education sciences

MDPI

Article

Seeing the Forest by Seeing the Trees: Using Student Surveys to Measure Instructional Practices

Sandra L. Laursen * and Tim Archie

Ethnography & Evaluation Research, University of Colorado Boulder, Boulder, CO 80309, USA; tim.archie@colorado.edu
* Correspondence: sandra.laursen@colorado.edu

Abstract

Efforts to improve undergraduate education in mathematics and other STEM fields often work with instructors to implement research-based instructional practices that emphasize active and collaborative learning. To measure the progress and outcomes of such initiatives, researchers need measurement tools that are versatile, meaningful, and inexpensive to use, to know what teaching practices are occurring. Because students spend a great deal of time observing class conditions, they are well positioned to report the teaching that they experienced. We report results from some 2400 student surveys on the use of active and collaborative learning (ACL) approaches in over 200 recitation sections of gateway courses in tertiary mathematics, physics, and computer science. We developed a set of survey items, TAMI-SS, and a compound measure based on the items, called S-ACL for Student-reported Active and Collaborative Learning, that reflects the extent of active and collaborative learning as reported by students. We find that S-ACL scores compare favorably with instructor surveys and observations, and with students' reports of their classroom experience using established measures. Moreover, S-ACL reflected departments' progress in implementing ACL in recitations. When focused on specific, observable classroom behaviors, student surveys of instructional practice can be used to measure the progress of instructional change initiatives in mathematics and similar fields.

Keywords: departmental transformation; instructional change; research-based instruction; gateway course; recitation; active and collaborative learning; measurement; tertiary mathematics education

check for updates

Academic Editors: Wendy Smith, Amy Been Bennett and Rachel Funk

Received: 10 October 2025
Revised: 6 December 2025
Accepted: 11 December 2025
Published: 18 December 2025

Citation: Laursen, S. L., & Archie, T. (2025). Seeing the Forest by Seeing the Trees: Using Student Surveys to Measure Instructional Practices. *Education Sciences*, *15*(12), 1712. https://doi.org/10.3390/educsci15121712

1. Introduction

By now, it is generally accepted that there are more effective approaches to tertiary STEM instruction than "continuous exposition by the teacher" (Bruff, 2015). While these improved approaches have a variety of names, they fit broadly under the umbrella of "active learning," where students learn through class activities and discussions that emphasize higher-order thinking and often involve group work. Despite variability among approaches and implementations, meta-analyses suggest that these approaches generally improve students' exam scores and course grades, relative to teaching that is heavy on instructor exposition (Freeman et al., 2014; Ruiz-Primo et al., 2011). When they are well structured, these approaches also tend to provide greater support for students from historically marginalized groups to achieve comparable academic success to peers from historically privileged groups (Theobald et al., 2020).

As Fairweather (2008) pointed out years ago, availability of vetted teaching and learning methods is not the bottleneck in applying active learning approaches to undergraduate STEM teaching. In mathematics, for example, instructors can draw on a variety of published curricula and well-described methods to adapt for their own courses. The bottleneck to widespread use of active learning is still instructors' willingness, readiness, and capacities to take this step, importantly combined with the institutional support they need to do it (Laursen et al., 2019). Some strategies can help: Recent research is demonstrating that individual instructors' practices can change significantly in response to substantive, well-designed professional development programs in their discipline (e.g., Archie et al., 2022; Manduca et al., 2017). And growing numbers of departments are engaging in collective efforts to strengthen teaching and learning for all students, especially in gateway courses, in mathematics (e.g., Cronin et al., 2025; W. M. Smith & Rasmussen, 2025; W. M. Smith et al., 2021) as in other STEM fields (e.g., Corbo et al., 2015; Greenhoot et al., 2020; Macdonald et al., 2019; Peteroy-Kelly et al., 2019; Wieman, 2017).

As collective change efforts take hold in disciplines and departments, there is a concomitant need for evaluation tools to measure collective progress in instructors' uptake of active learning approaches, and more generally to characterize teaching. A decade ago, a group of experts who were convened to discuss this problem (AAAS, 2013) identified a core set of measurement techniques: faculty surveys; faculty interviews; course artifacts or portfolios; student reports through surveys or interviews; and classroom observations. They identified strengths and limitations of each technique, recognizing differences in the suitability of each for different goals, such as documenting practice (e.g., in evaluating individual instructors for retention or promotion), improving practice (e.g., in recording improvements for institutional accreditation or assessing needs for professional development), or for research and evaluation (e.g., in linking specific teaching practices to student learning). For example, faculty surveys have long been used by the Higher Education Research Institute to track changes over time in college instruction (e.g., Eagan, 2016). In recent years, many scholars have developed new instruments, especially classroom observation protocols (e.g., Eddy et al., 2015; Hora, 2015; Hora & Ferrare, 2013; M. K. Smith et al., 2013) and faculty surveys (e.g., Landrum et al., 2017; Walter et al., 2016), with the intent to probe instructor practices in higher education and connect those practices to student outcomes. Our team has developed similar tools—an instructor survey and observation protocol, aligned to each other—for measuring instructional change as a result of professional development (Weston et al., 2021, 2023; Archie et al., 2022; Laursen et al., 2011, 2014; Laursen, 2024). In contrast, across the field, much less work has been done to draw on data from students to document the teaching they experience.

In this paper, we describe an approach to characterizing teaching from student survey data and illustrate its use in monitoring an institutional change initiative. Drawing on our experience with observations and instructor surveys, we developed a brief student survey to characterize features of the teaching that occurred in recitations of large courses in mathematics and two math-adjacent departments. We developed this survey in our role as evaluators of an institutional project to infuse active and collaborative learning (ACL) into those recitations. Importantly, we did not survey students primarily to measure outcomes, such as attitudes or learning, that might arise from the instructional reforms. Rather, we capitalized on students' everyday presence in the classroom as observers who could describe aspects of the teaching they directly experienced. We used their reports to characterize teaching across a large sample of class sections, then to measure change in teaching over time, as the initial desired outcome of the institutional project. The student survey focuses on the frequency of specific teaching practices and is aligned with other instruments that we have developed. Alongside this measure, we also probed students' perceptions of the classroom environment and a small set of student self-reported learning gains, to assess whether these experiences and outcomes were related to the teaching practices that students described, in line with expectations based on the literature.

Our title plays on a common metaphor, "can't see the forest for the trees," referring to someone who doesn't grasp the overall situation—represented by the forest—because they are so focused on the details, represented by the trees. Here, in contrast, we focus on the trees (the students) so that we can learn about the forest (the class): no single student has an authoritative perspective on the teaching taking place, but all can contribute information to a composite picture. Then, when we combine their many perspectives, we can say something about the class as a whole. The underlying philosophy of the measurement approach discussed here is that we can see the forest *because* we can see many trees.

This study sought to understand whether and how a student survey measure can be useful in characterizing teaching for evaluation and research studies, especially those seeking to measure change in teaching over time. We gathered three lines of evidence that speak to the construct validity and utility of the student survey measure. Our approach is nomological in philosophy, meaning that it emphasizes the relationships among our proposed new construct and pre-existing constructs that we expect, from theory, to be related (Cronbach & Meehl, 1955). First, we compared the student survey measure of teaching to established measures of student experience that we expect will be influenced systematically by the teaching the students received. We then compared the student survey measure to data that were obtained from other, independent methods of measuring teaching, and that we expect will yield similar characterizations. Finally, we examined the student survey data across groups of courses and over time. Because we know when and how the reform intervention had been applied to those courses, we expect particular patterns of differences over time and among groups. In each case, we are predicting and testing the relationships among constructs. In this paper, we describe results that speak to three research questions:

1. How does student-reported use of ACL practices relate to (a) their self-reported experience of the learning environment and (b) their self-reported learning gains?
2. How does student-reported use of ACL practices relate to instructor practices, from (a) instructor self-report and (b) observation?
3. How does student-reported use of ACL practices function as a metric for evaluating broad differences in teaching practice by groups and over time?

We address these questions using findings from a specific institutional study involving over 2400 student survey reports, at two time points, from the recitation sections of large introductory courses in mathematics, physics, and computer science, complemented by observations and instructor surveys from the same courses. We then discuss the general utility of student-reported teaching data for formative and summative evaluation of an institutional change project.

2. Rationale and Prior Work

In this section, we provide an overview of faculty surveys and observations as commonly used methods for studying teaching. The use case of interest to us is to measure change in selected characteristics of teaching after professional development or other interventions that promote instructors' use of research-based instruction. We then consider what is known about student surveys in comparison to instructor surveys and observations.

Instructor surveys are often used for measuring change in teaching because they can be efficiently administered to large numbers of people, such as participants in a professional development program or members of a department who are working on changing their teaching in some way. However, survey use poses challenges that include sampling, internal validity, and potential biases in instructor self-report, as Iverson (2012) discusses. Social desirability bias—perceived pressure to conform to others' expectations—is particularly concerning in self-reports of teaching when the goal is to make valid comparisons among courses or over time (e.g., Miller & Dumford, 2017). Iverson also points out the need for validation evidence for faculty survey instruments and emphasizes the potential of mixed methods study designs that support triangulation of survey data with other data. Williams et al. (2015) reviewed properties of several instruments for instructor self-report, many of which remain popular or have spawned variants. They assessed the strengths and limitations of the instruments, individually and collectively, and offered guidance on selecting an instrument.

Classroom observation is less well developed as a research method in higher education than in K-12 settings (Campbell, 2017), but is increasingly used by higher education researchers and evaluators for purposes such as evaluating outcomes of instructor professional development or assessing instructor use of research-aligned teaching practices. Observations offer richly detailed, naturalistic information about teaching, and can document behaviors of both students and instructors in real time as they unfold, but they are time-consuming and costly. Hora and Ferrare (2013) and Campbell (2017) have discussed strengths and limitations of classroom observation in higher education, with particular attention to validity and reliability concerns, including rater and inter-rater reliability, and the tradeoffs between structured and unstructured protocols. Weston et al. (2021) discuss additional challenges to reliability when time sampling is applied to make inferences about term-long courses. Anwar and Menekse (2021) review some of these instruments for their content focus, content validity, and practical features.

When student data are used in higher education studies, they are most often used to measure student response to their classroom experiences, or to assess learning and affective outcomes. For example, Wiggins et al. (2017) used surveys to examine two closely proximal outcomes immediately after a class session: students' engagement in specific activities and their self-assessed understanding of the concepts taught. Other studies have used end-of-course surveys to assess broader student responses to teaching, such as what aspects of the course helped students learn or not (e.g., Achen & Lumpkin, 2015; Borda et al., 2020; Jones, 2017), or to measure course-level outcomes such as learning (e.g., Andrews et al., 2011; Elliott et al., 2016) or attitudes (Achen & Lumpkin, 2015; Chiu et al., 2022; Elliott et al., 2016).

