics	R1 or R2	ECA	2.19 [1.44, 2.94]
T-RO2	Worksheets	Dialogic lecture with short GW activities	49 (large)	Public	No	Physics	Other	FCI	0.75 [0.50, 1.01]
T-RO3	Worksheets	Long GW activities	11 (small)	Public	No	Physics	Other	FCI	1.01 [0.26, 1.76]
T-RO*	Clickers	Clicker lecture	185 (large)	Public	Yes	Astronomy	R1 or R2	TOAST	0.96 [0.65, 1.27]
SU1	Worksheets	Long GW activities	48 (large)	Public	Yes	Physics	R1 or R2	...	...
SU2	Other groupwork	Short GW activities	54 (large)	Private	Yes	Physics	R1 or R2	...	...
SU3	Other groupwork	Short GW activities	135 (large)	Public	No	Physics	Other	FCI	1.13 [0.83, 1.42]
SU4	Other groupwork	Short GW activities	23 (small)	Private	No	Physics	Other	FCI	1.10 [0.58, 1.63]
SU5	Lecture	Long GW activities	99 (large)	Public	Yes	Physics	R1 or R2	FCI	0.93 [0.66, 1.20]
SU6	...	Clicker lecture	48 (large)	Private	No	Physics	Other	FCI	2.03 [1.57, 2.50]
SU7	Worksheets	Dialogic lecture with short GW activities	42 (small)	Public	Yes	Physics	R1 or R2	EMCS	1.07 [0.66, 1.47]

at least 40% of the nodes being assigned to the largest cluster. Triangulating all of these measures, we decided to use $\alpha = 0.05$.

3. Distribution of class sizes

Figure 8 shows the distribution of class sizes for the 30 courses in our study. The median of the distribution is 48 students. We considered courses with more than 45 students to be large and courses with fewer than 45 students to be small because there were no courses with 43–47 students and this partition created roughly even numbers of courses in each group: 16 large classes and 14 small classes. This partition is also consistent with that of Freeman *et al.*'s meta-analysis, which uses a threshold of 50 students to distinguish between small- and medium-sized classes [3].

4. Node-level measures by instruction type

Table VII provides the values for the points shown in Fig. 2(b), indicating the mean centrality measures for each COPUS code for the classroom observation networks assigned to each instruction type.

5. Concept inventory effect sizes by instruction type

Table VIII provides the values for the points shown in Fig. 5, indicating the Hedges' g values for the courses assigned to each instruction type.

6. Course and institution attributes

Table IX provides the course and institutional attributes, instruction type, and individual concept inventory effect size for each course included in the study. These attributes were used in our segregation analysis [Fig. 4(a)].

