

Measuring fidelity of implementation of named active learning methods in physics

Ibukunoluwa Bukola^{1,*}, Meagan Sundstrom¹, 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 19 December 2025; accepted 4 August 2026; published 2 October 2026)

Various active learning methods have been developed for introductory physics, and these methods are increasingly being adopted by instructors. However, instructors often do not implement these methods exactly as was originally intended by the developers, as they may face issues related to funding and institutional support for active learning and/or have different instructional contexts (e.g., student populations) and environments (e.g., physical classroom layouts) than the developers. Existing research does not sufficiently capture the range of variation in instructor implementation of established active learning methods, especially in comparison with high-fidelity implementations. In this study, we first identify the critical components (i.e., components without which the active learning method cannot be said to have been implemented) of three named active learning methods: Student-Centered Active Learning Environment with Upside-Down Pedagogies (SCALE-UP), Investigative Science Learning Environment (ISLE), and Tutorials. We then evaluate the fidelity with which 18 different introductory physics instructors implement these methods by analyzing classroom observations and comparing the extent to which these broader implementations use each critical component in their classroom to high-fidelity implementations. We find across all three active learning methods that broader implementations and high-fidelity implementations spend similar amounts of class time on critical components. At the same time, we observe substantial variation in the sequences of activities through which broader implementers operationalize these critical components (e.g., doing a few long activities versus many short activities). Finally, we find no clear relationship between fidelity of implementation (as measured by student and instructor behaviors during class) and student conceptual learning gains for our study's sample of instructors.

DOI: [10.1103/xcdt-txzh](https://doi.org/10.1103/xcdt-txzh)

I. INTRODUCTION

The benefits of active learning to student outcomes, such as conceptual learning [1,2] and persistence [3,4], are well-documented in the literature. Research also shows that over the past decade, the number of college-level physics instructors using research-based instructional strategies, including named active learning methods (e.g., Peer Instruction [5]), has significantly increased [6]. As these named methods become more widely implemented, instructors are modifying the methods based on their local

contexts, needs, and/or constraints [7]. Instructors, for example, may not receive appropriate funding [8] and/or institutional support [7,9,10] to facilitate their use of active learning. They may have different student populations, class sizes, and physical classrooms than the original developers [9,10]. Individual instructors also have their own teaching values and some amount of agency around their instructional choices [11].

Research has demonstrated that this flexibility of active learning methods to individual instructor modifications may help increase adoption and lead to improvements of the methods [12]. However, some modifications may depart too far from the established best practices [13]. Instructors, for example, may unknowingly omit *critical components* of the method, or components without which the method cannot be said to have been implemented [14]. Consequently, many existing studies on the effectiveness of research-based instructional strategies are met with concerns about the *fidelity* with which the instructors in the

*Contact author: ib396@drexel.edu

†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.

studies implemented the active learning methods they use, i.e., the extent to which instructors implemented the method in the way that the original developers intended [9,15,16]. Indeed, one study found that active learning does not improve student conceptual understanding more than traditional lecturing among a random set of instructors; rather, active learning may be more effective when implemented by science education researchers [17]. Together, these concerns and existing studies point to a growing need to better understand how instructors implement active learning methods and the effects of fidelity of implementation on student outcomes, such as conceptual learning.

Several previous studies have measured fidelity of implementation in undergraduate science courses [18–20]. Borrego and colleagues [19], for example, collected surveys from 387 engineering science faculty who used different research-based teaching practices. The authors measured fidelity as whether or not each instructor spent class time on each critical component of the research-based practices they implemented. The researchers found a wide range in fidelity across the examined active learning methods, spanning 11% to 80% of instructors implementing all critical components, and that methods with only one critical component exhibited higher fidelity than methods with multiple critical components.

The above studies, however, rely on survey responses from or interviews with instructors, which are prone to bias and may not accurately represent what instructors are actually doing in their classrooms [18]. Furthermore, to our knowledge, there are no studies that directly relate fidelity of implementation to student learning in the context of undergraduate science. The current study, therefore, uses direct classroom observations to characterize fidelity, leveraging the Classroom Observation Protocol for Undergraduate STEM (COPUS) [21] to accurately measure instructional practices. We measure the fidelity with which 18 introductory physics instructors (hereon referred to as “broader” implementations) implement three named active learning methods: Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) or Studio Physics, Investigative Science Learning Environment (ISLE), and Tutorials. We quantify fidelity in two ways: descriptively, comparing the fractions of class time spent on each activity in the COPUS to those of a high-fidelity implementation (including a focus on the COPUS activities that directly reflect each method’s critical components; we define high-fidelity implementations to be those at the site where the method was developed or based on the recommendation of experts, as in Ref. [22]), and with network analysis, comparing the temporal transitions between COPUS activities to those of a high-fidelity implementation [23]. We also directly relate these measures of fidelity to student understanding using concept inventory data.

We aim to address the following research questions:

1. How does the fraction of class time spent on critical components vary between high-fidelity and broader implementations of named active learning methods, as measured by relevant COPUS activities?
2. To what extent does a broad set of instructors implement named active learning methods with fidelity, based on full COPUS observations?
3. How, if at all, is fidelity of implementation (as measured by student and instructor behaviors during class using the COPUS) related to student conceptual learning?

II. BACKGROUND

In this section, we summarize existing literature about fidelity of implementation, critical components, and our analytic methods.

A. Fidelity of implementation

Fidelity of implementation has been extensively studied and defined in various fields, but only more recently in the Discipline-Based Education Research (DBER) community [24]. Broadly, fidelity of implementation is defined as a measure of closeness of an implemented intervention to the original intervention [25]. In the DBER context, Stains and colleagues define fidelity of implementation as “the extent to which the critical components of an intended educational program, curriculum, or instructional practice are present when that program, curriculum, or practice is enacted” [24] (p. 2). We operationalize this definition in the current study, where we consider active learning methods to be the relevant intervention.

As Stains and colleagues mention, identifying the critical components of an intervention (i.e., components without which the intervention cannot have been said to be carried out) is central to fidelity studies [14,16,19]. There are two types of critical components: structural and process (or instructional, as in Ref. [24]; Fig. 1). Structural

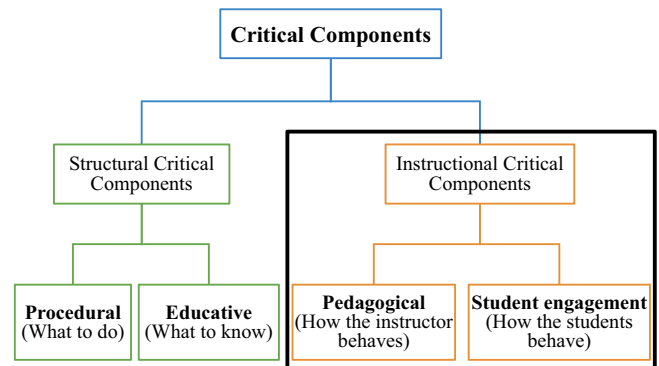


FIG. 1. Types of critical components as defined by Stains and colleagues [24]. The outlined black box highlights the focus of this study: instructional critical components.

critical components are related to program adherence (e.g., material covered, duration of intervention, and what knowledge the implementer must possess) [19,26]. Process or instructional critical components are related to how the intervention is implemented (e.g., instructor and student behaviors) [14,16]. In this study, we focus on the instructional critical components, as some studies have noted that these components more directly impact student outcomes [16,27]. We determine the instructional critical components of each active learning method through extensive literature review, similar to Ref. [19] (see further details in Methods).

Fidelity studies also require measurement and validation of fidelity of implementation. Fidelity can be measured using researcher or expert ratings from observations or documentation, user surveys, or interviews [19,24]. Researcher and expert ratings, however, are subjective and may be prone to biases, and user surveys and interviews may be inaccurate. In this study, we measure fidelity by applying a structured observation protocol to direct classroom observations and mapping the critical components to corresponding codes in the protocol (see further details in the next section and in Methods). We

validate fidelity by comparing the observations of broader implementations to those of high-fidelity implementations of the active learning methods, as suggested by prior work [16].

B. Classroom Observation Protocol for Undergraduate STEM (COPUS)

We measure fidelity by characterizing the instructor and student activities taking place during class using direct observation. Various classroom observation protocols have been developed with this aim [28]. These protocols are typically either open-ended or structured. Open-ended protocols expect observers to provide answers to certain prompts based on their observations, while structured protocols require observers to check off the presence of a specified list of activities. We aim to compare observations of different instructors; therefore, we opted to use a structured protocol for our study. We use the COPUS [21], a structured observation protocol with 25 specified codes to characterize both instructor (12 codes) and student (13 codes) behaviors in the classroom (Table I). COPUS observers check off codes that are present in two-minute time intervals.

TABLE I. Summary of COPUS codes for instructor and student activities. Full definitions can be found in Ref. [21].