However, this focus on outcomes represents a missed opportunity. Because students regularly spend time in classrooms, they can report back on what kind of teaching is actually taking place. This is importantly different from rating the quality or impact of that teaching. Indeed, in evaluative settings such as institutionally administered end-of-course questionnaires, student ratings of teaching are often derided as popularity contests, and researchers identify known, serious biases in how students rate teachers of varied gender, race and ethnicity that can lead to inequities in personnel decisions based on such ratings (Kreitzer & Sweet-Cushman, 2021; Heffernan, 2022). But Benton and Cashin (2014) argue, based on their detailed review of the literature, that "student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control, perhaps more so than any other data used for faculty evaluation" (p. 315). These characteristics suggest that student reports of teaching can offer valuable insight to studies of teaching change, when they are decoupled from teacher evaluation and when they are used in combination with other methods. In the K-12 arena, the Measures of Effective Teaching [MET] Project (2012) found clear connections between student reports about their teachers' practices and student achievement gains, highlighting the value of student reports within a larger portfolio of teaching measures. The key, they suggest, is asking "the right questions, in the right ways" (p. 2) and administering surveys in ways that minimize threats to validity.

Similarly, student reports of college teaching have been shown to align with instructor reports and observations, indicating that students can reliably estimate the frequency of teaching practices they experience in the classroom (Durham et al., 2018).

We have identified relatively few studies in higher education that have used student reports to characterize teaching methods. Rather, most studies using student survey data have examined student responses to or perceptions of the teaching they receive, or their reports of personal learning. Exceptions include an elegant study by Jones (2018) in the geosciences, who used student surveys to assess whether and how much students had experienced particular types of learning assignments, both in a specific course that they took, and in their major more generally, at graduation. Working with a major departmental initiative to improve teaching, and knowing which types of assignments were espoused and supported by the teaching initiative, Jones could detect increased prevalence of these strategies in students' reports over time, as students passed through the reformed courses and as more courses were reformed—thus demonstrating the reform initiative's impact on teaching practice across a department. Two studies of gateway courses in college mathematics adapted the faculty-directed PIPS survey (Walter et al., 2016) to ask students what percent of class time they spent on four key teaching strategies (Bressoud et al., 2015; Street et al., 2021). Various analyses used these items to measure teaching, for example, as a global indicator to characterize the overall prevalence of lecture (Larsen et al., 2015), and as a dependent variable in an analysis relating instruction to student attitudes (Apkarian et al., 2024). Finally, Durham et al. (2017) used student reports of teaching practices in biology and other STEM courses to detect course-level differences in teaching practices that reflected instructors' participation in instruction-focused professional development. Collectively, these examples show that student reports of teaching have much promise as measures of what teaching practices are taking place and how they may vary by context or time, in mathematics courses and in other fields.

3. Context for the Study

The study context was offered by Gateway2STEM, an NSF-supported project at George Mason University (GMU). The project sought to incorporate active and collaborative learning (ACL) into gateway courses in mathematics, physics and computer science (CS), focusing on recitation sections. Gateway courses are high enrollment courses that are foundational to a variety of educational pathways and high-risk for students (Koch, 2017). They have an outsized role in shaping students' paths: they may be gateways that open paths to STEM majors and careers, or to other majors that require STEM courses, or may serve as gatekeepers that restrict students' progress toward their own educational and career goals. In this study, the gateway courses were two two-semester calculus sequences, two two-semester sequences in introductory physics, and an introductory programming course, initially one semester long. In 2023, a new two-term sequence in CS was added to the original single-semester version and included in the study.

To develop ACL approaches suitable for their courses, all three departments gathered instructors in course-based communities of transformation (CCTs) to plan and implement an approach to ACL within recitations, and in some blended lecture-recitation courses (see Bryan et al., 2025; Bulancea et al., 2021; for descriptions of the CCT in mathematics). At GMU, recitation meetings offer a separate block of instruction that supports a lecture section. These 50-min sessions offered students review, practice and personal help with course material in smaller groups than the large lecture period, which met separately for 100–150 min. Blended classes did not differentiate lecture periods from recitation periods; most used a mixture of lecture-based and interactive modes during longer class blocks of 75 min twice a week. Certain mathematics and CS courses, and some physics sections, used the blended format.

The focus on recitations reflects both pragmatism and optimism in the project team's approach. Recitations were seen as a foot in the door to give instructors concrete experiences with ACL approaches, develop their expertise and interest, and build local teaching cultures that might spur broader changes. Implementing ACL in large courses was made more feasible through the involvement of graduate assistants, who teach recitation sections in mathematics and CS, and undergraduate learning assistants, used in all three departments in most of the targeted gateway sections. The CCTs chose a guided inquiry approach, where students would work in small groups on a sequence of tasks, variously referred to as worksheets (mathematics), tutorials (physics), or programming assignments (CS). Common tasks were prepared and recitation instructors were encouraged to use them; in practice, there was variation in whether instructors used the common tasks, adapted them, or wrote their own.

GMU is a large public institution with about 23,000 undergraduate students. Many are non-traditional students who are working, studying part-time, resuming education after military service, or preparing for a second career; many also represent the first generation of their families to attend college. The students are ethnically diverse and represent the population of the large metropolitan area where the university is located. The study samples reflected this diversity of backgrounds among GMU undergraduates.

Collectively, the gateway courses in this study served about 3000 students each fall term. Table 1 shows details of the courses and recitation sections and reflects both some homogeneity and some variability in the course contexts for ACL implementation.

Table 1. Characteristics of Gateway Course Recitation Sections Studied.

Course Characteristic	Subsample	Mathematics	Physics	CS	All
Gateway courses included in study	2021	4	4	1	9
	2023	4	4	3	11
Total # sections, recitation + blended	2021	41	41	30	112
	2023	42	34	18	94
Total enrollment	2021	1401	959	653	3013
	2023	1511	1051	903	3465
Lead instructors	rec *	GTAs	Term and tenured faculty	GTA pairs	
	blended	Term faculty	Term faculty	Term faculty	
Class length	rec	50 min (2 courses)	50 min	50 min	
	blended	75 min (2 courses)		75 min (1 course, 2023)	
Max. class size	rec	45	30	25 (2021) 60 (2023)	
	blended	35		60	

* rec denotes standalone recitations and is contrasted with blended courses as described in text.

We define some terms here as used in Table 1 and throughout the paper. CS used the term "labs" to describe the support sessions attached to their lecture courses, but here we refer to all these sections as *recitations*, to communicate that they served similar instructional purposes and to avoid confusion with science laboratory work. We use *instructors* to name all those with a classroom teaching role in recitations. Instructors leading recitation sections include *faculty* who may hold tenured or teaching-focused "term" appointments, and *graduate teaching assistants* (GTAs). Some courses also included undergraduate *learning assistants*, who were called *undergraduate teaching assistants* (UTAs) in CS. Like LAs in mathematics and physics sections, UTAs typically circulated during group work to answer questions and promote students' own thinking on in-class tasks. At the time of this study, GMU's LA program was robust and offered specialized training for LAs and UTAs consistent with research-based principles for such programs (Otero, 2015).

4. Materials and Methods

In this section, we provide details of the survey, observation and analysis methods.

4.1. Study Sample

The study sample was defined by the set of recitation sections (standalone recitations plus blended courses) offered for gateway courses in mathematics, physics, and computer science that were targeted by the Gateway2STEM project. The number of courses and sections in 2021 and 2023 is included in Table 1. The survey samples include college students and instructors of these sections, and the observation sample was drawn from sections taking place in the fall term of 2023. Student surveys were administered from mid-November to mid-January, and instructor surveys from mid-December to mid- or late January. Fall 2023 sections were observed in November. Written consent was obtained from survey participants as they began the survey.

4.2. Survey Instruments

The starting point for the study was an instructor survey we called the TAMI-IS, *Toolkit for Assessing Mathematics Instruction—Instructor Survey*. This set of items has been used before to probe instructors' teaching practices and measure change therein (Archie et al., 2022; Laursen et al., 2023) and aligns well with a related measure for classroom observations by external observers (Hayward et al., 2018; Weston et al., 2023). The core of the TAMI-IS instrument is a set of quantitative items in which instructors are asked to estimate the total percentage of time they spend on different teaching strategies, exclusive of exam time. Table 2 shows the list of teaching activities probed and the brief definitions provided to survey respondents.

Table 2. Teaching Activities and Definitions Provided in TAMI-SS Items.

Activity	Definition
Formal lecture	Little or no question and answer with students
Interactive lecture	Instructor asks questions during lectures; students answer with short or extended responses
Short explanations	Short explanations or summary by instructor (typically less than 5 min)
Whole class discussion	Students respond to peers in discussing course topics
Group work	Students work together in groups on problem sets or other activity
Individual work	Students work alone on problem sets or other activity
Student presentations	Students present or talk through problems on the board, projector, or online workspace
Assessment	Quiz or other assessment

Instructors responded with numerical estimates that were constrained to add up to 100% of class time. They could write in up to two additional practices labeled "other," but seldom did so. Additional multiple-choice and open-ended items asked instructors about other practices (e.g., assessment, technology), personal knowledge, skills, and beliefs about active learning, and departmental norms. These were used for project evaluation and are not discussed here.

The student survey was designed to be short and focused on what students experienced directly. The median duration for student responses in Qualtrics was less than 10 min. We used three blocks of items, which are summarized in Table 3 and described further below, and gathered some demographic indicators.

Table 3. Overview of Survey Item Blocks Used in this Study.

	Use of ACL Practices: TAMI-SS (Hayward et al., 2018; Weston et al., 2023)	**Learning Environment: WIHIC (Dorman, 2003; Hoover & Pelaez, 2008; Skordi & Fraser, 2019)**	**Learning Gains: SALG (Seymour et al., 2000; Laursen et al., 2011)**
Nature of items and scale	frequency of use for 5 practices: lecture, small group work, individual work, whole class discussion, student presentation How often in class? 0 = never, 1 = seldom, 2 = sometimes, 3 = often, 4 = almost always	frequency of experiences in 7 areas: student cohesiveness, teacher support, involvement, inquiry/investigation, task orientation, cooperation, equity (8 items each) How often in class? 0 = almost never to 4 = almost always	rate learning gains in 5 areas common to all 3 disciplines 0 = no gain to 4 = great gain rate "what helped me learn" for a variety of course features and activities 0 = no help to 4 = great help
Composite measure	S-ACL: mean for 3 practices—small group work, whole class discussion, and student presentation (range 0–4)	WIHIC scale: mean (range 0–4)	Gains scale: mean (range 0–4)

4.2.1. TAMI-SS, Student Survey

Items were adapted from our TAMI-IS (Instructor Survey) to measure the frequency of use of teaching practices, as perceived by students. Students were asked "How often does the following occur?" and responded to the same list of strategies and definitions as did instructors (Table 2), with the omission of quiz time and the write-in option. They rated how often they experienced each of the teaching practices on a five-point scale, "never" to "almost always."

