Not-so-Secret Watchers: What Students Can Tell Us About Teaching

Sandra Laursen
University of Colorado Boulder

Tim Archie
University of Colorado Boulder

Projects to improve teaching as a means to improve learning need good tools to measure shifts in the use of effective research-based instructional strategies. We describe a survey tool to assess active and collaborative learning practices in college STEM by asking students to report what practices are used commonly in their course. Student reports correlate with their experience of the learning environment, and with their self-reported learning outcomes. Student means by section are corroborated by instructor-reported use of these practices and external observations.

Keywords: teaching, professional development, survey measures, research-based instructional strategies (RBIS)

A basic question in education research is the linkage between teaching and learning. This is surprisingly difficult to measure, however. To study how teaching activities relate to learning outcomes, we must have ways to describe teaching in diverse settings (AAAS, 2013). This need surfaces, for example, in studies of interventions to improve teaching (e.g., Manduca, et al., 2017; Archie, et al., 2022). Efforts to increase college instructors' uptake of research-based instructional strategies (RBIS) need tools to assess the initial state of participants' teaching, to measure changes over time, and to know if their change strategies are working or not.

Surveys and observations are commonly used methods for studying teaching. Instructor surveys can be efficiently administered to large numbers of instructors, but they pose challenges including sampling, internal validity, and potential biases in instructor self-report (Iverson, 2012). Social desirability bias is particularly concerning in self-reports of teaching. Williams and coauthors (2015) review properties of existing instruments for instructor self-report.

Classroom observations offer detailed, naturalistic information about teaching, but they are time-consuming and costly. Hora and Ferrare (2013) and Campbell (2017) discuss strengths and limitations of classroom observation in higher education, with particular attention to validity and reliability concerns. Weston, Hayward and Laursen (2021) discuss additional challenges to reliability when time sampling is applied to make inferences about term-long courses.

As regular occupants of classrooms, students can offer insight about the teaching taking place, especially when combined with other measurement methods. In evaluative uses, student ratings of teaching are often derided as popularity contests with known, serious biases (Kreitzer & Sweet-Cushman, 2021; Heffernan, 2022). Others find 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" (Benton & Cashin, 2014, p. 315). The MET study (2012) found clear connections between student ratings of their teachers and student achievement gains: "Students know an effective classroom when they experience one." (p. 1) The key is asking "the right questions, in the right ways" (p. 1), probing aspects of teaching that students are qualified to evaluate, and doing so in ways that minimize threats to validity.

Mindful of these caveats, we developed a brief student survey to characterize the teaching in recitations of large STEM gateway courses, in the context of an institutional project to infuse active and collaborative learning into those recitations. Importantly, we did not survey students to measure course outcomes, such as attitude or learning. Rather, we capitalized on students' everyday presence in the classroom to report on the teaching they experienced, and we used their reports to describe and measure change in teaching over time. The simple student survey focuses on the frequency of specific teaching practices and is aligned with other instruments that we have developed (Hayward, et al., 2018; Weston, et al., 2021, 2023). Alongside this measure, we also measured students' perceptions of the classroom environment and a small set of student self-reported learning gains, to assess whether these experiences were related to the teaching practices that students described. We describe results for three research questions (RQ):

1. How does student-reported use of ACL practices relate to (a) their 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 can these student metrics inform formative and summative evaluation?

Conceptual Framework

We follow Neumann (2014) in viewing the purpose of higher education as learning "how to live a full and democratic life," exerting essential freedoms with the abilities "to argue and write, to inquire and imagine, and to act ethically and humanely" (pp. 249-250). This view contrasts with notions of the purpose of higher education as developing a labor force to compete in the global economy. Substantive learning requires that students encounter new subject-matter ideas and connect them to their existing personal knowledge. The teacher's job, Neumann argues, is to orchestrate and bring about these encounters. Doing this well requires sophistication, and thus we recognize teaching as a complex human activity that invokes individual beliefs, values, knowledge, and behaviors, and that takes place within nested cultural and operational contexts that include the classroom, the institution, and the wider community (Schoenfeld, 1998).

Our perspective on measurement draws on Guskey's framework for evaluating teacher professional development (PD; 2000, 2002). The evaluation framework begins with two near-term outcomes, teachers' experiences of the PD and the learning they take away, and it ends with student learning, the desired ultimate outcome. Connecting near-term outcomes to downstream learning are teacher context and implementation: how supportive is teachers' context to implement what they learn in PD, and how does their classroom behavior change in ways that will lead to stronger student learning. Both these mid-stage outcomes are more complex to measure, but influence whether and how the student learning outcomes are accomplished.