Code	Definition
Instructor	
Lec	Lecturing about course material to the whole class (e.g., presenting content or a problem solution)
RtW	Real-time writing on a board or document projector in front of the whole class
FUp	Following up on a clicker question or other group activity in front of the whole class, including showing student responses
PQ	Posing a nonclicker, nonrhetorical question to the whole class, including asking for student questions
CQ	Asking a clicker (or any electronic response system) question to the whole class, and students working on the question
AnQ_I	Listening to and answering questions from students while the entire class listens
MG	Moving and guiding ongoing student work during active learning task (also includes teaching and learning assistants)
1o1	One-on-one extended discussion with one or more individuals, while inattentive to other groups
D/V	Interaction with a demonstration, video, experiment, simulation, or animation
Adm	Administrative tasks, such as announcements on test dates or student evaluations
W	Waiting instead of interacting with or observing students
O	Other
Student	
L	Listening to the instructor lecture on course content and unrelated content, including taking notes
Ind	Individual thinking or individual problem solving, with or without explicit instructor prompting
CG	Answering clicker (or other electronic response system) question in groups of two or more
WG	Working in groups on worksheets that have been handed out
OG	Other group activity not included under clickers or worksheets (e.g., laboratory experiment, nonelectronic voting system)
AnQ_S	Individual student answering a question posed by the instructor or multiple students saying an answer at the same time
SQ	Individual student asking a question to instructor or another student in front of the whole class
WC	Whole-class discussion involving input from different students, usually moderated by the instructor
Prd	Predicting the outcome of demonstration or experiment individually or in groups
SP	Student presentation to the whole class, e.g., project presentations or presenting problem solutions
TQ	Students working on tests or quizzes either individually or in groups
W	Waiting, e.g., while the instructor is fixing technical issues or the instructor arrives late to class
O	Other

We use the COPUS in the current study because this protocol includes both instructor and student codes, and we aim to characterize both of these facets of instructional critical components (Fig. 1). Stains and colleagues, moreover, identify the COPUS as the only tool that “aligns with the measure of some of the potential critical components” [24] (p. 9) of educational interventions. Additionally, in this study, we use data from a combination of live and video-recorded classroom observations. COPUS is well suited for both types of observations and affords the ability to measure intercoder reliability.

COPUS was designed to facilitate descriptive studies about instructional practices [21], and many existing research studies use descriptive statistics to analyze COPUS data (e.g., calculating the fraction of two-minute time intervals that each code is present) [22,29]. Instructional practices, however, are more complex than descriptive measures may capture, as they involve both instructor and student behaviors that occur dynamically over time. To gain a more nuanced understanding of teaching practices, researchers have increasingly turned to more advanced analytical approaches, such as clustering, mixture modeling, and network analysis of COPUS data [23,30–32].

We use both descriptive measures and network analysis in the current study so that our measures of fidelity capture a rich representation of instructional practices. These two conceptualizations of fidelity are distinct, but complementary: descriptive measures tell us the extent to which certain activities are implemented, while network analysis tells us about the specific sequences in which these activities are implemented.

C. Network analysis

Network analysis is a method for analyzing complex systems [33,34]. Networks consist of nodes, which are the actors or entities in the system of interest, and edges, which indicate connections or relationships between the nodes. In DBER, network analysis has been used in multiple ways, though its primary application is in *social* network studies, where the researchers examine relationships among people (e.g., students). Commeford and colleagues, for example, used network analysis to examine how patterns of peer interactions change over the course of a semester in different active learning methods [22]. Other studies have used network analysis to study student reasoning when solving physics problems [35,36].

In this study, we apply network analysis to our COPUS observations to capture the complex and temporal nature of instruction, similar to Ref. [23]. We represent classroom activities (i.e., the COPUS codes) as nodes in our observation networks and the chronological transitions between activities as the edges. Here, we aim to compare broad implementations of active learning methods to high-fidelity implementations, and treating classroom observations as networks allows us to quantify fidelity by calculating a

single similarity measure between pairs of classroom observations (see further details in Methods).

III. METHODS

We examine fidelity of implementation of three well-established and widely implemented active learning methods used to teach introductory physics and astronomy: SCALE-UP, ISLE, and Tutorials. In this section, we describe the critical components of each of these methods as well as our data sources and analysis methods. De-identified data and analysis scripts can be found at Ref. [48].

A. Identifying critical components of active learning methods

Using the fidelity of implementation framework (Fig. 1) [14,16,24], we reviewed existing literature to determine the critical components of each method (Table II). This literature included the foundational papers describing each method and research studies examining various aspects of these methods. Importantly, the authors of one of these research studies had their descriptions of the active learning methods checked by the method developers, increasing the validity of their descriptions [22]. We considered a course element to be a critical component if the developers noted the element as part of the curriculum and/or if the aforementioned study described the element in their expert-validated description. The first two authors engaged in negotiated discussions until consensus was reached about the final set of critical components for each method, which are described in the subsections below.

We then identified all COPUS codes that would indicate the presence of each critical component by mapping the descriptions of the critical components onto the definitions of the COPUS activities. For example, if laboratory experiments were a critical component, we expected the instructor to be moving around and guiding group work (MG code in the COPUS) while the students work in groups (OG code in the COPUS because the group work here is experimental rather than in the form of clicker questions or worksheets). As COPUS was designed to capture student and instructor behaviors during class, critical components that are related to physical classroom structure, course structure (e.g., time allotted to lecture, laboratories, and recitations in a given week) or out-of-class activities have no corresponding COPUS codes and are beyond the scope of our analysis. Similar to identifying the critical components, the first two authors iteratively discussed the relevant COPUS codes until consensus was reached.

1. SCALE-UP

SCALE-UP was developed to “scale up” the Studio Physics pedagogy, which was originally intended for small

TABLE II. Critical components of the three examined active learning methods and corresponding instructor and student COPUS codes (see definitions in Table I). No student or instructor COPUS codes are assigned to critical components that occur outside of observed classroom activities or relate to course structures more broadly.

Method	Critical component	Instructor codes	Student codes
SCALE-UP [37–40]	Restaurant-style classrooms	...	...
	Lab, lecture, and recitation integrated together	...	...
	New material introduced through preclass readings and quizzes	...	...
	Students work on tangibles, ponderables, and experiments in groups	MG	OG
	Students present their answers	...	SP
	Class-wide discussions after completing activities	FUp, PQ	AnQ_S
ISLE [41–44]	Teaching assistant training	...	...
	Students complete readings after class	...	...
	Students devise explanations for observation experiments in groups	MG, D/V	OG
	Students create and predict outcomes of testing experiments in groups	MG	OG, Prd
	Students solve problems with classroom response system	CQ, MG	CG, Ind
Tutorials [45–47]	Teaching assistant training	...	...
	Students complete pretest and posttest	...	...
	Students complete worksheets in groups	MG	WG

classrooms [37–40]. One of the critical components of the SCALE-UP method is the physical classroom layout: the classrooms are often described as “restaurant-style,” where each table holds two to three groups of three to four students each. All course components (i.e., lectures, labs, and recitations) are integrated into one classroom meeting time, which often occurs three times per week for two hours. In-class time is focused on group work, where students work on “tangibles” (e.g., short hands-on activities), “ponderables” (e.g., conceptual, open-ended questions that encourage critical thinking, estimations, and problem solving), and lab activities [37]. Instructors serve as coaches during class time, circling through the classroom and helping students come up with answers to their own questions. After completing activities, students are encouraged to present their answers to the class, and their answers are then reviewed and discussed by peers and instructors. New material is typically introduced through preclass readings and assessed through quizzes completed before the class meetings.

2. ISLE

ISLE can be implemented in all components of a class, or it can be implemented only within the lab component [41–44]. Critical to this method is that students observe an experiment and provide multiple explanations (i.e., scientific hypotheses) for the observations and then conduct or design an experiment to test their explanations. Based on the results of their testing experiment, they either reject, revise, or accept their explanations. This cycle is designed to allow students engage in the scientific process as they learn physics. Students also engage in quantitative problem solving in ISLE classes, typically using elements of Peer Instruction [5], such as classroom response systems

(e.g., clickers). Students are assigned readings after in-class sessions to emphasize that scientific knowledge is sourced from scientific processes. Teaching assistants for ISLE classes usually receive training prior to class sessions. Worksheets that incorporate most of these aspects of ISLE have been developed and popularized for use in introductory physics labs [49].

3. Tutorials

Tutorials for Introductory Physics [45–47] and Astronomy [50] are typically implemented in either the recitation or lecture component of a course. Tutorials consist of students working on worksheets in groups, with the teaching assistant or instructor moving through groups to address student questions. The Tutorial worksheets are carefully designed to elicit, confront, and resolve common student misconceptions.

Meetings with teaching assistants are usually held before the Tutorials are implemented to familiarize them with the worksheets. Pretests are usually administered after relevant topics are covered during lecture, but before the Tutorials are implemented. After the Tutorials, a posttest is administered to students.

B. Data sources

This study uses data from two iterations of a national research project titled “Characterizing Active Learning Environments in Physics” (CALEP): classroom observations of high-fidelity implementations of each method (from the first iteration of the project, CALEP1), and classroom observations and concept inventory data from 18 adopters of the three methods (from the second iteration of the project, CALEP2).

1. CALEP1: High-fidelity implementations

In the first iteration of the project, researchers collected classroom observation data from one high-fidelity implementation of several active learning methods (Tables III and IV) [22]. The classroom observations captured all class meetings within one week and were conducted live and in-person by a single observer using the COPUS. The high-fidelity implementation sites were chosen as either the institution where the active learning method was developed, or an institution recommended by instructors with extensive research experience on the methods. Further details of the data collection procedures for the CALEP1 study can be found in Ref. [22]. In the current study, we analyze the COPUS data from the SCALE-UP, ISLE, and Tutorials courses collected in this prior work.