4.2.2. What Is Happening in This Class (WIHIC)

The items in this survey measure student perceptions of the learning environment, with 8 items in each of 7 scales listed in Table 4. The instrument has been used in many published studies (e.g., Dorman, 2003; Fraser, 2023; Hoover & Pelaez, 2008; Skordi & Fraser, 2019), and we have found the WIHIC to be informative about classroom environments in other evaluation studies.

Table 4. WIHIC Scales and Sample Items.

WIHIC Scale	**Scale Description**	**Sample Items**
Student cohesiveness	Extent to which students know, help and are supportive of one another	I make friends among students in this class. In this class, I get help from other students.
Teacher support	Extent to which the teacher helps, befriends, trusts and shows interest in students	The teacher checks in with me. The teacher's questions help me to understand.

Table 4. *Cont.*

WIHIC Scale	Scale Description	Sample Items
Involvement	Extent to which students have attentive interest, participate in discussions, perform additional work and enjoy the class	I give my opinions during class discussions. I explain my ideas to other students.
Inquiry/investigation	Extent to which the skills and processes of inquiry and their use in problem solving and investigation are emphasized	I try new approaches to test my ideas. I explain the meaning of statements, diagrams and graphs.
Task orientation	Extent to which it is important to complete activities planned & to stay on the subject matter	I know the goals for this class. I pay attention during this class.
Cooperation	Extent to which students cooperate with one another on learning tasks	I learn from other students in this class. When I work in groups in this class, there is teamwork.
Equity	Extent to which the teacher treats students equitably	I am treated the same as other students in this class. My work receives as much praise as other students' work.

4.2.3. Student Assessment of Their Learning Gains (SALG)

The SALG asks students to assess and report on their own learning and on the degree to which specific aspects of the course have contributed to that learning (Seymour et al., 2000). For this study, our shortened version of this instrument asked students about a few learning gains from the course, generalized so that the same gains could be probed in mathematics, physics and CS. We have previously used these items to document students' self-reported learning outcomes and to understand what may account for linkages between student outcomes and their teaching and learning experiences (e.g., Laursen et al., 2011, 2014). Additional SALG items probing what aspects of the course helped student learning, such as types of class and lab activities, assessments, and resources, were used for project evaluation and are not discussed here.

For all three blocks of student measures—TAMI-SS, WIHIC, and SALG—we asked the GMU PI team to check our assumptions, note disciplinary differences, and identify language that would make sense to GMU students, as a safeguard to content validity. For example, slight modifications were made to adapt WIHIC items for the forms of inquiry, investigation and problem-solving typically used in each discipline. Questions for CS students asked about "lab" instead of "recitation" experiences, following the department's terminology for the non-lecture periods. All such changes were minor language adjustments (e.g., using "inquiry" in mathematics and "investigation" in physics); no items were removed, substituted or combined, and we did not separately validate the modified items.

Finally, identity-related items were collected to learn whether responses varied among demographic subgroups, especially among student groups that have often faced higher risk of non-completion in STEM courses and degree programs. These items asked students' class year, transfer student status, parental education, gender identity, race and ethnicity, international status, and current number of courses.

4.3. Survey Data Collection

The online survey was administered using Qualtrics near the end of the fall term in both 2021 and 2023. Surveys were distributed by recitation instructors using a customized Qualtrics URL and scannable QR code that could be emailed, posted to the course LMS, and/or projected on a slide. We provided text for instructors to share with students to explain the survey and suggested that they provide 10 min in class for students to respond, and we sent survey reminders by email or through the project director in each department.

Across courses, 90–95% of students reported attending recitation “most” or “all” of the time. Thus, their reports of what happens in class are expected to reflect cumulative experience. In all, 2424 students responded at least in part. Overall student response rates were 36% in 2021 (914 useable responses) and 46% in 2023 (1510 useable responses). In the combined data set from both years, 1088 students responded at least in part, and 754 reached the end of the survey (this does not mean they answered all or any questions). The largest number of usable responses was 827, on WIHIC items. For most analyses, we combine the two years of data, to answer questions about the survey items’ behavior and meaning. For answering RQ3, we separated the data by time.

About 33% of respondents used the QR code to respond to the survey; this and the clustering of time stamps on the student surveys confirm that some (but not all) sections were given class time to respond. These sections are better represented in the sample than those not given class time to respond. Estimated response rates by section ranged from 96% to 0, varying greatly by how the survey was shared and communicated with students. Typically, all recitations taught by the same instructor had similar response rates. Because of this bias in response rate, we do not directly compare courses or sections.

In general, the student demographics reflected GMU’s population and showed differences in course enrollment that were already known to the PI team and CCTs, and that are common in STEM courses. For example, women were more concentrated in math and physics sequences that required lower levels of mathematics preparation, and they were underrepresented in CS. Part-time and transfer students were also more concentrated in math and physics sequences that required lower levels of mathematics preparation.

4.4. Observation Methods and Analysis

In fall 2023, one author (S.L.) visited the campus and observed 22 class sections targeted by the project, including 16 recitations (taught by GTAs and faculty) and 6 blended sections (taught by faculty) across all three departments. Sections were chosen to maximize the observer’s in-class time across four on-campus days, making sure to sample each course in the study. The field notes included the stop and start time for different activities (e.g., lecture, instructions, quiz, group work), using the same base activities as the surveys (Table 2). The field notes also captured attendance, visual indicators of student diversity, instructor and student behaviors and interactions, the flow of class activities, and verbatim examples of instructor and student remarks. Qualitative analysis included repeated holistic reading and annotation of the notes about each section, guided by literature on effective group work as a focus of the institution’s ACL intervention (e.g., Hodges, 2018; Johnson et al., 2014; Liljedahl, 2020). Some of these data were used for project evaluation and are not reported in detail here, but we used the field notes to assess the general level of student engagement and participation, especially during group work.

Here we focus on the use of class time, which is directly comparable to survey data, and not on other aspects of instruction. We computed elapsed minutes and percentage of class time used for different activities, and computed simple descriptive statistics. We sum the times for individual “seat work” and group work because in some sections, students had the choice to work alone or with others. Given small sub-sample sizes, we did not separate observations by department or course, to protect instructor identities.

4.5. Analysis Methods and Creation of S-ACL Score

For analysis, we simplified the categories for all measures, student survey, instructor survey, and observation. Our survey instruments distinguish types of instructor talk by their function: *lecture* is pre-prepared instructor talk, distinct from short, responsive *explanations* such as signposting or mini-lectures that are directly prompted by student questions or difficulties. Variants of lecture include formal lecture (with little or no exchange of questions and answers, Q&A, with students) and informal lecture (with some Q&A). Most instructors used either formal or interactive lecture; a few used segments of each. In general, we combined the two forms of lecture for analysis, and we do not rely on these numerical estimates. In practice, we find that instructors do not respond consistently about formal and interactive lecture, and often overestimate how interactive their lectures are. Thus, in instructor responses it is most accurate to combine all types of instructor talk. Students tend to consider all instructor talk to be "lecture," thus offering higher estimates of lecture time, whereas in observations we distinguish class business such as making announcements and handing back papers, using the category "other" for time that is not instruction and is unavailable for student interaction. This evident complexity of defining "lecture" consistently is one reason we prefer to use a dependent variable that is based on non-lecture time.

With this perspective, we computed three composite variables. The student ACL score (S-ACL) is the mean frequency of the three ACL strategies: class discussion, student group work, and student presentation, as reported by one student (scale 0–4). The section ACL score (section S-ACL) is the mean S-ACL score for sections with at least 10 responses. The instructor ACL score (I-ACL) is the total percentage of class time for the same three activities (scale 0–100%). For each, higher values indicate more frequent use of the ACL strategies.

Internal reliability of WIHIC subscales was verified with all subscales showing a Cronbach's alpha of 0.9 or greater. Independent sample t-tests were used for statistical comparisons of courses targeted or not by the intervention. For most analyses, we combined data across both years to increase statistical power and learn about the behavior of the item. To answer RQ3, we separated data sets by year. We conducted a series of initial analyses to test for differences in ACL and WIHIC scores by student characteristics including gender, race, ethnicity, first-generation college student, and transfer status. As we found no statistically significant differences by any student characteristics, subsequent statistical analyses did not include these variables.

5. Results

In this section we present descriptive statistics for the three scales listed in Table 3, then discuss results relevant to each of the research questions RQ1–RQ3.

5.1. Descriptive Statistics

Table 5 provides the mean and average standard deviation for each composite item. ACL practices were reported to have "sometimes" occurred on average and the distribution was moderately platykurtic, while student-reported WIHIC items to have occurred "seldom" to "sometimes" on average, and the scale was negatively skewed. Students reported "moderate" to "good" gains on average and this scale was moderately negatively skewed.

Table 5. Descriptive Statistics for S-ACL, WIHIC, and Gains.

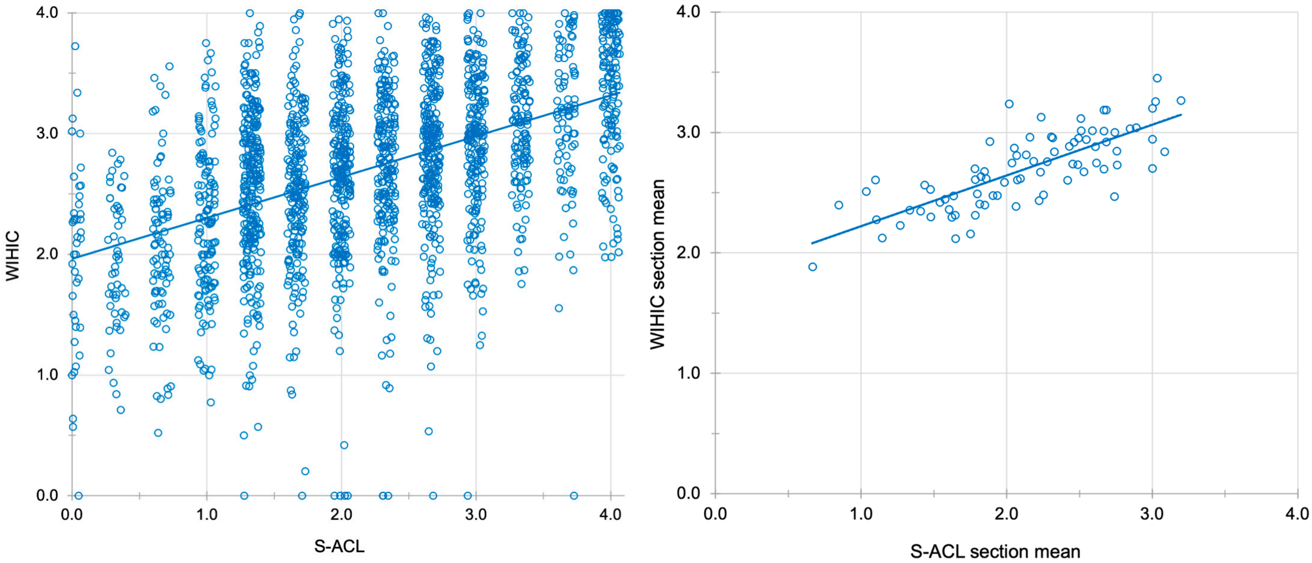
Scale	N	Mean	SE	SD	Skew	Skew SD	Kurtosis	Kurtosis SD	Alpha Reliability
S-ACL	2056	2.15	0.02	1.04	−0.07	0.05	−0.68	0.11	NA *
WIHIC	2236	2.70	0.01	0.70	−0.48	0.05	0.17	0.10	0.97
Gains	1973	2.82	0.02	0.92	−0.64	0.06	−0.07	0.11	0.91

* Cronbach's alpha was not calculated.