Here we focus on development of a tool to measure elements of implementation of active and collaborative learning, a suite of RBIS linked to student learning (Freeman et al., 2014). While our tool yields a picture of teaching much simpler than Neumann's portrayal of complex classroom interplay, both have at heart teachers' choices about how to use class time to more or less actively orchestrate students' encounters with subject matter (Campbell et al., 2017). Effective use of active learning techniques focuses students' attention on course material and creates opportunities to connect to their own knowledge and learning needs, as students engage with each other and the instructor over the course material. We do not assume that teachers are equally skillful at implementing these approaches, nor indeed at lecturing, but we do assume that their choices are agentic and that they matter for student experience and outcomes.

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 recitation sections of gateway courses in mathematics, physics and computer science

(CS). Gateway courses are foundational to a variety of educational pathways, high in enrollment, and high-risk for students (Koch, 2017). Here the gateway courses were two two-term calculus sequences, two two-term sequences in introductory physics, and introduction to programming in CS. In 2023 a new two-term intro sequence in CS was added.

Recitations were targeted as a foot in the door, seeking to give instructors concrete experiences with ACL approaches, develop their expertise and interest, and build local teaching cultures that would spur broader changes. In large courses, ACL was implemented by graduate assistants who teach recitations in math and CS; term and tenure-track faculty, who teach physics presentations and blended recitation/lecture courses in math and CS. Each department chose a guided inquiry approach, in which students worked in small groups on sequenced tasks, called worksheets (mathematics), tutorials (physics), or programming assignments (CS). Common tasks were prepared and recitation instructors were encouraged but not required to use them.

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 are in the first generation of their families to attend college. The students are ethnically diverse, like the metro area where GMU is located.

Research Methods

We surveyed students and instructors about their experiences of recitation across the whole term, and observed single class sessions for a sample of recitation sections and blended courses.

Surveys and Observations

The short survey (≤10 min) focused on student experiences. We used three blocks of items (Table 1)—two existing instruments plus our instrument, called the TAMI-SS (Toolkit for Assessing Mathematics Instruction: Student Survey) in keeping with our other TAMI instruments. The instructor survey (TAMI-IS) asked teachers to estimate their use of the same five teaching practices, by percentage of total class time, across the whole course.

Table 1: Overview of Survey Item Blocks Used in this Study

Use of ACL practices: TAMI-SS	Learning environment: WIHIC	Learning gains: SALG
(Hayward, Weston & Laursen, 2018; Weston, Laursen & Hayward, 2023) frequency of 5 teaching practices: lecture, small group work, individual work, student presentation, whole class discussion (0=never to 4=almost always) S-ACL = mean for 3 practices: small group work, student presentation, whole class discussion (0-4)	(Dorman, 2003; Hoover & Pelaez, 2008; Skordi & Fraser, 2019) frequency of experiences in 7 areas (8 items each): student cohesiveness, teacher support, involvement, inquiry/ investigation, task orientation, cooperation, equity (0=almost never to 4=almost always) WIHIC: mean (0-4)	(Seymour et al., 2000; Laursen et al., 2011) learning gains in 5 areas common to all 3 disciplines (0=no gain to 4=great gain) Gains: mean (0-4)

Surveys were administered online near the end of fall term, to 112 sections in 9 courses in 2021 and to 94 sections of 11 courses in 2023, to over 3000 students each time. Student surveys were distributed by recitation instructors to protect student anonymity. We gave instructors a script to explain the survey and asked them to provide class time to respond.

In all, about 2200 students responded at least in part. Overall student response rates were 36% in 2021 and 46% in 2023, ranging by section from 96% to 0, depending on how the survey was shared with students. Because of this variation in response, we do not directly compare courses. We were able to match 25 instructor responses to student sections across both years.

We also conducted on-site observations for single class sessions of 16 recitation sections and six blended sections. Along with detailed field notes, we recorded the elapsed time for the same core teaching activities as on the surveys (lecture, small group work, etc.). Summing the time for each activity enabled comparisons with instructors' self-report and student report.

Data Analysis

We computed three composite variables. The student ACL score (S-ACL) is the mean frequency of class discussion, student group work, and student presentation, for one student. The section ACL score is the mean S-ACL score for sections with ≥10 responses. The instructor ACL score (I-ACL) is the total percentage of class time for the same three activities. 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 items. To answer RQ3, we separated data sets by year.

Results

We used descriptive statistics to report on the behavior of the survey items, and used correlations and linear mixed models to answer RQs1-2.

Descriptive Statistics

Table 2 provides the mean and standard deviation for each composite item. On average, ACL practices were reported to occur 'seldom' and the distribution was moderately platykurtic, while student-reported WIHIC items occurred 'seldom' to 'sometimes' on average, and the scale was negatively skewed. Student reported 'moderate' to 'good' gains on average, and this scale was moderately negatively skewed.