TABLE III. Types of institutions represented in the two data sources. For CALEP1, instructors came from three unique institutions in the United States (one per method) [22]. For CALEP2, the 18 instructors came from 16 different institutions in the United States. Carnegie Classifications of Research Activity are from 2025: R1 indicates “Very High Research Spending and Doctorate Production,” R2 indicates “High Research Spending and Doctorate Production,” and RCU indicates “Research Colleges and Universities.”

Type of institution	CALEP1	CALEP2
Public/Private	3/0	9/7
Highest degree awarded		
Associate’s	0	1
Bachelor’s	0	2
Master’s	0	6
Ph.D.	3	7
Carnegie classification		
R1	3	5
R2	0	3
RCU	0	4
Hispanic-serving institution	0	3

TABLE IV. Class sizes (i.e., number of enrolled students) by active learning method. For the high-fidelity implementation of Tutorials, the recitation-only and whole-class observations come from the same course, so the number of enrolled students is the same. For broader implementations, the values indicate means, with standard deviations in parentheses.

Method	High-fidelity implementation	Broader implementations
SCALE-UP	71	64.1 (38.8)
ISLE	28	24.0 (9.8)
Tutorials (recitation-only)	171	30.7 (19.0)
Tutorials (whole-class)	171	38.3 (22.4)

2. CALEP2: Broader implementations

As part of the current iteration of the project [23,32], we recruited 18 introductory physics and astronomy instructors across the United States who self-reported using SCALE-UP, ISLE, or Tutorials in their introductory physics or astronomy course. Instructors were recruited through the research team’s personal networks, advertisements on the American Physical Society website, and/or identification through grant advisory board members. We targeted our recruitment at college-level physics instructors broadly; however, most of the instructors who chose to participate were either physics education researchers or high consumers of physics education research. All courses took place during the fall 2023, spring 2024, or fall 2024 semesters.

Seven SCALE-UP instructors, four ISLE instructors, and seven Tutorials instructors participated in the study, and these instructors’ courses spanned a wide range of institution types and class sizes (Tables III and IV). One Tutorials course was astronomy, and all other courses were physics. Within the Tutorials courses, there were two types of implementation: whole-class implementations (four courses), where the Tutorials worksheets were implemented as part of lecture sections, and recitation-only implementations (three courses), where the Tutorials worksheets were only implemented in recitation sections. We analyzed these different implementations separately, as noted in the Results.

We collected two types of data from each instructor: video recordings of class sessions and pre- and postsemester concept inventory scores. Instructors recorded class sessions using their own camera, Zoom (a cloud-based platform with recording features), or a camera we sent to them by mail. They recorded three consecutive classroom meetings to capture a typical week in their course, as recommended in prior literature [31]. All 18 instructors recorded three full class sessions and are included in our fidelity analysis (i.e., the first and second research questions).

Instructors also collected concept inventory data through research-validated assessments administered online at the beginning and end of the semester. Instructors chose a concept inventory that best aligned with their course content. Most instructors used the Force Concept Inventory [51] or Force and Motion Concept Evaluation [52] (see Table VII in the Appendix). Out of the 18 instructors, 14 had at least 50% of enrolled students who took both the pre- and postsemester concept inventory (i.e., had matched responses) and are included in our analysis of student conceptual learning (i.e., the third research question). Each instructor received \$1000 as compensation for participating in the study.

C. Data analysis

1. COPUS coding

As mentioned, we used the COPUS to capture student and instructor behaviors during class. The CALEP1

TABLE V. Summary of the data used in our study by active learning method: the subset of observations from the CALEP1 COPUS data included in our analysis, the number of courses with three video observations in the CALEP2 data, and the number of CALEP2 courses with concept inventory data.

Method	CALEP1 classroom observations	CALEP2 courses with three video observations	CALEP2 courses with concept inventory data
SCALE-UP	3 class sessions	7	5
ISLE	1 lecture and 1 recitation	4	3
Tutorials (recitation-only)	1 recitation	3	3
Tutorials (whole-class)	1 lecture and 1 recitation	4	3

observations were conducted live as part of a prior study [22], so we used the existing COPUS coding of those courses. These observations, however, included a mix of course components (i.e., lectures and recitations) and, in some cases, several different sections of students conducting the same activity (e.g., several recitation sections of a large-enrollment course using Tutorials). To ensure that the CALEP1 and CALEP2 data were comparable for our fidelity analysis, we used a subset of the CALEP1 observations with course structures that most aligned with those of the observations collected for CALEP2 in our analysis (Table V). In cases where the CALEP1 data included observations of different sets of students completing the same activity (i.e., recitation sections of both ISLE and Tutorials), we used a random number generator to select one observation to include in the current analysis (no CALEP2 instructors recorded multiple sections of students doing the same activity). For Tutorials, the recitation section selected from the CALEP1 data was used as the high-fidelity comparison for the recitation-only implementations in CALEP2. We aggregated this same section's observation with a lecture section observation when comparing CALEP1 to the CALEP2 whole-class implementations of Tutorials to increase comparability (the CALEP2 whole-class implementations embedded Tutorials worksheets within their lecture sections). The CALEP1 data also included two different implementations of ISLE: lab only and whole-class (i.e., ISLE implemented in lecture, recitations, and labs). We only used the whole-class implementation data from CALEP1, as its structure was more similar to that of the ISLE courses in CALEP2.

The second, third, and fourth authors applied the full COPUS to the 54 classroom video recordings collected in CALEP2 (i.e., three video recordings for each of the 18 instructors). First, all three authors independently coded one video per active learning method. Then, the authors iteratively met to discuss disagreements and individually re-coded the videos until an interrater reliability of over 80% was reached for each method (as measured with Cohen's Kappa [53]; see values in Ref. [32]). After reaching this level of agreement, we randomly assigned one of these three authors to code each video in its entirety.

We included all 25 COPUS activities in our coding; however, we excluded three instructor codes (1o1, O,

and W; Table I) and four student codes (WC, TQ, W and O; Table I) from the current analysis because these codes were rarely marked as being present (i.e., each of these codes was observed in less than 2% of all two-minute time intervals in the dataset). We also aggregated the observations for each instructor to get a more holistic representation of their instructional choices. For the CALEP1 data, we aggregated lecture and recitation observations for ISLE and the whole-class implementation of Tutorials. For the CALEP2 data, we combined each instructor's three video recordings (all CALEP2 instructors recorded the same course component for all three observations).

2. Constructing classroom observation networks

For each set of observations considered in our comparisons (i.e., four from CALEP1 and 18 from CALEP2, Table V), we constructed two classroom observation networks: one for instructor activities and one for student activities. The network nodes were the COPUS codes. The network edges were directed, with arrows indicating a chronological transition from one code to another (i.e., pointing from one code to the code that happened afterward). As in Ref. [23], we considered a transition between codes to be when a code was not present in a two-minute time interval and then became present in the following two-minute time interval. We created an edge from all codes in the previous time interval to the newly appearing code(s) in the following interval. The network edges were also weighted as the number of occurrences of each transition divided by the total number of observed two-minute time intervals (i.e., to normalize for different class durations). To illustrate, the example COPUS observation in Table VI would be represented as a network with nine nodes (one per code) with the following edges and edge weights:

- Adm $\rightarrow$ Lec (weight = $1/5 = 0.2$)
- Adm $\rightarrow$ RtW (weight = $1/5 = 0.2$)
- Lec $\rightarrow$ CQ (weight = $2/5 = 0.4$)
- RtW $\rightarrow$ CQ (weight = $1/5 = 0.2$)
- Lec $\rightarrow$ FUp (weight = $1/5 = 0.2$)
- CQ $\rightarrow$ FUp (weight = $1/5 = 0.2$)
- FUp $\rightarrow$ CQ (weight = $1/5 = 0.2$)

TABLE VI. Toy example of COPUS coding for only instructor codes for 10 min of class time. Each row represents one two-minute time interval, with 1 indicating the presence of the behavior and 0 indicating that the behavior did not occur in that interval.

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

Though not directly part of the network structure, we also calculated the fraction of class time spent on each COPUS code as the fraction of the observed two-minute time intervals that the code was present. We represented these fractions of class time in the network diagrams in the node colors (see Fig. 4). For the example observation in Table VI, the fraction of class time spent on Lec is 0.6 (observed in 3 out of 5 total time intervals), on CQ is 0.4 (observed in 2 out of 5 total time intervals), on RtW, FUp, and Adm is 0.2 (observed in 1 out of 5 total time intervals), and is 0 for the remaining instructor codes that are not observed in any time interval.

3. Measuring and validating fidelity

To address our first research question, we compared the fraction of class time spent on each COPUS code relevant to the identified critical components between the high-fidelity (CALEP1) and broader (CALEP2) implementations (Table II). We conducted this analysis descriptively by comparing the high-fidelity implementation measurement to the distribution of broader implementations measurements for each active learning method.