Cronbach's alpha coefficient was used as a measure of the internal consistency for Gains and WIHIC scales. ACL practices are discrete, mutually exclusive practices that can be used independently, thus there is not an expected relationship or response pattern. Alpha coefficients for Gains and WIHIC scales were from 0.91 and 0.97, respectively, indicating high reliability.

5.2. Results for RQ1, Relationship of Student-Reported ACL to Their Experiences

To answer RQ1(a), "How does student-reported use of ACL practices relate to their experience of the learning environment?" we tested the relationship between S-ACL and WIHIC scales. From a correlation analysis of all student cases, we found a moderate positive correlation (0.51) between S-ACL and WIHIC, indicating that students reported more positive environments in courses that they also perceived to include more active learning (see Figure 1). Next, we tested the consistency of this relationship across disciplines and found similar, low to moderate correlations across math, physics, and CS, with some variation (0.40, 0.54, and 0.33, respectively).



(**a**) (**b**)

Figure 1. WIHIC score correlates positively with S-ACL, shown for (**a**) individual students and (**b**) section means (N $\geq$ 10).

Lastly, we conducted a linear mixed model to test this relationship with WIHIC scale as the dependent variable, S-ACL as a fixed effect, and course section as a second-order random effect. Controlling for the influence of section-level data is necessary as section responses ranged from 1 to 169, which would bias the correlations based on student reports from larger sections. To do this, we excluded student responses from sections with fewer than 10 responses, as small class sets have shown to be less reliable than larger class sizes (Benton & Cashin, 2014). The resulting analysis included 1640 student observations from 93 course sections. This analysis revealed a moderate positive effect of ACL practices (captured by S-ACL) on WIHIC reports (0.33, t (73) = 18.64, $p < 0.001$) when controlling for the effect of course sections.

RQ1(b) asks, "How does student-reported use of ACL practices relate to students' self-reported learning gains?" To address this question, we tested the relationship between student reports of the ACL practices they experienced and their perceived learning gains. A correlation analysis of all student cases revealed a statistically significant moderate positive correlation (0.25) between S-ACL and Gains, indicating that students report more learning gains in courses that include more active learning. Next, we tested the consistency of this relationship across disciplines and found similar minimal correlations for math and physics students, 0.25 and 0.27, respectively. We found no statistically significant relationship for CS.

We then conducted a linear mixed model to test this relationship with Gains as the dependent variable, S-ACL as a fixed effect, and course section as a second-order random effect. Again, we excluded student responses from sections with fewer than 10 responses. This analysis revealed a moderate positive effect (0.23, t (86) = 8.93, $p < 0.001$) of ACL practices on student reports of their learning gains when controlling for the effects of class sections.

5.3. Results for RQ2, Relationship of Student-Reported ACL to Instructor Practice

RQ2 asks, "How does student-reported use of ACL practices relate to instructor practices?" We have two sources of evidence about instructor practice to examine this question, instructor surveys in which they report their own practices, and external observations of some sections.

First, we tested the relationship between student- and instructor-reported use of ACL practices, by instructor and course section. For sections where we had both instructors' self-report and at least 10 student responses (n = 28), we compared the section mean of S-ACL for each recitation section with the corresponding instructor response, I-ACL (Figure 2). A correlation analysis of mean section S-ACL responses and instructor responses revealed a statistically significant strong positive correlation (0.60, $p < 0.05$), indicating that students' and instructors' reports of the ACL practices in a given class section are well aligned. This relationship is shown in Figure 2.

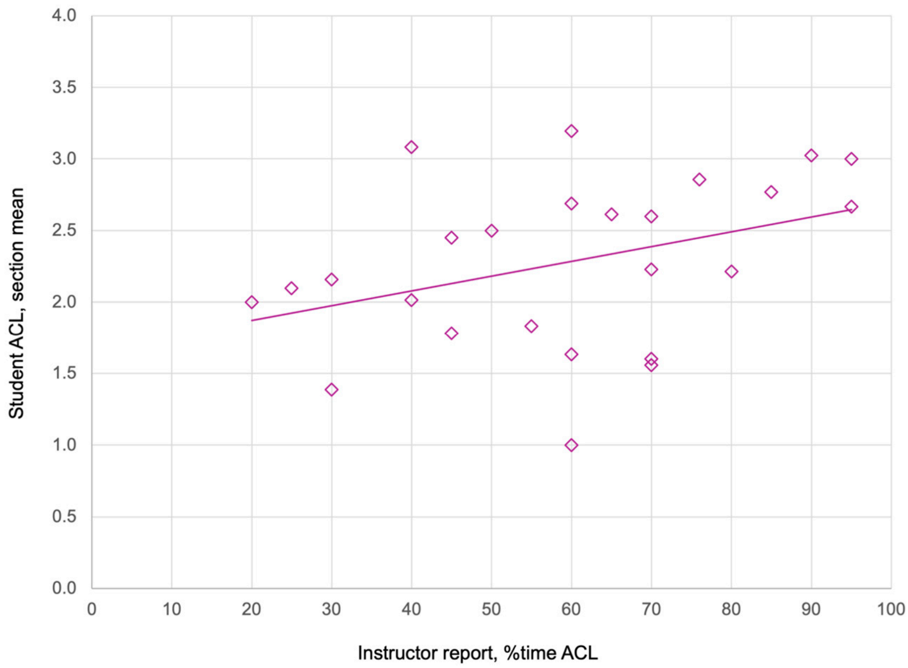


Figure 2. Section S-ACL correlates positively with I-ACL for sections where both types of data are reported.

Similarly, we tested the relationship between instructor-reported I-ACL, and the observed proportion of time spent in ACL practices from a single observation (Figure 3). The data set includes 16 sections where we had the section instructor's self-report.

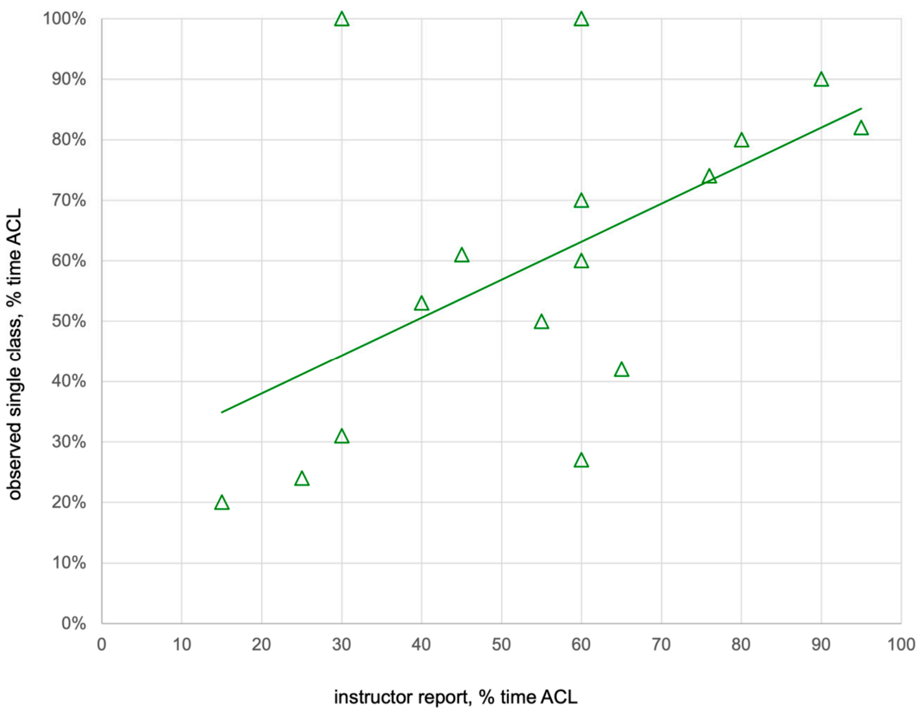


Figure 3. Observed percentage of ACL class time correlates positively with I-ACL for sections where both types of data are reported.

We confirmed that these observation subsets represent the larger data set well. For example, the mean for the full set of 22 sections is 62% time spent in ACL, the same as for the subsets where we have multiple data types. Differences between the two class types, 70% of time spent in ACL for recitations and 46% for blended sections, reflect differences in the purpose of these class blocks. In general, there is good though non-significant correlation between what instructors say they do, and what we saw them do (0.41, $p = 0.16$). The high outliers in Figure 3 are two sections led by GTAs; both used the whole of the observed class period for a guided inquiry task, but reported using much less time for ACL on average.

5.4. Results for RQ3, Use in Comparing Groups for Program Evaluation

Research Question 3 asks, "How does student-reported use of ACL practices function as a metric for evaluating differences by groups and over time?" Tables 6 and 7 summarize S-ACL means over time and by department and indicate the magnitude of the changes for various comparisons by group and/or by time. These results are graphed in Figure 4.

Table 6. Comparison of Mean S-ACL Scores by Survey Year and Department, using all 2021 Target Courses as the Standard.

	2021 Target		Comparison				
Comparison	**M**	**SD**	**M**	**SD**	***t***	***p***	**Cohen's d**
2021 target vs. 2021 nontarget	2.33	1.04	1.24	1.04	11.84	<0.001	1.05
2021 target vs. all 2023	2.33	1.04	2.17	0.98	3.29	<0.001	0.16
2021 target vs. MATH 2023	2.33	1.04	2.62	0.87	4.48	<0.001	0.29
2021 target vs. PHYS 2023	2.33	1.04	2.12	1.01	3.48	<0.001	0.20
2021 target vs. CS 2023	2.33	1.04	1.74	0.84	8.93	<0.001	0.61

Table 7. Comparison of Mean S-ACL Scores by Department, Over Time.