-
- [1] Doug Lombardi, Thomas F. Shipley, Astronomy Team, Biology Team, Chemistry Team, Engineering Team, Geography Team, Geoscience Team, and Physics Team, The curious construct of active learning, *Psychol. Sci. Publ. Interest* **22**, 8 (2021).
 - [2] Richard 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).
 - [3] Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410 (2014).
 - [4] Elli J. Theobald, Mariah J. Hill, Elisa Tran, E. Nicole Arroyo Sweta Agrawal, Shawn Behling, Nyasha Chambwe, Dianne Laboy Cintrón, Jacob D. Cooper, Dunster Gideon *et al.*, Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6476 (2020).
 - [5] Joshua Von Korff, Benjamin Archibeque, K. Alison Gomez, Tyrel Heckendorf, Sarah B. McKagan, Eleanor C. Sayre, Edward W. Schenk, Chase Shepherd, and Lane Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, *Am. J. Phys.* **84**, 969 (2016).
 - [6] Eric Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
 - [7] Catherine H. Crouch, Jessica Watkins, Adam P. Fagen, and Eric Mazur, Peer Instruction: Engaging students one-on-one, all at once, in *Research-Based Reform of University Physics* (2007), Vol. 1, pp. 40–95, [10.1119/RevPERv1.1.3](#).
 - [8] Eugenia Etkina and Alan Van Heuvelen, Investigative science learning environment—A science process approach to learning physics, in *Research-Based Reform of University Physics* (2007), Vol. 1, pp. 1–48, [10.1119/RevPERv1.1.1](#).
 - [9] Eugenia Etkina, Millikan award lecture: Students of physics—listeners, observers, or collaborative participants in physics scientific practices?, *Am. J. Phys.* **83**, 669 (2015).
 - [10] Charles Henderson and Melissa H. Dancy, Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics, *Phys. Rev. ST Phys. Educ. Res.* **3**, 020102 (2007).
 - [11] Charles Henderson, Melissa Dancy, and Magdalena Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020104 (2012).
 - [12] Melissa H. Dancy and Charles Henderson, Experiences of new faculty implementing research-based instructional strategies, *AIP Conf. Proc.* **1413**, 163 (2012).
 - [13] Yessi Affriyenni, Helen Georgiou, and Noah Finkelstein, Navigating the adoption of research-based instructional strategies within the complex nature of higher education, *Phys. Rev. Phys. Educ. Res.* **21**, 020124 (2025).
 - [14] Kathleen T. Foote, Xaver Neumeyer, Charles Henderson, Melissa H. Dancy, and Robert J. Beichner, Diffusion of research-based instructional strategies: The case of SCALE-UP, *Int. J. STEM Educ.* **1**, 10 (2014).
 - [15] Charles Henderson and Melissa H. Dancy, Impact of physics education research on the teaching of introductory quantitative physics in the united states, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020107 (2009).
 - [16] Melissa Dancy, Charles Henderson, Naneh Apkarian, Estrella Johnson, Marilyne Stains, Jeffrey R. Raker, and Alexandra Lau, Physics instructors' knowledge and use of active learning has increased over the last decade but most