To address our second research question, we calculated four similarity metrics for each pair of high-fidelity and broader implementation within the same active learning method: two that compare the fraction of class time spent on each COPUS code (one measure for instructor codes and one measure for student codes, including all COPUS codes), and two that compare the structures of the classroom observation networks (one measure for instructor networks and one measure for student networks, including all COPUS codes). We used cosine similarity as our similarity metric, as in Ref. [23], because cosine similarity is less sensitive to network sparsity than other similarity indices, such as the Jaccard index [54,55], and a few of the networks in our study contained very few edges. Cosine similarity measures the similarity between a pair of vectors, with the cosine of the angle between the vectors being a measure of their effective similarity. The dimensionality of the vectors compared in our calculations was the same across all instructors and active learning methods

because all observations used the same set of COPUS codes. Cosine similarity values range from -1 to 1 , with values closer to -1 indicating dissimilar vectors, 0 indicating orthogonal vectors, and 1 indicating identical vectors [56]. Mathematically,

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

for two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$.

To validate fidelity based on fractions of class time spent on each COPUS code, the two vectors were defined as the set of these fractions for the high-fidelity and broader implementation being compared. In the case of the observation in Table VI, for example, the relevant vector of the fractions of two-minute intervals spent on each code (starting with Lec and ending with Adm) would be $\mathbf{v} = [0.6, 0.2, 0.2, 0, 0.4, 0, 0, 0, 0.2]$. A second vector would be constructed in the same way for the comparison observation (e.g., the high-fidelity implementation), and these two vectors would be input to the above equation. Higher values of cosine similarity indicate that the two instructors had similar distributions of class time spent on the COPUS codes.

To validate fidelity based on classroom observation networks (i.e., capturing the transitions between COPUS codes), the two vectors were defined as the set of network edge weights for the high-fidelity and broader implementations being compared: $\mathbf{v} = [W(\text{Lec} \rightarrow \text{RtW}), W(\text{Lec} \rightarrow \text{FUp}), \dots, W(\text{Lec} \rightarrow \text{Adm}), W(\text{RtW} \rightarrow \text{Lec}), W(\text{RtW} \rightarrow \text{FUp}), \dots, W(\text{Adm} \rightarrow \text{D/V})]$, where W is the edge weight. Higher values of cosine similarity indicate that the two instructors had similar sequences of COPUS codes (e.g., spending long amounts of time on certain activities versus cycling through different activities frequently).

We acknowledge that while COPUS is a useful tool for measuring procedures and patterns of classroom activities, it does not measure other relevant aspects of fidelity, such as pedagogical or epistemological goals and the extent to which course materials are research-based. Our interpretations of cosine similarity values only relate to fidelity as measured by student and instructor behaviors during class time.

4. Relating fidelity to student conceptual learning

For each of the 14 broader implementations with concept inventory data (Table V), we calculated an effect size using Hedges' g to measure student conceptual learning gains. Hedges' g is the standardized difference between students' mean pre- and postsemester concept inventory scores, only including students with matched responses. Hedges' g is identical to Cohen's d , but includes a correction factor to account for small sample sizes (i.e., small-enrollment physics courses) [57]. Hedges' g has been shown to be more suitable for comparing effect sizes

across courses that include both large and small sample sizes [57], as we have in this study (Table IV).

To address our third research question, we descriptively examined the relationship between fidelity of implementation (i.e., the four cosine similarity metrics described above) and effect sizes using scatter plots.

IV. RESULTS

Here, we present our findings by research question.

A. Fidelity of implementation: Critical components

Across all three active learning methods, the broad set of instructors implement the critical components to a similar extent as the high-fidelity implementers (Fig. 2). In the subsections below, we describe more specific findings for each method.

1. SCALE-UP

Between the high-fidelity and broader implementations of SCALE-UP, instructors spend similar fractions of class time on all critical components: for the bolded codes, the red diamonds representing the high-fidelity implementations fall within the gray distributions representing the broader implementations in Fig. 2. One possible exception to this pattern is student presentations to the whole class (SP), as the high-fidelity implementation of SCALE-UP did not implement this component at all. It is possible that in this course, students presented their solutions or ideas in the form of answers to instructor questions (red diamond is on the upper end of the gray distribution for AnQ_S in Fig. 2).

Other subtle differences include the broader implementations of SCALE-UP spending less class time on instructor follow-up to activities and more time on students doing group work than the high-fidelity implementation (red diamond is on the upper end of the gray distribution for FUp and the lower end of the gray distribution for OG in Fig. 2). The latter may be explained by the prevalence of clicker questions in the high-fidelity implementation (red diamond is much higher than the gray distribution for CQ and CG in Fig. 2), indicating that different modes of group work were used in that implementation.

2. ISLE

With the exception of the instructor moving and guiding student group work, most of the critical components of ISLE have low prevalence in both the high-fidelity and broader implementations (low values for red diamonds and black dots for all bolded codes except MG in Fig. 2). Instead, there is a relatively high prevalence of students completing worksheets in groups in both the high-fidelity and broader implementations, with a slightly higher prevalence of worksheets in the high-fidelity implementation

than the broader implementations (red diamond is on the upper end of the gray distribution for WG in Fig. 2). As mentioned earlier, the method experts designed worksheets for use in ISLE classrooms [49]. These worksheets include prompts that elicit several of the critical components (e.g., students making predictions about demonstrations). It is possible that both the high-fidelity and broader implementations used these ISLE-specific worksheets or operationalized the critical components through their own worksheets; however, we did not have access to the course materials to check this.

Another notable difference is that broader implementations of ISLE implemented individual student work more than the high-fidelity implementation (red diamond is on the lower end of the gray distribution for Ind in Fig. 2). Individual work is often coded alongside working on clicker questions in groups if there are some students working alone, suggesting the high-fidelity implementation engaged all students in small group discussions when completing these types of questions.

3. Tutorials

In both the whole-class and recitation-only versions of Tutorials, the high-fidelity and broader implementations have high and comparable prevalence of the two critical components: the instructor moving and guiding group work and the students completing worksheets in groups (similarly high values of red diamonds and black dots for MG and WG codes in Fig. 2 for each version). In the recitation-only version of Tutorials, the instructor moved and guided the group worksheets in the high-fidelity implementation more than in the broader implementations (red diamond is above the gray distribution for MG in Fig. 2). In general, worksheets are used less frequently in the whole-class version of Tutorials than the recitation-only version, which aligns with the goals of each of these versions of the method (i.e., Tutorials supplementing lecture and other activities in the whole-class version versus Tutorials being the main activity of recitation section in the recitation-only version; higher values of red diamonds and gray distributions for WG in the recitation-only than whole-class version in Fig. 2).

B. Fidelity of implementation: Full COPUS observations

We use cosine similarity to compare both the proportions of class time spent on each COPUS code (including all COPUS codes, not only those related to critical components, because some critical components may be operationalized through different mediums such as worksheets) and the classroom observation networks (which capture temporal transitions between all COPUS codes) between high-fidelity and broader implementations of each active learning method (Fig. 3).