Comparison	2021 M	2021 SD	2023 M	2023 SD	t	p	Cohen's d
Mathematics	2.54	0.91	2.62	0.87	1.31	0.190	0.10
Physics	1.82	1.15	2.12	1.01	4.32	<0.001	0.29
CS	0.89	0.97	1.74	0.84	6.33	<0.001	0.99

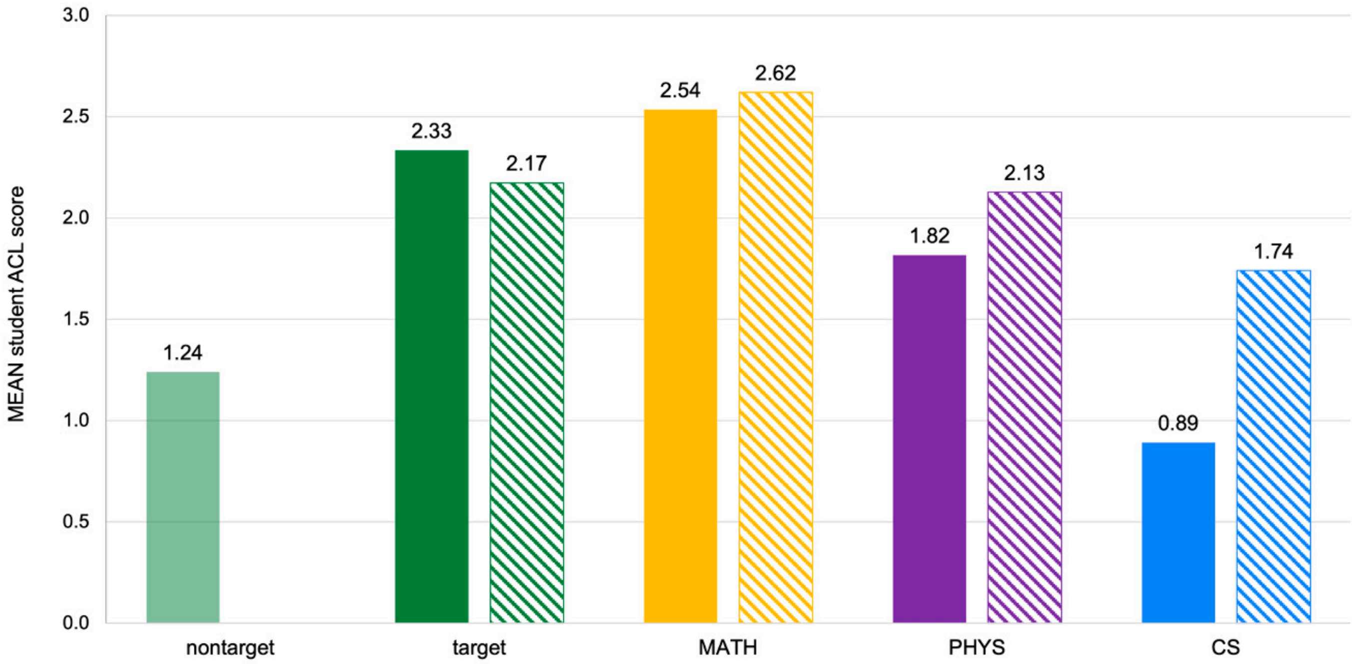


Figure 4. Project-wide and department means for S-ACL in 2021 (solid), and 2023 (stripes), for target and non-target courses and by department.

Figure 4 shows two types of comparisons that use the S-ACL as a metric for project evaluation. On the far left is a time-based comparison, mean S-ACL for all target courses that were actively involved in the project versus nontarget courses that had not yet implemented ACL in 2021. Here the S-ACL metric clearly distinguishes courses that had (dark green) and had not (light green) implemented ACL, with an effect size of 1.05.

Figure 4 also compares means for 2021 (solid) and 2023 (stripes), for the total sample and for each department's set of gateway courses. Because ACL was implemented in all these gateway courses by 2023, there is no non-target group for that year. The S-ACL shows that, by 2023, the average S-ACL had risen to the same level as the 2021 target courses. Breaking out this trend by department, we see expected changes over time in departmental teams' engagement in the project:

- The CCT in math had already implemented ACL in target courses in 2021 and sustained that work in 2023 (small effect size of 0.10).
- Physics' implementation increased from 2021 to 2023 (modest effect size of 0.29).
- CS was not actively implementing ACL in 2021, but had begun to do so by 2023 (large effect size of 0.99).

In each case, mean S-ACL increases as the number of course sections implementing ACL increases.

5.5. Results from Classroom Observation

Table 8 presents a summary of classroom activities observed in the 22 gateway recitation sections that were visited for observation. Group and/or individual work was consistently and substantially included in both recitations and blended sections. At least some of this work occurred in every class visited, across multiple gateway courses in math, physics, and CS. For true recitations, both GTA-led and faculty-led sections used substantial class time for group and/or individual problem-solving (70%) vs. quite modest lecture time (14%). In blended lecture-recitation classes, time for lecture was more prevalent (47%) but still well balanced against student work, on average (46%).

Table 8. Observed Class Activities for True Recitation and Blended Class Meetings.

		True Recitation Sections (16)		Blended Lec/Rec Sections (6)	
Definition	**Activity**	**Number of Sections**	**Mean % of Class Time Spent (Std. Dev.)**	**Number of Sections**	**Mean % of Class Time Spent (Std. Dev.)**
instructor gives lecture (formal + interactive)	total lecture	6	15 (23)	6	50 (21)
instructor summary, ≤5 min	short explanations	7	8 (11)	1	1 (3)
discussion in which students respond to peers	whole class discussion	0	0	0	0
students work on problems alone or in small groups	total group + individual work	16	71 (23)	6	48 (19)
students present their work to the whole class	student presentations	0	0	0	0
announcements & other class business	total other	6	7 (11)	0	0

Time spent on quizzes was removed from totals, consistent with the way survey questions were asked.

The observations also showed a range of teaching approaches and student engagement among the 22 class sessions observed. Qualitative analysis of field notes showed patterns in instructors' choices in implementing and facilitating student work that also shaped the nature and quality of student engagement in ACL (Laursen, 2024). One observed variation was in the instructor's ability to draw out and respond to student thinking, which typically depends on multiple factors, including their teaching experience with active learning approaches, their content knowledge, and their pedagogical content knowledge (Auerbach et al., 2018). The observations also identified three sets of instructor choices for facilitating guided inquiry work that influenced student engagement significantly:

- Whether students wrote on whiteboards vs. on paper at table
- Whether and when students were required to hand something in, alone or as a group
- Whether instructors circulated assertively and how they interacted.

Most of the classrooms we observed were well-equipped spaces well equipped for interactive pedagogies, with mobile chairs and tables for small group work and multiple, large whiteboard spaces. In general, we observed stronger engagement when students worked at the whiteboards rather than on paper. For example, they were more likely to collaborate actively—e.g., adding ideas of their own or corrections to others' work on the board—and less likely to be silent, disengaged, or work alone. The strongest cases of engagement occurred when all students were asked to work in groups at the whiteboards, and were not given a choice of whether to work alone or work in groups at their tables (except in classes where whiteboard space was limiting, where students used stand-alone portable whiteboards at their tables).

The field notes also reflected stronger engagement in ACL when students had moderate accountability for the guided inquiry tasks assigned in the recitation. Instructors' approaches varied: for example, some instructors had students turn in evidence of their work as a group, such as a photograph of their whiteboard, and gave credit for participating rather than getting all the answers correct. This approach prioritizes group interdependence through social accountability of students to their peers. Other instructors had students turn in very similar problems individually, as homework, within the 2–3 days. In these cases, the recitation work itself was not graded but it was clear that doing and learning from the recitation work assigned in class helped with the graded, out-of-class homework assignment. When accountability was highest, such as a requirement to complete a fixed number of problems by the end of recitation time, time pressure was also high, and students tended to work alone. Many relied on TAs and LAs when they did seek help, and fairly few spoke to a peer. These students said they sat with trusted peers or study partners to whom they could direct questions but would not be slowed down by having to explain to someone else.

Finally, we observed higher engagement in classes where instructors and learning assistants circulated assertively and facilitated actively. Observed active facilitation behaviors included asking open-ended questions that required students explain their thinking, encouraging them to stay on task, and praising students' effort and dialogue rather than assessing correctness. We saw instructors demonstrate patience with the slow pace of group work and attend to equity by disengaging with more talkative groups or students who were trying to extract answers and avoid thinking critically. Instructors who regularly visited each group—not just responding to raised hands—had better success in keeping students accountable to making progress on the tasks and engaging with each other.

6. Discussion

In this section we discuss the key results and their implications.

6.1. Comparing Multiple Measurements of Teaching

We begin by discussing the results from comparing multiple methods to measure the teaching delivered in these recitations for gateway courses in mathematics, physics and CS. Individually, students' reports of class time do vary, reflecting their different abilities to estimate this as well as personal variations in how they spend class time. From pilot work, for example, we learned that students report any kind of instructor talk as "lecture," including non-content talk such as announcements. Students also tend to report more individual seat work than instructors say they assign, perhaps because students work out ideas for themselves during lectures or work alone for a bit before they join a group task. That is, there are real differences between the single distribution of class activities that instructors report, based on their planning for the class and the tasks they assign, and the variety of things that any student may be doing at the same time. However, the variability of student reports is counter-balanced by averaging: Using reports from sections with at least 10 students, as we did, yields a section average that is generally in solid agreement with what their instructors report, as seen in Figure 2.

Our findings are comparable to those from a recent study by Testa et al. (2025). These authors found that students were more accurate than instructors in estimating the percentage of class time used in active learning, as compared to observations by trained evaluators that are treated as a reference criterion, when students, instructors and observers all reported on the same, single class session. Testa and coauthors found statistically significant correlations near 0.4 between student reports and observations, but correlations were lower and non-significant for instructor reports vs. observations. These authors suggest that students were good at judging active learning because such methods were commonly used in their program, and students were exposed to many variants of instruction by taking multiple courses. Instructors, on the other hand, were less experienced with active learning overall, less familiar with variations of active learning other than their own usage, and may have overestimated their use of active learning approaches due to the general positive culture around active teaching in their organization.

Like their study, our student-instructor comparison suggests that instructors may report higher use of ACL approaches than what students report. In Figure 2, several instructors report using 80–100% ACL, while student means do not exceed 3 on the S-ACL scale, or "often," when we might consider 80–100% a closer match to 4, "almost always." That is, we'd expect a steeper slope for the relationship in Figure 2 if students and instructors agreed, on average.

However, comparison of the instructor reports with observation (Figure 3) does not suggest such a mismatch. Rather, it points out the importance of time sampling if researchers are using observations to characterize teaching. The two high outliers—cases where we observed nearly 100% ACL in a single class, but instructors reported much less use on average—and two low outliers, where instructors report near 60% active learning on average, but we saw somewhat less, may all reflect particular choices made by instructors to meet their students' learning needs on the single day we visited. Both kinds of variability illustrate the risks of mischaracterizing general practices from a single day's observation data, and highlight the fallacy of assuming that observations are "accurate" while self-report is biased. Weston et al. (2021) discuss this issue at greater length.