- still lecture too much, *Phys. Rev. Phys. Educ. Res.* **20**, 010119 (2024).
- [17] Maura Borrego, Stephanie Cutler, Michael Prince, Charles Henderson, and Jeffrey E. Froyd, Fidelity of implementation of research-based instructional strategies (RBIS) in engineering science courses, *J. Eng. Educ.* **102**, 394 (2013).
- [18] Melissa Dancy, Charles Henderson, and Chandra Turpen, How faculty learn about and implement research-based instructional strategies: The case of peer instruction, *Phys. Rev. Phys. Educ. Res.* **12**, 010110 (2016).
- [19] Erin Scanlon, Brian Zamarripa Roman, Elijah Ibadlit, and Jacquelyn J. Chini, A method for analyzing instructors' purposeful modifications to research-based instructional strategies, *Int. J. STEM Educ.* **6**, 12 (2019).
- [20] Linda E. Strubbe, Adrian M. Madsen, Sarah B. McKagan, and Eleanor C. Sayre, Beyond teaching methods: Highlighting physics faculty's strengths and agency, *Phys. Rev. Phys. Educ. Res.* **16**, 020105 (2020).
- [21] Tessa M. Andrews, Michael J. Leonard, Clinton A. Colgrove, and Steven T. Kalinowski, Active learning not associated with student learning in a random sample of college biology courses, *CBE Life Sci. Educ.* **10**, 394 (2011).
- [22] Madhavi Vishnubhotla, Ahsan Chowdhury, Naneh Apkarian, Estrella Johnson, Melissa Dancy, Charles Henderson, Alexandra Claire Lau, Jeffrey R. Raker, and Marilyne Stains, "I use IBL in this course" may say more about an instructor's beliefs than about their teaching, *Int. J. Res. Undergrad. Math. Educ.* **10**, 87 (2024).
- [23] Diane Ebert-May, Terry L. Derting, Janet Hodder, Jennifer L. Momsen, Tammy M. Long, and Sarah E. Jardeleza, What we say is not what we do: Effective evaluation of faculty professional development programs, *BioScience* **61**, 550 (2011).
- [24] Chandra Turpen and Noah D. Finkelstein, Not all interactive engagement is the same: Variations in physics professors' implementation of peer instruction, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020101 (2009).
- [25] Michelle K. Smith, Erin L. Vinson, Jeremy A. Smith, Justin D. Lewin, and MacKenzie R. Stetzer, A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions, *CBE Life Sci. Educ.* **13**, 624 (2014).
- [26] Anna K. Wood, Ross K. Galloway, Robyn Donnelly, and Judy Hardy, Characterizing interactive engagement activities in a flipped introductory physics class, *Phys. Rev. Phys. Educ. Res.* **12**, 010140 (2016).
- [27] Kelley Commeford, Eric Brewre, and Adrienne Traxler, Characterizing active learning environments in physics using network analysis and classroom observations, *Phys. Rev. Phys. Educ. Res.* **17**, 020136 (2021).
- [28] I. Bukola, M. Sundstrom, J. Gambrell, O. Ross, A. L. Traxler, and E. Brewre, Multi-institutional assessment of Peer Instruction implementation and impacts using the Framework for Interactive Learning in Lectures, presented at PER Conf. 2025, Washington, DC, [10.1119/perc.2025.pr.Bukola](https://doi.org/10.1119/perc.2025.pr.Bukola).
- [29] Laura K. Weir, Megan K. Barker, Lisa M. McDonnell, Natalie G. Schimpf, Tamara M. Rodela, and Patricia M. Schulte, Small changes, big gains: A curriculum-wide study of teaching practices and student learning in undergraduate biology, *PLoS One* **14**, e0220900 (2019).