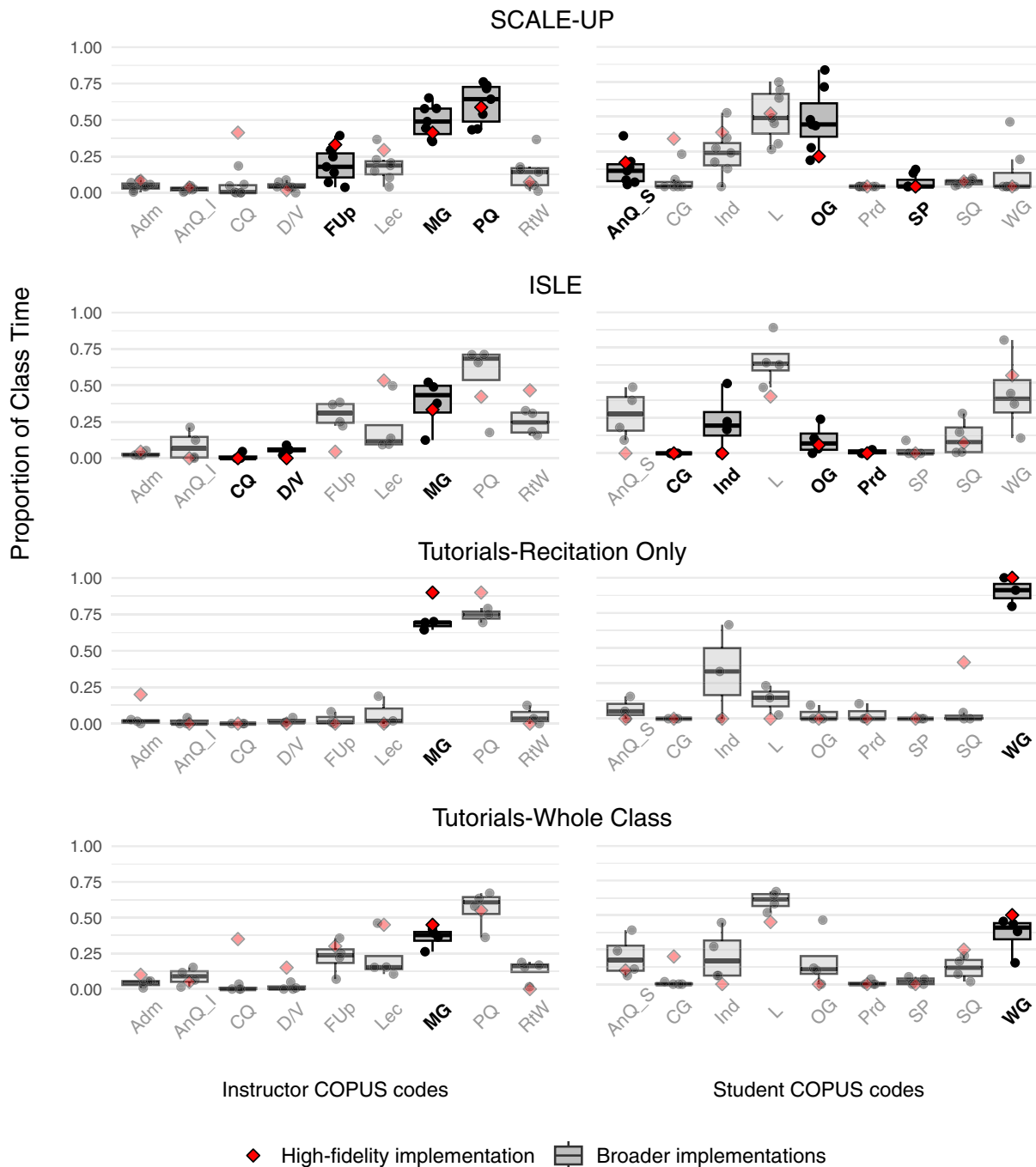


FIG. 2. The proportion of two-minute time intervals spent on each COPUS code in high-fidelity and broader implementations of each active learning method. The boxplots indicate interquartile range, with the bold lines within the boxplots representing the medians and whiskers denoting 1.5 times the interquartile range. The COPUS codes for the critical components (Table II) are bolded, and their corresponding boxplots are nontransparent.

1. Proportion of time spent on classroom activities

With regard to the proportion of class time spent on each COPUS code, there is high fidelity for all broader implementations of all three methods [cosine similarity values close to one for both instructor and student codes in each active learning method in Fig. 3(a)]. Only considering the student codes, there is comparable fidelity across all

active learning methods [the gray distributions overlap for student codes in Fig. 3(a)]. Only considering the instructor codes, there is variation in fidelity across methods: the recitation-only version of Tutorials has the highest fidelity, followed by SCALE-UP, the whole-class version of Tutorials, and then ISLE. This pattern of fidelity is largely consistent with that found in our descriptive comparison of

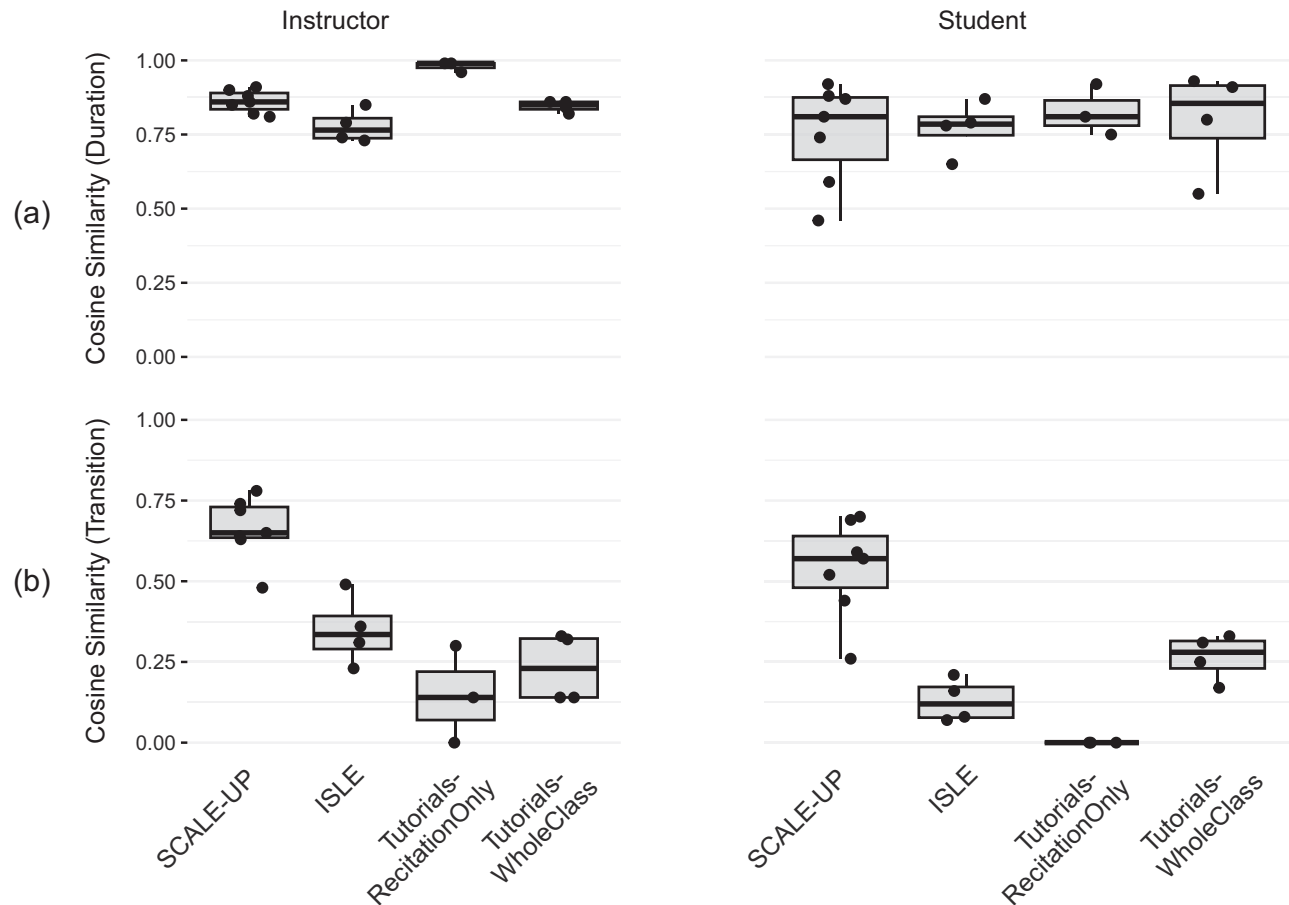


FIG. 3. Cosine similarity values comparing (a) the proportion of class time spent on all COPUS activities and (b) patterns of transitions between all COPUS activities (i.e., using the edge weights of classroom observation networks) in broader implementations to high-fidelity implementations of each active learning method. The gray boxplots indicate interquartile range, with the bold lines representing the medians and whiskers denoting 1.5 times the interquartile range.

the prevalence of critical components presented in the previous section.

Additionally, there is more variation in fidelity within each active learning method for student codes than instructor codes [for each method, the gray distribution for student codes spans a wider range than the gray distribution for the instructor codes in Fig. 3(a)]. This suggests that the same instructor code may correspond to different student codes in different implementations. For example, if an instructor poses a question or activity to the whole class, PQ, this student activity may be in the form of worksheets, WG, or other group work, OG.

2. Sequences of classroom activities

With regard to classroom observation networks, which capture the chronological patterns of activities, we observe lower cosine similarity values across all active learning methods than the fidelity measured with regard to proportions of class time spent on each activity [for each method, cosine similarity values are lower in Fig. 3(b) than Fig. 3(a)]. This indicates that even when instructors spend

similar amounts of time on certain classroom activities, those activities are executed in very different ways (e.g., short, discontinuous sequences of activities versus long, continuous activities). For example, Fig. 4 shows the classroom observation networks for the high-fidelity implementation of SCALE-UP and one broader implementation of SCALE-UP with high cosine similarity values for duration fidelity and low cosine similarity values for transition fidelity (only for the instructor codes). We see that the node colors, which indicate the proportion of time spent on the codes, are mostly comparable between the two networks (with the exceptions of clicker questions, CQ, and following up, FUp). The broad implementation [Fig. 4(b)], however, generally has thinner edge weights than the high-fidelity implementation [Fig. 4(a)], indicating that the broad implementation includes extended time periods of activities and the high-fidelity implementation cycles through shorter versions of similar activities.