6.2. Comparing Student Reports of Teaching with Their Experiences of Learning

A second type of evidence comes from the relationship of student survey reports of ACL to their experiences of the classroom environment and their learning. Students' reports of the classroom learning environment (using WIHIC) were positively related to their reports of active learning teaching practices, as seen in Figure 1. That is, as the use of active learning teaching practices increases, students rate the classroom environment more favorably. The correlation here is quite striking, evident in the individual S-ACL data (Figure 1a) and especially in the section means (Figure 1b). This is noteworthy because prior research has found positive classroom environments to be associated with more positive student attitudes and higher learning achievement (see reviews by Fraser, 2012, 2023). As Fraser points out, "This research has revealed consistent, strong and positive associations between the quality of students' learning environments and their cognitive and affective outcomes" (Fraser, 2023, p. 4).

We also found a positive correlation between students' reports of active learning teaching practices and their perceived learning gains: as students report more experiences with active learning, they report greater gains in their learning. These findings are consistent with prior research that shows that active learning is positively associated with student achievement in mathematics and other STEM fields (Berssanette & de Francisco, 2021; Castillo et al., 2024; Crouch & Mazur, 2001; Deslauriers et al., 2019; Duran et al., 2024; Freeman et al., 2014; Hake, 1998; Kramer et al., 2023); and, in mathematics, with students' own sense of mastery of mathematical ideas and thinking skills (Laursen et al., 2014).

6.3. Sources of Variability

The results when S-ACL is averaged over larger aggregates (Figure 4, Tables 6 and 7) demonstrate the usefulness of the TAMI-SS and derived S-ACL measure to probe differences in instruction, such as change over time or differences in course settings. In particular, we did not observe differences in S-ACL scores by student characteristics, but we did observe differences in scores that were reflective of students' classroom contexts. For example, we found no differences in S-ACL scores by student gender, race, or ethnicity, but we did find differences between courses that had undergone the intended teaching reforms and courses that were not (yet) reformed, as Figure 4 shows. Thus, we argue, the observed differences in S-ACL scores provide evidence that the project's work had the desired overall effect of increasing instructors' ACL use, and these changes are not attributed to the student characteristics of a given course section.

Aggregated at the department level, the data do show differences in instructor practices and/or student response among mathematics, physics and CS courses. While it is tempting to interpret these as disciplinary in nature, we ascribe most of this variation to CS' later start on ACL implementation, which was still in progress at the second measurement point. Qualitative analyses of students' write-in comments suggest that mathematics and physics students were more open to ACL practices, perhaps due to greater exposure and acculturation in their departments (Reeves et al., 2023). At our second survey point, both faculty practice and student response in CS were still evolving, and we did not have a third survey opportunity. It is possible that a study of large, multi-site sample of courses with established ACL practices could identify sustained discipline-based patterns.

A different source of variation is the teaching that occurred. The departmental S-ACL means in 2023 show that the desired instructional changes were becoming widespread—yet there was still notable variation among sections of the same course, showing that the desired reforms had not been uniformly adopted. This variation is readily seen in the observation data, where we observed a range of teaching approaches and student engagement among the 22 class sessions. Because many of the sections met in classrooms with good whiteboard space, we could corroborate the findings of Liljedahl (2020) about the "massive" merits of vertical non-permanent surfaces for group work: Easy editing and erasing promotes risk-taking, and working at a vertical surface prevents students from disengaging. Likewise, our findings on behaviors of instructors align with those in the "thinking classrooms" that Liljedahl (2020) studied, especially the instructors who circulated steadily and proactively to listen to students and ask them what he calls "keep-thinking questions," designed to promote students' thinking rather than assess the rightness of their answer. Finally, our observations about accountability structures mirror comments by Clark and Talbert (2023) on how student motivation is strengthened by providing feedback, especially as students are developing or practicing skills. Being accountable for doing the work, not for getting the right answer, shifts the focus of in-class work to learning rather than grading. Liljedahl (2020) likewise emphasizes the benefits of rebranding homework as "check-your-understanding" questions to emphasize purpose and thought over scoring or ranking.

In previous studies of inquiry-based mathematics at the tertiary level (e.g., Laursen et al., 2014) we have often seen students presenting their own work to peers, which then often prompted whole-class discussion. This was more common in upper-level mathematics courses where students have passed the beginner level in both expertise and confidence in mathematics. At GMU, student presentations were not observed in this set of gateway course recitations. This reflects the CCTs' choices to use a guided inquiry approach with emphasis on collaborative learning, rather than one more centered on individual students. That is, the observed uses of time are generally consistent with the project's strategy to devote significant recitation time to student work with peers on inquiry tasks in a collaborative, instructor-supported environment. This approach, focused on meaningful practice with concepts already introduced elsewhere, may be well suited to recitations, and seems to be easier for novice instructors to execute, in comparison with approaches that introduce new material through inquiry.

All three of the facilitation structures that stood out in observations are straightforward decisions where specific, proactive choices by instructors were seen to be more effective in engaging students: use of vertical writing surfaces, modest task accountability that emphasizes learning and social accountability to peers rather than grading, and proactive facilitation. All three approaches could be selected as course-wide tactics and applied course-wide by instructors of all experience levels. They place priority on using recitation time for students to think and explain their thinking in ways that generate and test knowledge, that are supported by instructor guidance and support, and that offer both cognitive and sociocultural benefits (Hodges, 2018). These are relative lightweight structures, applied here to recitation alone, but the results corroborate other scholars' findings about the benefits of high structure in active learning courses. Instructor facilitation and assessment choices that increase interaction and support peer accountability can reinforce time on task to learn the material and strengthen students' sense of being part of a community of learners, especially for students with less prior preparation (Freeman et al., 2011; Eddy & Hogan, 2014; see also McCrea et al., 2016, for a review of studies on facilitation of group work). Thus, these findings point to some additional steps the project team could take to improve student experiences more consistently across sections.

6.4. Affordances and Limitations of the TAMI-SS for Research and Evaluation

As a means to characterize teaching for evaluation and research, the TAMI-SS student survey and the derived S-ACL measure have multiple affordances. Data about teaching practices are based on multiple observers—the students—attending class and reporting on their experiences across a full academic term, avoiding the sampling errors arising from single or small-N observations. Setting a minimum threshold (e.g., 10 responses per section) helps to ensure that class or section means are representative. An attendance measure can be used, as we did, to check that the reports are based on actual student experience. The measure is simple and focused on teachers' use of class time, unlike end-of-course instruments that ask students to evaluate the quality of teaching and thus readily invoke implicit or explicit biases. The TAMI-SS is quick for students to complete and we found very few examples in which it appeared students had raced through, giving the same response to every item. The survey can be used with other survey items (such as the WIHIC or SALG in this study), or in combination with exams, grades, and other outcome measures, if linked to identifiers. Even without student outcome measures such as student learning measures, student-provided evidence *about teaching* can be powerful in gathering evidence about the implementation of an instructional intervention.

As is the case for any student-focused data collection, both net numbers and response rates are critical. Even when overall response rates approach 50%, variable response levels by course or section can limit the use of this measure in making comparisons among courses or over time. Such comparisons are often essential for making arguments about the implementation progress of a desired teaching change. While non-response bias from a given section is mitigated by including only sections with ten or more student responses in analyses, this may mean that some sections are not represented at all in course-level analyses, and systematic error may arise from these omissions. Use of multiple data sources can help to confirm that the included sections represent the range of practices across all course sections, but to eliminate any potential bias, future studies should seek to maximize response rates across all course sections. In particular, when comparing student responses to instructor self-report, as in Figure 3, total response rates are important, but even more so is the extent of matching between student section and instructor response, that is, the section-by-section response rate. While we chose (for logistical reasons) to contact students through their instructors rather than to acquire the identifiable information needed to contact them directly, it is likely that instructor participation was one bottleneck in reaching students. Some strategies to increase response rate include making the survey available through the course learning management system and encouraging instructors to invite their students to complete the survey (NSSE, n.d.). In our study, providing a QR code and a few minutes of dedicated class time to complete the study significantly boosted student response. Proactive, personal communication by the project's liaison to the instructors in their department was impactful in both encouraging participation in the survey and arranging a classroom access to boost student response. Paid incentives to instructors (e.g., gift cards) were only modestly productive, and were more effective with GTAs than with faculty.

Efforts to reduce variation of response rate by course and section can also help to reduce bias in who responds. Instructors can be a positive influence in encouraging students to respond, but because of the power differential between students and instructors, it is important to take strong steps to protect student anonymity and confidentiality, especially if identifiers are also gathered for linking survey data to student outcomes such as exam scores, course grades, or persistence in their educational program. Likewise, it is important to engage with instructors and ensure that their participation is not threatened by concerns about how findings will be shared or used. In particular, we recommend against using this tool to evaluate or study individual instructors. Rather, it is useful to indicate progress on efforts to support instructors to shift their practices toward research-based instructional strategies, at an institutional scale (Laursen et al., 2019).

7. Conclusions

We demonstrate the utility of a new student survey instrument, TAMI-SS, and scale, S-ACL, for characterizing teaching. In this study, the S-ACL was used to measure the *presence* of instruction aimed at active and collaborative learning in college mathematics, physics and computer science. We found that S-ACL data also correlated positively and strongly with two established survey measures that reflect the *quality* of instruction, from students' points of view. The S-ACL score correlated positively with students' ratings of the learning environment using the WIHIC scale, and with students' self-reported learning gains, based on several SALG items. Classes with higher S-ACL means—those where, on average, students reported that more active and collaborative learning occurred—were also reported to be more engaging, challenging, and interactive, and as generating more learning. Students' individual ratings on the TAMI-SS vary, but when they are combined into a composite measure and averaged across meaningful groups, the ratings appear to provide a trustworthy picture of the instruction students are collectively experiencing.

These qualities, and the simplicity of the measure, give it value for studies of tertiary mathematics instruction and efforts to change it within large courses, departments, or subdisciplines. As a measure of the teaching that students experience, the S-ACL can be used by researchers and evaluators to discern patterns of difference in pedagogical approaches, to assess variability across a multi-section course, and to measure change after professional development or a course reform effort. In this role, as a metric for assessing progress on an institutional effort to increase use of active and collaborative learning, the S-ACL composite from TAMI-SS items was responsive to change over time and sensitive to differences in implementation. These features make it a valuable tool for studies that seek to enhance understanding of how best to support instructors in moving toward inclusive, evidence-based teaching that better supports mathematics and other STEM students in meeting their educational goals. In highlighting the importance of large and representative samples, and some challenges to achieving these, we offer experiential advice to others who may use the instrument for these purposes and reinforce our warning against the use of the TAMI-SS to evaluate or compare teaching practices of individual instructors.