- [30] M. Sundstrom, J. Gambrell, C. Green, A. Traxler, and E. Brewre, Relative benefits of different active learning methods to conceptual physics learning, *Nat. Phys.* (2026), [10.1038/s41567-026-03307-2](https://doi.org/10.1038/s41567-026-03307-2).
- [31] Lacy M. Cleveland, Jeffrey T. Olimpo, and Sue Ellen DeChenne-Peters, Investigating the relationship between instructors' use of active-learning strategies and students' conceptual understanding and affective changes in introductory biology: A comparison of two active-learning environments, *CBE Life Sci. Educ.* **16**, ar19 (2017).
- [32] Georgianne L. Connell, Deborah A. Donovan, and Timothy G. Chambers, Increasing the use of student-centered pedagogies from moderate to high improves student learning and attitudes about biology, *CBE Life Sci. Educ.* **15**, ar3 (2016).
- [33] Karen S. McNeal, Min Zhong, Nick A. Soltis, Lindsay Doukopoulos, Elijah T. Johnson, Stephanie Courtney, Akilah Alwan, and Mallory Porch, Biosensors show promise as a measure of student engagement in a large introductory biology course, *CBE Life Sci. Educ.* **19**, ar50 (2020).
- [34] Saira Anwar and Muhsin Menekse, A systematic review of observation protocols used in postsecondary STEM classrooms, *Rev. Educ.* **9**, 81 (2021).
- [35] Daiyo Sawada, Michael D. Piburn, Eugene Judson, Jeff Turley, Kathleen Falconer, Russell Benford, and Irene Bloom, Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol, *School Sci. Math.* **102**, 245 (2002).
- [36] Camille L. Wainwright, Lawrence Flick, and Patricia Morrell, The development of instruments for assessment of instructional practices in standards-based teaching, *J. Math. Sci.* **6**, 21 (2003), https://scholarscompass.vcu.edu/jmsce_vamsc/vol6/iss1/6.
- [37] Alene H. Harris and Monica Farmer Cox, Developing an observation system to capture instructional differences in engineering classrooms, *J. Eng. Educ.* **92**, 329 (2003).
- [38] Anne L. Kern, Tamara J. Moore, and F. Caglin Akillioglu, Cooperative learning: Developing an observation instrument for student interactions, in *Proceedings of the 37th Annual Frontiers in Education Conference—Global Engineering: Knowledge without Borders, Opportunities without Passports* (IEEE, New York, 2007), pp. T1D–1.
- [39] Matthew Tadashi Hora and Joseph J. Ferrare, Instructional systems of practice: A multidimensional analysis of math and science undergraduate course planning and classroom teaching, *J. Learn. Sci.* **22**, 212 (2013).
- [40] Matthew T. Hora, Toward a descriptive science of teaching: How the TDOP illuminates the multidimensional nature of active learning in postsecondary classrooms, *Sci. Educ.* **99**, 783 (2015).
- [41] Michelle K. Smith, Francis H. M. Jones, Sarah L. Gilbert, and Carl E. Wieman, The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices, *CBE Life Sci. Educ.* **12**, 618 (2013).
- [42] Aditi Kothiyal, Rwitajit Majumdar, Sahana Murthy, and Sridhar Iyer, Effect of think-pair-share in a large CS1 class:

- 83% sustained engagement, in *Proceedings of the 9th Annual International ACM Conference on International Computing Education Research* (Association for Computing Machinery, New York, NY, 2013), pp. 137–144.
- [43] Cynthia J. Finelli, Matthew DeMonbron, Prateek Shekhar, Maura Borrego, Charles Henderson, Michael Prince, and Cindy K. Waters, A classroom observation instrument to assess student response to active learning, in *Proceedings of the 2014 IEEE Frontiers in Education Conference (FIE)* (IEEE, New York, 2014), pp. 1–4.
- [44] Travis J. Lund, Matthew Pilarz, Jonathan B. Velasco, Devasmita Chakraverty, Kaitlyn Rosploch, Molly Undersander, and Marilyne Stains, The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice, *CBE Life Sci. Educ.* **14**, ar18 (2015).
- [45] Kameryn Denaro, Brian Sato, Ashley Harlow, Andrea Aebersold, and Mayank Verma, Comparison of cluster analysis methodologies for characterization of classroom observation protocol for undergraduate STEM (COPUS) data, *CBE Life Sci. Educ.* **20**, ar3 (2021).
- [46] Tong Wan, Ashley A. Geraets, Constance M. Doty, Erin K. H. Saitta, and Jacquelyn J. Chini, Characterizing science graduate teaching assistants' instructional practices in reformed laboratories and tutorials, *Int. J. STEM Educ.* **7**, 30 (2020).
- [47] Corbin M. Campbell, Alberto F. Cabrera, Jessica Ostrow Michel, and Shikha Patel, From comprehensive to singular: A latent class analysis of college teaching practices, *Res. High. Educ.* **58**, 581 (2017).
- [48] M. Stains *et al.*, Anatomy of STEM teaching in North American universities, *Science* **359**, 1468 (2018).
- [49] Kelley Commeford, Eric Brewe, and Adrienne Traxler, Characterizing active learning environments in physics using latent profile analysis, *Phys. Rev. Phys. Educ. Res.* **18**, 010113 (2022).
- [50] Timothy J. Weston, Sandra L. Laursen, and Charles N. Hayward, Measures of success: Characterizing teaching and teaching change with segmented and holistic observation data, *Int. J. STEM Educ.* **10**, 24 (2023).
- [51] Robert M. Erdmann and Marilyne Stains, Classroom as genome: Using the tools of genomics and bioinformatics to illuminate classroom observation data, *CBE Life Sci. Educ.* **18**, es1 (2019).
- [52] Matthew T. Hora and Joseph J. Ferrare, Remeasuring postsecondary teaching: How singular categories of instruction obscure the multiple dimensions of classroom practice, *J. Coll. Sci. Teach.* **43**, 36 (2014).
- [53] Jesper Bruun, Linda Udby, and Pia Jensen Ray, Network analyses of student interactions with online textbook problems, *Eur. J. Phys.* **46**, 025704 (2025).
- [54] Meagan Sundstrom, Comparing Active Learning Methods (2025), <https://github.com/msundstrom33/ComparingActiveLearningMethods>.
- [55] Lillian C. McDermott, *Tutorials in Introductory Physics* (Prentice Hall, Upper Saddle River, NJ, 2002).
- [56] J. P. Adams, E. E. Prather, and T. F. Slater, *Lecture-Tutorials for Introductory Astronomy* (Prentice Hall, Upper Saddle River, NJ, 2005).
- [57] Robert J. Beichner, Jeffery M. Saul, David S. Abbott, Jeanne J. Morse, Duane Deardorff, Rhett J. Allain, Scott W. Bonham, Melissa H. Dancy, and John S. Risley, The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project, in *Research-Based Reform of University Physics* (2007), Vol. 1, pp. 2–39, [10.1119/RevPERv1.1.4](https://doi.org/10.1119/RevPERv1.1.4).
- [58] Eric Brewe, Modeling theory applied: Modeling instruction in introductory physics, *Am. J. Phys.* **76**, 1155 (2008).
- [59] Patricia Heller, Ronald Keith, and Scott Anderson, Teaching problem solving through cooperative grouping, *Am. J. Phys.* **60**, 627 (1992).
- [60] Patricia Heller and Mark Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, *Am. J. Phys.* **60**, 637 (1992).
- [61] David Hestenes, Malcolm Wells, and Gregg Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
- [62] Susan Ramlo, Validity and reliability of the force and motion conceptual evaluation, *Am. J. Phys.* **76**, 882 (2008).
- [63] David Hestenes and Malcolm Wells, A mechanics baseline test, *Phys. Teach.* **30**, 159 (1992).
- [64] Jayson Nissen, Robin Donatello, and Ben Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, *Phys. Rev. Phys. Educ. Res.* **15**, 020106 (2019).
- [65] J. Richard Landis and Gary G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33**, 159 (1977).
- [66] Robert P. Dalka and Justyna P. Zwolak, Network analysis of graduate program support structures through experiences of various demographic groups, *Phys. Rev. Phys. Educ. Res.* **20**, 020106 (2024).
- [67] G. P. Lepage Ryan Tapping and N. Holmes, Visualizing patterns in CSEM responses to assess student conceptual understanding, *presented at PER Conf. 2018, Washington, DC*, pp. 419–422, [10.1119/perc.2018.pr.Tapping](https://doi.org/10.1119/perc.2018.pr.Tapping).
- [68] Gerd Kortemeyer, Extending item response theory to online homework, *Phys. Rev. ST Phys. Educ. Res.* **10**, 010118 (2014).
- [69] Tor Ole B. Odden, Alessandro Marin, and Marcos D. Caballero, Thematic analysis of 18 years of physics education research conference proceedings using natural language processing, *Phys. Rev. Phys. Educ. Res.* **16**, 010142 (2020).
- [70] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi, Comparing methods for comparing networks, *Sci. Rep.* **9**, 17557 (2019).
- [71] Laura Steinert and H. Ulrich Hoppe, A comparative analysis of network-based similarity measures for scientific paper recommendations, in *Proceedings of the 3rd European Network Intelligence Conference (ENIC)*, 2016 (IEEE, New York, 2016), pp. 17–24.
- [72] Cole Walsh, Daniyar Kushaliev, and Natasha G. Holmes, Connecting the dots: Student social networks in introductory physics labs, *presented at PER Conf. 2020*, pp. 557–562, [10.1119/perc.2020.pr.Walsh](https://doi.org/10.1119/perc.2020.pr.Walsh).
- [73] Meagan Sundstrom, David G. Wu, Cole Walsh, Ashley B. Heim, and N. G. Holmes, Examining the effects of lab instruction and gender composition on intergroup interaction networks in introductory physics labs, *Phys. Rev. Phys. Educ. Res.* **18**, 010102 (2022).