Unlike fidelity based on duration of class activities, there is substantial variation in cosine similarity values

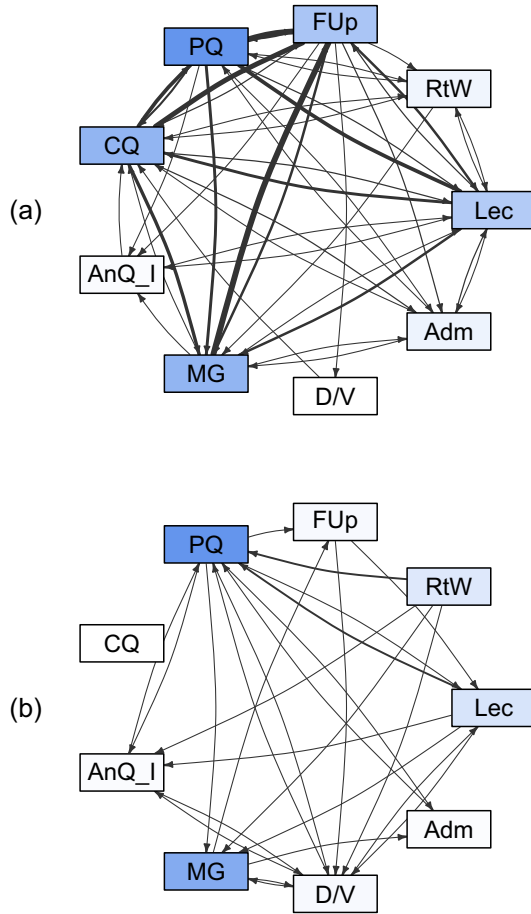


FIG. 4. Example classroom observation networks, only considering instructor codes, for (a) the high-fidelity implementation of SCALE-UP and (b) one broader implementation of SCALE-UP with high cosine similarity values for duration fidelity and low cosine similarity values for transition fidelity. Node color represents the fraction of observed two-minute time intervals spent on each code, with darker shades indicating larger fractions. Edges point from an initial code to the code that occurs in the following two-minute time interval of the COPUS observation. Edge width indicates the number of transitions occurring between those two codes normalized by the total number of two-minute time intervals observed. Networks for all courses in this study are available at Ref. [48].

based on transitions between class activities across the active learning methods. SCALE-UP seems to be implemented closer to the high-fidelity implementation than the other active learning methods when we consider the chronological patterns of activities [gray distributions are higher for SCALE-UP than the other methods for both instructor and student codes in Fig. 3(b)]. We note that some cosine similarity values are equal to zero for the recitation-only version of Tutorials because those classroom observation networks are very sparse, as the class sessions often involve students doing worksheets for the whole class with very few transitions to or from other activities.

C. Relating fidelity of implementation to student conceptual learning

There is no clear relationship between fidelity of implementation of the observed active learning methods, as measured by either duration of class time spent on each COPUS code or the transitions between COPUS codes captured in the classroom observation networks, and student conceptual learning [Fig. 5]. For fidelity measured using duration of activities, all cosine similarity values are fairly high, so we may not have enough variation of fidelity in our data to determine if any relationship exists [Fig. 5(a)]. For fidelity measured using transitions between activities, we have more variation in cosine similarity values, but there is no clear upward, downward, or flat trend in the scatterplot [Fig. 5(b)]. This suggests a preliminary finding that SCALE-UP has both higher fidelity in terms of sequences of activities and higher student learning gains than the other methods (green dots are clustered in the top right of Fig. 5(b) for both instructor and student codes), but we cannot make any strong conclusions about particular active learning methods based on our small sample size.

V. DISCUSSION

In this study, we measured the fidelity of implementation of 18 introductory physics instructors using one of three named active learning methods: SCALE-UP, ISLE, and Tutorials. We did so using direct classroom observations and the COPUS, which is likely more accurate than self-reported instructor practices collected through survey or interviews [18,19]. We analyzed COPUS data in two ways: first, by measuring the extent to which the critical components of each method were implemented by the 18 instructors as compared with high-fidelity implementers, and second, by comparing the full COPUS observations of these broader implementations to those of high-fidelity implementations. We also expand upon existing studies of fidelity in undergraduate science courses by relating fidelity to student conceptual learning.

A. Fidelity of implementation of named active learning methods

We found that broader implementations of active learning mostly use the critical components of the methods. Previous studies have measured fidelity solely based on whether or not each critical component is present in the implementation; by this definition, most of the instructors in our study implemented the method with fidelity [18,19]. However, we also measured the fractions of class time spent on all COPUS codes. These fine-grained measurements again highlighted that broader implementations used the critical components, and other classroom activities, to similar extents as the high-fidelity implementations. These results may suggest that active learning methods are generally being implemented as intended

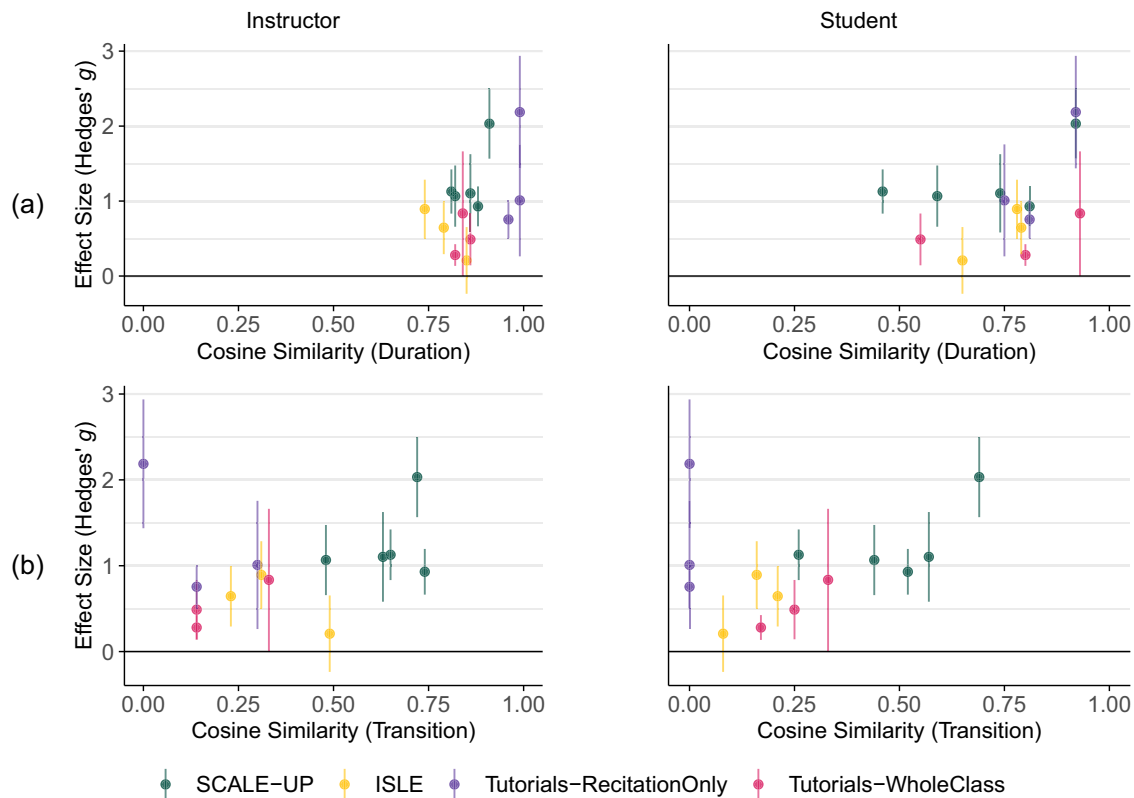


FIG. 5. Effect sizes of student concept inventory scores versus (a) duration fidelity, measured using the proportion of class time spent on all COPUS codes, and (b) transition fidelity, measured using classroom observation networks (Fig. 3). Dots indicate Hedges' g values, and error bars indicate 95% confidence intervals.

by the developers; however, this pattern could be due to the nature of our study sample. Most of the instructors in this study are currently conducting physics education research, previously conducted physics education research (i.e., during graduate school), or are high consumers of physics education research. Prior work has shown that instructors trained in science education research may teach differently, and have more positive impacts on student outcomes, than a random sample of instructors, possibly biasing our results [17]. It is possible that we observed less variation in the use of critical components and all COPUS activities within our sample than exists among the wider population of physics instructors. Existing literature, however, suggests that most typical physics faculty (i.e., nonphysics education researchers) use traditional teaching methods rather than active learning strategies [58], so our sample of users of named active learning methods is likely representative. Still, we encourage future work to identify and measure fidelity of implementation of a larger set of college physics instructors who implement named active learning methods to validate our findings.

We also observed that in some active learning methods, instructors spend a relatively high proportion of time on noncritical component codes, for example posing questions (PQ) in ISLE and the recitation-only version of

Tutorials. Although these codes do not directly reflect critical components of the methods, they are central to active learning instruction and facilitate the use of critical components (e.g., moving and guiding groupwork, MG, and posing questions are often coded concurrently).

At the same time, we observed differences in the sequences of activities through which broader implementers used these critical components (i.e., based on classroom observation networks that capture the chronological transitions between activities [23]). That is, while broader implementation instructors spend similar proportions of time on activities as the high-fidelity implementations, they differ in how they distribute these activities across class time. One of the SCALE-UP instructors, for example, used a few long time periods to implement groupwork activities. The high-fidelity implementation of SCALE-UP, on the other hand, implemented the same types of activities in many short time periods. These observed differences in sequences of activities may be due to instructor choices alone and/or other factors, such as different course structures or formats (e.g., lecture versus recitation sections) and the topical arc of the physics content covered in that week. Future research should continue to characterize these different instructional patterns across a broader sample and determine the method developers' intentions for specific ways of implementing

the activities (i.e., to validate the fidelity of transitions between activities in addition to the critical components).

Interestingly, when considering these chronological patterns of COPUS codes, broader implementations of SCALE-UP were more similar to the high-fidelity implementation of SCALE-UP than the broader implementations of ISLE and Tutorials with their respective high-fidelity implementations. This pattern is surprising because both ISLE and Tutorials have more specific curricula (e.g., with carefully designed worksheets) than SCALE-UP, which places stronger emphasis on the physical classroom layout [59]. One hypothesis is that the structural features of SCALE-UP classrooms inherently constrain instructional choices, allowing for consistent instructor enactment of critical components even without highly prescriptive materials. Indeed, the broader implementations of ISLE and Tutorials courses took place in a wide variety of classrooms, possibly leading to a wide range in instructional practices. Another explanation is that instructors who adopt SCALE-UP receive more focused training or institutional support related to its implementation, which could lead to closer alignment with high-fidelity implementations. We recommend for future studies to examine these possibilities for variation in activity sequences across the active learning methods.