Author Contributions: Conceptualization, S.L.L. and T.A.; Methodology, S.L.L. and T.A.; Validation, T.A.; Formal analysis, S.L.L. and T.A.; Investigation, S.L.L. and T.A.; Resources, S.L.L.; Data curation, S.L.L. and T.A.; Writing—original draft, S.L.L. and T.A.; Writing—review and editing, S.L.L. and T.A.; Visualization, S.L.L. and T.A.; Supervision, S.L.L.; Project administration, S.L.L.; Funding acquisition, S.L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by US National Science Foundation grant numbers DUE-1821440 and DUE-1821704. Pilot work for the research was supported by the Spencer Foundation under award 201500163. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and reviewed by the Institutional Review Board of the University of Colorado Boulder (protocol code 18-0581, designated Exempt 14 November 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: A restricted version of the data presented in this study is available to qualified researchers upon request from the corresponding author, due to the presence of identifiable information.

Acknowledgments: At GMU, we thank Jessica Rosenberg, Bob Sachs, Mark Snyder, Melinda Ryan, Nischchal Thapa Magar and Marco Brizzolara for their advice and help in identifying samples and gathering data. We thank them as well as Jill Nelson, Jaime Lester and Joanna Jauchen for valuable conversations. We appreciate the interest and candor of the GMU students, instructors, graduate teaching assistants, and undergraduate learning assistants who took part in the project and in the study. At CU Boulder, we thank Tim Weston and Chuck Hayward for helpful discussions.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ACL	Active and collaborative learning
CCT	Course-based communities of transformation
CS	Computer science
GMU	George Mason University
GTA	Graduate teaching assistant
I-ACL	Instructor-reported ACL score (survey measure)
LA	Learning assistant
PIPS	Postsecondary Instructional Practices Survey
RQ	Research question
S-ACL	Student-reported ACL score (survey measure)
SALG	Student Assessment of their Learning Gains (survey measure)
STEM	Science, technology, engineering and mathematics
TA	Teaching assistant
TAMI-IS	Toolkit for Assessment of Mathematics Instruction—Instructor Survey
TAMI-SS	Toolkit for Assessment of Mathematics Instruction—Student Survey
UTA	Undergraduate teaching assistant
WIHIC	What is happening in this class (survey measure)

References

AAAS (American Association for the Advancement of Science). (2013). *Describing and measuring undergraduate STEM teaching practices*. Available online: https://drive.google.com/file/d/1Du7sCx3g3p21C9LwgnyJvmA5E9EosEU2/view?usp=sharing (accessed on 10 December 2025).

Achen, R. M., & Lumpkin, A. (2015). Evaluating classroom time through systematic analysis and student feedback. *International Journal for the Scholarship of Teaching and Learning*, *9*(2), n2. [CrossRef]

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active learning not associated with student learning in a random sample of college biology courses. *CBE—Life Sciences Education*, *10*(4), 394–405. [CrossRef] [PubMed]

Anwar, S., & Menekse, M. (2021). A systematic review of observation protocols used in postsecondary STEM classrooms. *Review of Education*, *9*(1), 81–120. [CrossRef]

Apkarian, N., Voigt, M., Hagman, J. E., Tremaine, R., Street, C., Martinez, A. E., & Guglielmo, J. (2024). Critical, intersectional, quantitative analyses of instructional practices and changes in undergraduate students' math affect. *International Journal of Research in Undergraduate Mathematics Education*, 1–32. [CrossRef]

Archie, T., Hayward, C. N., Yoshinobu, S., & Laursen, S. L. (2022). Investigating the linkage between professional development and mathematics instructors' use of teaching practices using the theory of planned behavior. *PLoS ONE*, *17*(4), e0267097. [CrossRef]

Auerbach, A. J., Higgins, M., Brickman, P., & Andrews, T. C. (2018). Teacher knowledge for active-learning instruction: Expert-novice comparison reveals differences. *CBE—Life Sciences Education*, *17*(1), ar12. [CrossRef]

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In *Higher education: Handbook of theory and research* (pp. 279–326). Springer.

Berssanette, J. H., & de Francisco, A. C. (2021). Active learning in the context of the teaching/learning of computer programming: A systematic review. *Journal of Information Technology Education: Research*, *20*, 201–220. [CrossRef]

Borda, E., Schumacher, E., Hanley, D., Geary, E., Warren, S., Ipsen, C., & Stredicke, L. (2020). Initial implementation of active learning strategies in large, lecture STEM courses: Lessons learned from a multi-institutional, interdisciplinary STEM faculty development program. *International Journal of STEM Education*, *7*(1), 1–18. [CrossRef]

Bressoud, D. M., Mesa, V., & Rasmussen, C. L. (Eds.). (2015). *Insights and recommendations from the MAA national study of college calculus*. MAA Press.

Bruff, D. (2015, September 15). In defense of continuous exposition by the teacher. *Agile Learning Blog*. Available online: https://derekbruff.org/?p=3126 (accessed on 10 December 2025).

Bryan, T., Bulancea, G., Crossin, K., Jauchen, J. G., Nelson, M., Sachs, R., & Sausville, C. (2025). Healthy vulnerable community and institutional change. *PRIMUS*, *35*(4–5), 444–459. [CrossRef]

Bulancea, G., Granfield, P., Jauchen, J., Love, J., Nelson, M., Sachs, R., & Sausville, C. (2021). A community of grassroots leaders: Leveraging faculty networks to create change. *PRIMUS*, *31*(3–5), 627–642. [CrossRef]

Campbell, C. M. (2017). An inside view: The utility of quantitative observation in understanding college educational experiences. *Journal of College Student Development*, *58*(2), 290–299. [CrossRef]

Castillo, A. J., Durán, P., Fuller, E., Watson, C., Potvin, G., & Kramer, L. H. (2024). Student attitudes in an innovative active learning approach in calculus. *International Journal of Mathematical Education in Science and Technology*, *55*(3), 730–758. [CrossRef]

Chiu, P. H. P., Im, S. W. T., & Shek, C. H. (2022). Disciplinary variations in student perceptions of active learning classrooms. *International Journal of Educational Research Open*, *3*, 100131. [CrossRef]

Clark, D., & Talbert, R. (2023). *Grading for growth: A guide to alternative grading practices that promote authentic learning and student engagement in higher education*. Routledge.

Corbo, J. C., Reinholz, D. L., Dancy, M. H., & Finkelstein, N. (2015). Departmental Action Teams: Empowering faculty to make sustainable change. In *2015 physics education research conference proceedings* (pp. 91–94). American Association of Physics Teachers.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. [CrossRef]

Cronin, A., Moore-Russo, D., Bjorkman, K., & Mills, M. (2025). A synthesis of mathematical departmental change efforts: Framing, foci, and future opportunities. *PRIMUS*, *35*, 571–591. [CrossRef]

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970–977. [CrossRef]

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, *116*(39), 19251–19257. [CrossRef]

Dorman, J. P. (2003). Cross-national validation of the What Is Happening In this Class? (WIHIC) questionnaire using confirmatory factor analysis. *Learning Environments Research*, *6*(3), 231–245. [CrossRef]

Duran, P. A., Castillo, A. J., Watson, C., Fuller, E., Potvin, G., & Kramer, L. H. (2024). Student attitudes and achievement in active learning calculus. *International Journal of Mathematical Education in Science and Technology*, *55*(3), 759–780. [CrossRef]

Durham, M. F., Knight, J. K., Bremers, E. K., DeFreece, J. D., Paine, A. R., & Couch, B. A. (2018). Student, instructor, and observer agreement regarding frequencies of scientific teaching practices using the Measurement Instrument for Scientific Teaching-Observable (MISTO). *International Journal of STEM Education*, *5*, 31. [CrossRef]

Durham, M. F., Knight, J. K., & Couch, B. A. (2017). Measurement Instrument for Scientific Teaching (MIST): A tool to measure the frequencies of research-based teaching practices in undergraduate science courses. *CBE—Life Sciences Education*, *16*(4), ar67. [CrossRef] [PubMed]

Eagan, K. (2016). *Becoming more student-centered? An examination of faculty teaching practices across STEM and non-STEM disciplines between 2004 and 2014*. Higher Education Research Institute. Available online: https://sloan.org/storage/app/media/files/STEM_Higher_Ed/STEM_Faculty_Teaching_Practices.pdf (accessed on 10 December 2025).

Eddy, S. L., Converse, M., & Wenderoth, M. P. (2015). PORTAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE—Life Sciences Education*, *14*(2), ar23. [CrossRef]

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE—Life Sciences Education*, *13*, 453–468. [CrossRef]

Elliott, E. R., Reason, R. D., Coffman, C. R., Gangloff, E. J., Raker, J. R., Powell-Coffman, J. A., & Ogilvie, C. A. (2016). Improved student learning through a faculty learning community: How faculty collaboration transformed a large-enrollment course from lecture to student centered. *CBE—Life Sciences Education*, *15*(2), ar22. [CrossRef]

Fairweather, J. (2008). *Linking evidence and promising practices in science, technology, engineering, and mathematics (STEM) undergraduate education*. Board of Science Education, National Research Council, The National Academies. Available online: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072637.pdf (accessed on 10 December 2025).

Fraser, B. J. (2012). Classroom learning environments: Retrospect, context and prospect. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 1191–1239). Springer.

Fraser, B. J. (2023). The evolution of the field of learning environments research. *Education Sciences*, *13*, 257. [CrossRef]

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. [CrossRef]

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education*, *10*(2), 175–186. [CrossRef]

Greenhoot, A. F., Aslan, C., Chasteen, S., Code, W., & Sherman, S. B. (2020). Variations on embedded expert models: Implementing change initiatives that support departments from within. In K. White, A. Beach, N. Finkelstein, C. Henderson, S. Simkins, L. Slakey, M. Stains, G. Weaver, & L. Whitehead (Eds.), *Transforming institutions: Accelerating systemic change in higher education* (Chapter 22). Pressbooks. Available online: http://openbooks.library.umass.edu/ascnti2020/chapter/greenhoot-etal/ (accessed on 10 December 2025).

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. [CrossRef]

Hayward, C., Weston, T., & Laursen, S. (2018). First results from a validation study of TAMI: Toolkit for assessing mathematics instruction. In A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, & S. Brown (Eds.), *Proceedings of the 21st annual conference on the research in undergraduate mathematics education, San Diego, CA, USA* (pp. 727–735). Mathematical Association of America, SIGMAA on RUME. Available online: http://sigmaa.maa.org/rume/RUME21.pdf (accessed on 10 December 2025).

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education*, *47*(1), 144–154.

Hodges, L. C. (2018). Contemporary issues in group learning in undergraduate science classrooms: A perspective from student engagement. *CBE—Life Sciences Education*, *17*(2), es3. [CrossRef]

Hoover, M. A., & Pelaez, N. J. (2008). Blood circulation laboratory investigations with video are less investigative than instructional blood circulation laboratories with live organisms. *Advances in Physiology Education*, *32*(1), 55–60. [CrossRef]

Hora, M. T. (2015). Toward a descriptive science of teaching: How the TDOP illuminates the multidimensional nature of active learning in postsecondary classrooms. *Science Education*, *99*(5), 783–818. [CrossRef]

Hora, M. T., & Ferrare, J. J. (2013). *A review of classroom observation techniques in postsecondary settings* (WCER Working Paper 2013-1). University of Wisconsin–Madison, Wisconsin Center for Education Research. Available online: https://wcer.wisc.edu/publications/abstract/wcer-working-paper-no.-2013-01 (accessed on 10 December 2025).