- [74] Nicholas J. Foti, James M. Hughes, and Daniel N. Rockmore, Nonparametric sparsification of complex multi-scale networks, *PLoS One* **6**, e16431 (2011).
- [75] Martin Rosvall and Carl T. Bergstrom, Maps of random walks on complex networks reveal community structure, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1118 (2008).
- [76] Jesper Bruun and Ian G. Bearden, Time development in the early history of social networks: Link stabilization, group dynamics, and segregation, *PLoS One* **9**, e112775 (2014).
- [77] Sonsoles López-Pernas, Santtu Tikka, and Mohammed Saqr, Mining patterns and clusters with transition network analysis: A heterogeneity approach, in *Advanced Learning Analytics Methods: AI, Precision and Complexity* (Springer Nature, Cham, Switzerland, 2025).
- [78] Brock Murdoch and Paul W. Guy, Active learning in small and large classes, *Account. Educ.* **11**, 271 (2002).
- [79] Jessica Gurevitch and Larry V. Hedges, Statistical issues in ecological meta-analyses, *Ecology* **80**, 1142 (1999).
- [80] Jacob Cohen, Quantitative methods in psychology: A power primer, *Psychol. Bull.* **112**, 155 (1992).
- [81] Herbert M. Turner III and M. Bernard Robert, Calculating and synthesizing effect sizes, *Contemp. Issues Commun. Sci. Disord.* **33**, 42 (2006).
- [82] Rachel E. Scherr and Andrew Elby, Enabling informed adaptation of reformed instructional materials, *AIP Conf. Proc.* **883**, 46 (2007).
- [83] George Kinnear, Steph Smith, Ross Anderson, Thomas Gant, Jill R. D. MacKay, Pamela Docherty, Susan Rhind, and Ross Galloway, Developing the FILL+ tool to reliably classify classroom practices using lecture recordings, *J. STEM Educ. Res.* **4**, 194 (2021).
- [84] Karen Cummings, Jeffrey Marx, Ronald Thornton, and Dennis Kuhl, Evaluating innovation in studio physics, *Am. J. Phys.* **67**, S38 (1999).
- [85] Cedric Linder, Jesper Bruun, Arvid Pohl, and Burkhard Priemer, Relationship between semiotic representations and student performance in the context of refraction, *Phys. Rev. Phys. Educ. Res.* **20**, 010103 (2024).
- [86] M. E. J. Newman Aaron Clauset and Cristopher Moore, Finding community structure in very large networks, *Phys. Rev. E* **70**, 066111 (2004).
- [87] Laura McCullough and David Meltzer, Differences in male/female response patterns on alternative-format versions of the Force Concept Inventory, *presented in PER Conf. 2001, Rochester, New York*, 10.1119/perc.2001.pr.013.
- [88] Jing Han, Kathleen Koenig, Lili Cui, Joseph Fritchman, Dan Li, Wanyi Sun, Zhao Fu, and Lei Bao, Experimental validation of the half-length Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **12**, 020122 (2016).
- [89] Chandralekha Singh and David Rosengrant, Multiple-choice test of energy and momentum concepts, *Am. J. Phys.* **71**, 607 (2003).
- [90] Lin Ding, Designing an energy assessment to evaluate student understanding of energy topics, Ph.D. thesis, North Carolina State University, 2007.
- [91] Erin M. Bardar, Edward E. Prather, Kenneth Brecher, and Timothy F. Slater, Development and validation of the light and spectroscopy concept inventory, *Astron. Educ. Rev.* **5**, 103 (2007).
- [92] Stephanie J. Slater, The development and validation of the test of astronomy standards (TOAST), *J. Astron. Earth Sci. Educ.* **1**, 1 (2014).