B. Fidelity of implementation and student conceptual learning

Our analysis revealed no clear relationship between fidelity of implementation and conceptual learning gains, as measured using concept inventories. We did, however, identify preliminary evidence that SCALE-UP courses tend to be implemented with higher fidelity (particularly as measured using the classroom observation transition networks) than the other methods. This may (at least partly) explain results from another study of these data documenting significantly higher student learning gains in SCALE-UP than ISLE [32]. Perhaps fidelity of implementation is necessary, but not sufficient for learning gains (i.e., other factors, such as class size, student preparation, and homework assignments, may also matter) [23]. This proposition aligns with the findings of prior work related to “reinvention” of research-based instructional strategies, where the researchers found a correlation between fidelity and outcomes of social innovation programs, and that local additions to the critical components further enhance impacts to outcomes [60].

Due to the nature and size of our sample, however, these findings relating fidelity to student learning should be considered exploratory and largely descriptive of the specific instructors in our current dataset. The limited variation in duration fidelity, for instance, is likely due to the vast teaching and physics education research experience of the instructors included in our study. More research

is needed to better understand the role of fidelity in the effectiveness of active learning methods to student outcomes in undergraduate physics (and science) courses, including a dataset with more instructors and more student outcomes (i.e., attitudes, identity, and experimental skills) than the study presented here. If there is a relationship between fidelity and student outcomes, this would indicate that preserving the critical components and the specific ways they are operationalized (i.e., activity sequences) is necessary for improving student outcomes. If there is no strong relationship between fidelity of implementation and student outcomes, this would imply that instructors can flexibly choose the structure in which they implement the critical components without sacrificing effectiveness.

C. Limitations and future work

While this study offers valuable insights to the fidelity of implementation of named active learning methods in introductory physics and astronomy, there are several limitations that may be addressed in future research. First, we used extensive literature review and negotiated discussions among the research team to identify the critical components of each active learning method and their mapping to COPUS codes. Future research that validates these sets of critical components, their corresponding COPUS codes, the fraction of class time instructors should spend on these codes, and the intended flow of or transitions between activities (e.g., through developer and/or expert interviews) will provide further context for our claims.

Second, we have assumed that high-fidelity implementation data from the first iteration of the project are indeed high fidelity. This is a plausible assumption because these observations were largely conducted at the development sites of the methods in our study; however, it is possible that over time (i.e., between the time of development and the implementation we analyzed), fidelity of implementation might have changed [61,62]. Relatedly, the previous researchers did not conduct some of the high-fidelity observations for full class sessions (i.e., they observed 20 min segments of some class sessions), though they noted that the observed subset of the full class time was representative [22,30]. Future studies should conduct multiple, full-length classroom observations of high-fidelity implementations and/or aim to observe more than one high-fidelity implementation of each named active learning method to validate our findings [63,64].

Finally, in terms of the COPUS coding, the high-fidelity and broader implementation data were collected and coded several years apart, so we could not establish interrater reliability across the coders involved in each iteration of the project. This may have led to subtle differences in the ways the codes were applied to the observations in each dataset; however, the COPUS was designed for consistent

application across many observers with minimal training, reducing the likelihood that small variations in coding meaningfully affected the overall patterns in our analysis [21]. Additionally, the COPUS may not capture the full range and nuance of classroom activities (e.g., when student predictions are embedded in group worksheets and when the same instructor activity coincides with multiple different student activities), and notably does not capture specific content covered, the extent to which course materials are research-based, or epistemic goals. Therefore, our analysis likely overlooks several aspects of the courses that are related to fidelity. We recommend for future work to examine possible improvements to the COPUS or other observation protocols and/or supplement direct classroom observations with other data, such as instructor interviews and course materials, to capture the full scope of fidelity.

VI. CONCLUSIONS

We have showcased a novel approach to measuring fidelity that we hope other researchers will use to better understand the gray area often associated with fidelity of implementation of active learning strategies. In our study of 18 introductory physics instructors using SCALE-UP, ISLE, or Tutorials, we found generally high fidelity of implementation and no clear relationship between fidelity and student conceptual learning. Future studies should validate and improve our approach and measure fidelity and its impacts across a more diverse set of instructors, including those outside of science education research, those in other scientific disciplines, and those outside of the United States. Such studies will help to narrow down the conditions under which active learning methods are the most effective for student outcomes, which will inform best practices for undergraduate science education.

ACKNOWLEDGMENTS

We thank the instructors and students who participated in our study. We also thank David Esparza, Ian Olivant, and Bruna Schons Ribeiro for their feedback on this manuscript. This work is supported by the National Science Foundation Grants No. 2111128 and No. 2139899, and the Cotswold Foundation Postdoctoral Fellowship at Drexel University.

DATA AVAILABILITY

The data that support the findings of this article are openly available [48].

APPENDIX

1. Concept inventories

Table VII summarizes the concept inventories used to measure student learning by active learning method.

TABLE VII. Number of broader implementation courses using each concept inventory by active learning method. Concept inventory data were not collected for the high-fidelity implementations.

Method	SCALE-UP	ISLE	Tutorials
Force concept inventory [51]	4	2	3
Force and motion concept evaluation [52]	0	1	1
Energy and momentum conceptual survey [65]	1	0	0
Energy concept assessment [66]	0	0	1
Light and spectroscopy concept inventory [67]	0	0	1

[1] 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).

[2] P. T. Terenzini, A. F. Cabrera, C. L. Colbeck, J. M. Parente, and S. A. Bjorklund, Collaborative learning vs. lecture/discussion: Students’ reported learning gains, *J. Eng. Educ.* **90**, 123 (2001).

[3] J. M. Braxton, W. A. Jones, A. S. Hirschy, and H. V. Hartley III, The role of active learning in college student persistence, *New Dir. Teach. Learn.* **2008**, 71 (2008).

[4] D. Miller, J. Deshler, T. McEldowney, J. Stewart, E. Fuller, M. Pascal, and L. Michaluk, Supporting student success and persistence in STEM with active learning approaches in emerging scholars classrooms, *Front. Educ.* **6**, 667918 (2021).

[5] Eric Mazur, *Peer Instruction: A User’s Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).

[6] M. Dancy, C. Henderson, N. Apkarian, E. Johnson, M. Stains, J. R. Raker, and A. 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).

[7] Y. Affriyenni, H. Georgiou, and N. Finkelstein, Navigating the adoption of research-based instructional strategies within the complex nature of higher education, *Phys. Rev. Phys. Educ. Res.* **21**, 020124 (2025).

[8] J. Gess-Newsome, S. A. Southerland, A. Johnston, and S. Woodbury, Educational reform, personal practical