Iverson, E. (2012). *2012 AAAS measurement workshop—Faculty surveys white paper*. Commissioned white paper for the AAAS Measurement of Teaching Practices meeting. Available online: http://serc.carleton.edu/files/serc/commissioned_paper_faculty_sur.pdf (accessed on 10 December 2025).

Johnson, D. W., Johnson, R. T., & Smith, K. A. (2014). Cooperative learning: Improving university instruction by basing practice on validated theory. *Journal on Excellence in College Teaching*, *25*(3&4), 85–118.

Jones, F. (2017). Impact assessment of a department-wide science education initiative using students' perceptions of teaching and learning experiences. *Assessment & Evaluation in Higher Education*, *42*(5), 772–787.

Jones, F. (2018). Comparing student, instructor, classroom and institutional data to evaluate a seven-year department-wide science education initiative. *Assessment & Evaluation in Higher Education*, *43*(2), 323–338.

Koch, A. K. (2017). It's about the gateway courses: Defining and contextualizing the issue. In *Improving teaching, learning, equity, and success in gateway courses: New directions for higher education* (Vol. 180, pp. 11–17). Wiley.

Kramer, L., Fuller, E., Watson, C., Castillo, A., Oliva, P. D., & Potvin, G. (2023). Establishing a new standard of care for calculus using trials with randomized student allocation. *Science*, *381*(6661), 995–998. [CrossRef]

Kreitzer, R. J., & Sweet-Cushman, J. (2021). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*, *20*(1), 73–84. [CrossRef]

Landrum, R. E., Viskupic, K., Shadle, S. E., & Bullock, D. (2017). Assessing the STEM landscape: The current instructional climate survey and the evidence-based instructional practices adoption scale. *International Journal of STEM Education*, *4*, 25. [CrossRef]

Larsen, S. L., Glover, E., & Melhuish, K. (2015). Beyond good teaching. In D. M. Bressoud, V. Mesa, & C. L. Rasmussen (Eds.), *Insights and recommendations from the MAA national study of college calculus* (pp. 93–106). MAA Press.

Laursen, S. L. (2024). *On the ground with active and collaborative learning in STEM gateway courses: Findings from classroom observations, year 6* (Report to George Mason University Gateway 2 STEM Project). University of Colorado Boulder.

Laursen, S. L., Andrews, T., Stains, M., Finelli, C. J., Borrego, M., McConnell, D., Johnson, E., Foote, K., Ruedi, B., & Malcom, S. (2019). *Levers for change: An assessment of progress on changing STEM instruction*. American Association for the Advancement of Science. Available online: https://www.aaas.org/resources/levers-change-assessment-progress-changing-stem-instruction (accessed on 10 December 2025).

Laursen, S. L., Archie, T., Weston, T. J., Hayward, C. N., & Yoshinobu, S. (2023). A measurement hat trick: Evidence from a survey and two observation protocols about instructional change after intensive professional development. In S. Cook, B. Katz, & D. Moore-Russo (Eds.), *Proceedings of the 25th annual conference on research in undergraduate mathematics education, Omaha, NE, USA*. Available online: http://sigmaa.maa.org/rume/RUME25_Proceedings.pdf (accessed on 10 December 2025).

Laursen, S. L., Hassi, M. L., Kogan, M., Hunter, A. B., & Weston, T. (2011). *Evaluation of the IBL mathematics project: Student and instructor outcomes of inquiry-based learning in college mathematics* (Report to the educational advancement foundation and the IBL mathematics centers). University of Colorado Boulder, Ethnography & Evaluation Research, Assessment & Evaluation Center for Inquiry-Based Learning in Mathematics. Available online: https://www.colorado.edu/eer/sites/default/files/attached-files/iblmathreportall_050211.pdf (accessed on 10 December 2025).

Laursen, S. L., Hassi, M. L., Kogan, M., & Weston, T. J. (2014). Benefits for women and men of inquiry-based learning in college mathematics: A multi-institution study. *Journal for Research in Mathematics Education*, *45*(4), 406–418. [CrossRef]

Liljedahl, P. (2020). *Building thinking classrooms in mathematics, grades K-12: 14 teaching practices for enhancing learning*. Corwin Press.

Macdonald, R. H., Beane, R. J., Baer, E. M., Eddy, P. L., Emerson, N. R., Hodder, J., Iverson, E. R., McDaris, J. R., O'Connell, K., & Ormand, C. J. (2019). Accelerating change: The power of faculty change agents to promote diversity and inclusive teaching practices. *Journal of Geoscience Education*, *67*(4), 330–339. [CrossRef]

Manduca, C. A., Iverson, E. R., Luxenberg, M., Macdonald, R. H., McConnell, D. A., Mogk, D. W., & Tewksbury, B. J. (2017). Improving undergraduate STEM education: The efficacy of discipline-based professional development. *Science Advances*, *3*(2), e1600193. [CrossRef] [PubMed]

McCrea, R., Neville, I., Rickard, D., Walsh, C., & Williams, D. (2016). *Facilitating group work: A guide to good practice*. Technological University Dublin. Available online: https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1007&context=ltcpgdprp (accessed on 10 December 2025).

Measures of Effective Teaching [MET] Project. (2012). *Asking students about teaching: Student perception surveys and their implementation* (Policy and Practice Summary). Bill & Melinda Gates Foundation. Available online: https://usprogram.gatesfoundation.org/news-and-insights/usp-resource-center/resources/asking-students-about-teaching-student-perception-surveys-and-their-implementation (accessed on 10 December 2025).

Miller, A., & Dumford, A. (2017). *Social desirability bias*. FSSE Psychometric Portfolio. Available online: https://scholarworks.iu.edu/dspace/bitstreams/651dad9e-35e6-4820-af21-254faf32f7d7/download (accessed on 10 December 2025).

NSSE (National Survey of Student Engagement). (n.d.). *Maximizing your number of respondents effectively and ethically*. Available online: https://nsse.indiana.edu/nsse/administering-nsse/promoting-nsse-on-campus/maximizing-your-number-of-respondents-effectively-and-ethically.html (accessed on 10 December 2025).

Otero, V. K. (2015). Nationally scaled model for leveraging course transformation with physics teacher preparation. In C. Sandifer, & E. Brewe (Eds.), *Recruiting and educating future physics teachers: Case studies and effective practices* (pp. 107–116). American Physical Society. Available online: https://cdn.sanity.io/files/4wurd6lm/production/290c6570a2d6110db5943c0cd4b26e81558c42c0.pdf (accessed on 10 December 2025).

Peteroy-Kelly, M., Brancaccio-Taras, L., Awong-Taylor, J., Balser, T., Jack, T., Lindsay, S., Marley, K., Romano, S., Uzman, J. A., & Pape-Lindstrom, P. (2019). A qualitative analysis to identify the elements that support department level change in the life sciences: The PULSE Vision & Change Recognition Program. *PLoS ONE*, *14*(5), e0217088.

Reeves, P. M., Cavanagh, A. J., Bauer, M., Wang, C., & Graham, M. J. (2023). Cumulative cross course exposure to evidence-based teaching is related to increases in STEM student buy-in and intent to persist. *College Teaching*, *71*(1), 66–74. [CrossRef]

Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R., & Shepard, L. A. (2011). Impact of undergraduate science course innovations on learning. *Science*, *331*(6022), 1269–1270. [CrossRef]

Seymour, E., Wiese, D., Hunter, A., & Daffinrud, S. M. (2000, March 27). *Creating a better mousetrap: On-line student assessment of their learning gains*. National Meeting of the American Chemical Society, San Francisco, CA, USA. Available online: https://www.colorado.edu/eer/sites/default/files/attached-files/seymoursalgbettermousetrap2000.pdf (accessed on 10 December 2025).

Skordi, P., & Fraser, B. J. (2019). Validity and use of the What Is Happening In this Class?(WIHIC) questionnaire in university business statistics classrooms. *Learning Environments Research*, *22*(2), 275–295. [CrossRef]

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE—Life Sciences Education*, *12*(4), 618–627. [CrossRef]

Smith, W. M., & Rasmussen, C. (2025). Infusing active learning into precalculus and calculus: Mathematics department transformation in challenging times. *PRIMUS*, *35*(4–5), 329–337. [CrossRef]

Smith, W. M., Rasmussen, C., & Tubbs, R. (2021). Introduction to the special issue: Insights and lessons learned from mathematics departments in the process of change. *PRIMUS*, *31*(3–5), 239–251. [CrossRef]

Street, C., Apkarian, N., Gehrtz, J., Tremaine, R., Barron, V., Voigt, M., & Hagman, J. E. (2021). X-PIPS-M Data Summary. *arXiv*, arXiv:2111.01795.

Testa, G. M. G., Souza, M. B. D. O., Paes, Â. T., & Magdalon, J. (2025). Students outperform faculty in recognizing the use of active learning methods. *Advances in Physiology Education*, *49*(2), 356–365. [CrossRef]

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., . . . Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, *117*(12), 6476–6483. [CrossRef]

Walter, E. M., Henderson, C. R., Beach, A. L., & Williams, C. T. (2016). Introducing the Postsecondary Instructional Practices Survey (PIPS): A concise, interdisciplinary, and easy-to-score survey. *CBE—Life Sciences Education*, *15*(4), ar53. [CrossRef] [PubMed]

Weston, T. J., Hayward, C. N., & Laursen, S. L. (2021). When seeing is believing: Generalizability and decision studies for observational data in evaluation and research on teaching. *American Journal of Evaluation*, *42*(3), 377–398. [CrossRef]

Weston, T. J., Laursen, S. L., & Hayward, C. N. (2023). Measures of success: Characterizing teaching and teaching change with segmented and holistic observation data. *International Journal of STEM Education*, *10*(1), 24. [CrossRef]

Wieman, C. (2017). *Improving how universities teach science: Lessons from the science education initiative*. Harvard University Press.

Wiggins, B. L., Eddy, S. L., Wener-Fligner, L., Freisem, K., Grunspan, D. Z., Theobald, E. J., Timbrook, J., & Crowe, A. J. (2017). ASPECT: A survey to assess student perspective of engagement in an active-learning classroom. *CBE—Life Sciences Education*, *16*(2), ar32. [CrossRef]

Williams, C. T., Walter, E. M., Henderson, C., & Beach, A. L. (2015). Describing undergraduate STEM teaching practices: A comparison of instructor self-report instruments. *International Journal of STEM Education*, *2*, 18. [CrossRef]