- theories, and dissatisfaction: The anatomy of change in college science teaching, *Am. Educ. Res. J.* **40**, 731 (2003).
- [9] K. T. Foote, X. Neumeyer, C. Henderson, M. H. Dancy, and R. J. Beichner, Diffusion of research-based instructional strategies: The case of SCALE-UP, *Int. J. STEM Educ.* **1**, 10 (2014).
 - [10] C. Henderson and M. 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] L. E. Strubbe, A. M. Madsen, S. B. McKagan, and E. C. Sayre, Beyond teaching methods: Highlighting physics faculty's strengths and agency, *Phys. Rev. Phys. Educ. Res.* **16**, 020105 (2020).
 - [12] Everett Rogers, *Diffusion of Innovations*, 5th ed. (Free Press, New York, NY, 2003).
 - [13] Julia Willison, Erin M. Scanlon, and Jacquelyn J. Chini, Examining faculty choices while implementing the Next Gen PET curriculum through revealed causal mapping, in *Proceedings of the Physics Education Research Conference (PERC)* (American Association of Physics Teachers, College Park, MD, 2023), pp. 391–396, [10.1119/perc.2023.pr.Willison](https://doi.org/10.1119/perc.2023.pr.Willison).
 - [14] J. Century, M. Rudnick, and C. Freeman, A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge, *Am. J. Eval.* **31**, 199 (2010).
 - [15] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
 - [16] C. T. Mowbray, M. C. Holter, G. B. Teague, and D. Bybee, Fidelity criteria: Development, measurement, and validation, *Am. J. Eval.* **24**, 315 (2003).
 - [17] T. M. Andrews, M. J. Leonard, C. A. Colgrove, and S. T. Kalinowski, Active learning not associated with student learning in a random sample of college biology courses, *CBE—Life Sci. Educ.* **10**, 394 (2011).
 - [18] M. Dancy, C. Henderson, and C. 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] M. Borrego, S. Cutler, M. Prince, C. Henderson, and J. E. Froyd, Fidelity of implementation of research-based instructional strategies (RBIS) in engineering science courses, *J. Eng. Educ.* **102**, 394 (2013).
 - [20] E. Scanlon, B. Zamarripa Roman, E. Ibadlit, and J. J. Chini, A method for analyzing instructors' purposeful modifications to research-based instructional strategies, *Int. J. STEM Educ.* **6**, 12 (2019).
 - [21] M. K. Smith, F. H. M. Jones, S. L. Gilbert, and C. 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).
 - [22] K. Commeford, E. Brewwe, and A. Traxler, Characterizing active learning environments in physics using network analysis and classroom observations, *Phys. Rev. Phys. Educ. Res.* **17**, 020136 (2021).
 - [23] M. Sundstrom, J. Gambrell, C. Green, A. L. Traxler, and E. Brewwe, Beyond named methods: A typology of active learning based on classroom observation networks, *Phys. Rev. Phys. Educ. Res.* **22**, 010143 (2026).
 - [24] M. Stains and T. Vickrey, Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices, *CBE—Life Sci. Educ.* **16**, rm1 (2017).
 - [25] C. L. O'Donnell, Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research, *Rev. Educ. Res.* **78**, 33 (2008).
 - [26] B. Harn, D. Parisi, and M. Stoolmiller, Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools?, *Exceptional Child.* **79**, 181 (2013).
 - [27] R. Gersten, L. S. Fuchs, D. Compton, M. Coyne, C. Greenwood, and M. S. Innocenti, Quality indicators for group experimental and quasi-experimental research in special education, *Exceptional Child.* **71**, 149 (2005).
 - [28] S. Anwar and M. Menekse, A systematic review of observation protocols used in postsecondary STEM classrooms, *Rev. Educ.* **9**, 81 (2021).
 - [29] M. K. Smith, E. L. Vinson, J. A. Smith, J. D. Lewin, and M. K. R. Stetzer, A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions, *CBE—Life Sci. Educ.* **13**, 624 (2014).
 - [30] K. Commeford, E. Brewwe, and A. Traxler, Characterizing active learning environments in physics using latent profile analysis, *Phys. Rev. Phys. Educ. Res.* **18**, 010113 (2022).
 - [31] M. Stains, J. Harshman, M. K. Barker, S. V. Chasteen, R. Cole, S. E. DeChenne-Peters, M. K. Eagan Jr., J. M. Esson, J. K. Knight, F. A. Laski, M. Levis-Fitzgerald, C. J. Lee, S. M. Lo, L. M. McDonnell, T. A. McKay, N. Michelotti, A. Musgrove, M. S. Palmer, K. M. Plank, T. M. Rodela, E. R. Sanders, N. G. Schimpf, P. M. Schulte, M. K. Smith, M. Stetzer, B. Van Valkenburgh, E. Vinson, L. K. Weir, P. J. Wendel, L. B. Wheeler, and A. M. Young, Anatomy of STEM teaching in North American universities, *Science* **359**, 1468 (2018).
 - [32] M. Sundstrom, J. Gambrell, C. Green, A. L. Traxler, and E. Brewwe, Relative benefits of different active learning methods to conceptual physics learning, *Nat. Phys.*, **1** (2026).
 - [33] D. Z. Grunspan, B. L. Wiggins, and S. M. Goodreau, Understanding classrooms through social network analysis: A primer for social network analysis in education research, *CBE—Life Sci. Educ.* **13**, 167 (2014).
 - [34] Eric Brewwe, The roles of engagement: Network analysis in physics education research, in *Getting Started in PER* (American Association of Physics Teachers, College Park, MD, 2018), Vol. 2, [10.1119/RevPERv2.5.1](https://doi.org/10.1119/RevPERv2.5.1).
 - [35] Madelen Bodin, Mapping university students' epistemic framing of computational physics using network analysis, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010115 (2012).
 - [36] J. C. Speirs, M. K. R. Stetzer, and B. A. Lindsey, Utilizing network analysis to explore student qualitative inferential reasoning chains, *Phys. Rev. Phys. Educ. Res.* **20**, 010147 (2024).

- [37] Robert Beichner, *The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project* (American Association of Physics Teachers, College Park, MD, 2007), [10.1119/RevPERv1.1.4](#).
- [38] Robert J. Beichner, Jeffery M. Saul, Rhett J. Allain, Duane L. Deardorff, and David S. Abbott, Introduction to SCALE-UP: Student-centered activities for large enrollment university physics, in *Paper presented at the Annual Meeting of the American Association for Engineering Education* (ERIC, St. Louis, Missouri, 2000).
- [39] Robert J. Beichner and J. Saul, Student-centered activities for large-enrollment university physics (SCALE-UP), in *Proceedings of the Sigma Xi Forum on the Reform of Undergraduate Education* (Sigma Xi, The Scientific Research Society, Research Triangle Park, NC, 1999), pp. 43–52.
- [40] Robert Beichner, The SCALE-UP project: A student-centered active learning environment for undergraduate programs, *An invited white paper for the National Academy of Sciences* (2008).
- [41] Eugenia Etkina, Alan Van Heuvelen *et al.*, Investigative science learning environment—a science process approach to learning physics, *Res.-Based Reform Univ. Phys.* **1**, 1 (2007).
- [42] E. Etkina, D. T. Brookes, and G. Planinsic, The investigative science learning environment (ISLE) approach to learning physics, *J. Phys.: Conf. Ser.* **1882**, 012001 (2021).
- [43] E. Etkina, S. Murthy, and X. Zou, Using introductory labs to engage students in experimental design, *Am. J. Phys.* **74**, 979 (2006).
- [44] Eugenio Tufino, Pasquale Onorato, and Stefano Oss, Exploring active learning in physics with ISLE-based modules in high school, *J. Phys. Conf. Ser.* **2950**, 012021 (2025).
- [45] Lillian C. McDermott, Peter S. Shaffer *et al.*, *Tutorials in Introductory Physics* (Prentice Hall, Upper Saddle River, NJ, 2002), Vol. 2.
- [46] Noah D. Finkelstein and Steven J. Pollock, Replicating and understanding successful innovations: implementing tutorials in introductory physics, *Phys. Rev. ST Phys. Educ. Res.* **1**, 010101 (2005).
- [47] J. L. Docktor and José P. Mestre, Synthesis of discipline-based education research in physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020119 (2014).
- [48] <https://github.com/ibBukola/FidelityOfActiveLearningMethods>.
- [49] <https://sites.google.com/site/scientificabilities/>.
- [50] J. P. Adams, E. E. Prather, and T. F. Slater, in *Lecture-Tutorials for Introductory Astronomy* (Prentice Hall, Upper Saddle River, NJ, 2005).
- [51] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [52] S. Ramlo, Validity and reliability of the force and motion conceptual evaluation, *Am. J. Phys.* **76**, 882 (2008).
- [53] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33**, 159 (1977).
- [54] Cole Walsh, Daniyar Kushaliev, and Natasha G. Holmes, Connecting the dots: Student social networks in introductory physics labs, in *PER Conf.* (American Association of Physics Teachers, College Park, MD, 2020), pp. 557–562, [10.1119/perc.2020.pr.Walsh](#).
- [55] M. Sundstrom, D. G. Wu, C. Walsh, A. 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).
- [56] S. Al-Otaibi, N. Altwoujry, A. Alqahtani, L. Aldheem, M. Alqhatani, N. Alsuraiby, S. Alsaif, and S. Albarrak, Cosine similarity-based algorithm for social networking recommendation, *Int. J. Electr. Comput. Eng.* **12**, 1881 (2022).
- [57] H. M. Turner III and R. M. Bernard, Calculating and synthesizing effect sizes, *Contemp. Issues Commun. Sci. Disord.* **33**, 42 (2006).
- [58] C. Henderson and M. H. Dancy, Physics faculty and educational researchers: Divergent expectations as barriers to the diffusion of innovations, *Am. J. Phys.* **76**, 79 (2008).
- [59] https://www.physport.org/methods/Section.cfm?G=SCALE_UP&S=What.
- [60] Craig H. Blakely, Jeffrey P. Mayer, Rand G. Gottschalk, Neal Schmitt, William S. Davidson, David B. Roitman, and James G. Emshoff, The fidelity-adaptation debate: Implications for the implementation of public sector social programs, in *A Quarter Century of Community Psychology: Readings from the American Journal of Community Psychology* (Springer, Boston, MA, 2002), pp. 163–179, [10.1007/978-1-4419-8646-7_10](#).
- [61] John P. Kotter, Leading change: Why transformation efforts fail, in *Museum Management and Marketing*, edited by R. Sandell and Robert R. Janes (Routledge, London and New York, 2007), pp. 20–29, [10.4324/9780203964194](#).
- [62] K. Foote, A. Knaub, C. Henderson, M. Dancy, and R. J. Beichner, Enabling and challenging factors in institutional reform: The case of SCALE-UP, *Phys. Rev. Phys. Educ. Res.* **12**, 010103 (2016).
- [63] T. J. Lund, M. Pilarz, J. B. Velasco, D. Chakraverty, K. Rosploch, M. Undersander, and M. 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).
- [64] L. K. Weir, M. K. Barker, L. M. McDonnell, N. G. Schimpf, T. M. Rodela, and P. M. Schulte, Small changes, big gains: A curriculum-wide study of teaching practices and student learning in undergraduate biology, *PLoS One* **14**, e0220900 (2019).
- [65] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, *Am. J. Phys.* **71**, 607 (2003).
- [66] Lin Ding, Designing an energy assessment to evaluate student understanding of energy topics, Ph.D., North Carolina State University, 2007.
- [67] E. M. Bardar, E. E. Prather, K. Brecher, and T. F. Slater, Development and validation of the light and spectroscopy concept inventory, *Astron. Educ. Rev.* **5**, 103 (2007).