Table 2. 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 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 are not expected to show an expected relationship or response pattern connecting individual practices. Therefore, we did not measure the internal consistency of ACL practices. The internal consistency of both Gains and WIHIC scales was good, with alpha coefficients of 0.91 and 0.97 respectively.

RQ 1a: How does student-reported use of ACL practices relate to learning environment?

We tested the relationship between student ACL (S-ACL) and WIHIC scales. From correlation analysis of all student cases, we found a moderate positive correlation (0.51), indicating that students report more positive environments in courses that include more active learning. We also tested the consistency of this relationship across disciplines and found similar, low to moderate correlations across math, physics, and CS (0.40, 0.54, 0.33, respectively). Figure 1 shows both the correlation of individual scores for all student responses that include responses to every item in the scale (L), and of section means for sections with at least 10 students (R).

Lastly, we conducted a linear mixed model to test this relationship with WIHIC scale as the dependent variable, student 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. For this analysis, we excluded student responses from sections with fewer than 10 responses, as larger class sizes have been shown to be more reliable (Benton & Cashin, 2014). The resulting analysis of 1640 student responses from 93 course sections revealed a moderate positive effect (0.33, t (73) = 18.64, $p < 0.001$) of ACL practices on WIHIC, when controlling for the random effect of course sections.

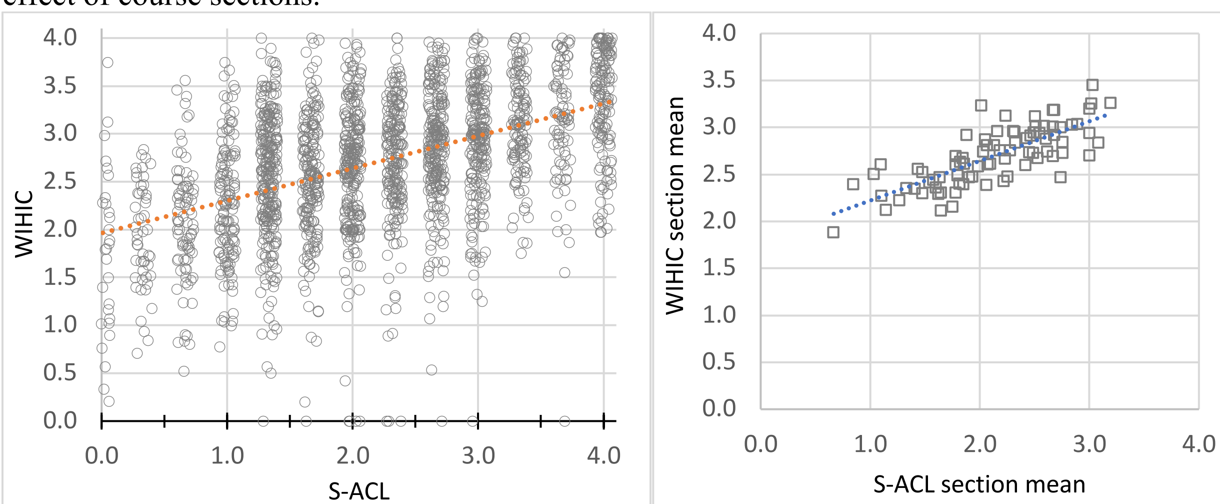

Figure 1. WIHIC score correlates positively with S-ACL, for individual students (L) and section means (N≥10) (R).

RQ 1b: How does student-reported use of ACL practices relate to students' learning gains?

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 ACL and gains, indicating that students report greater learning gains in courses that include more active learning. We tested the consistency of this relationship across disciplines and found similar minimal correlations for math and physics, 0.25, and 0.27, respectively, but no statistically significant relationship for CS.

We then conducted a linear mixed model to test this relationship with learning gains as the dependent variable, ACL as a fixed effect, and course section as a second-order random effect. As above, 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.

RQ 2a-b: How does student-reported use of ACL practices relate to instructor practices?

We tested the relationship between student- and instructor-reported use of ACL practices, by instructor and course, for the subset of cases where both types of data were available. We compared the mean student reports of ACL practices for each recitation section (section S-ACL) with that section's instructor response (I-ACL). A correlation analysis of section S-ACL and I-ACL 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 are well aligned.

Similarly, we tested the relationship between instructor-reported I-ACL, and observed proportion of time spent in ACL practices from a single class observation. In general, there is good correlation between what instructors say they do, and what we saw them do (0.76, $p = 0.16$). The outliers in Figure 2 (R) are two GTAs who used the whole observed class session for a guided inquiry task, but reported using much less time for ACL on average.

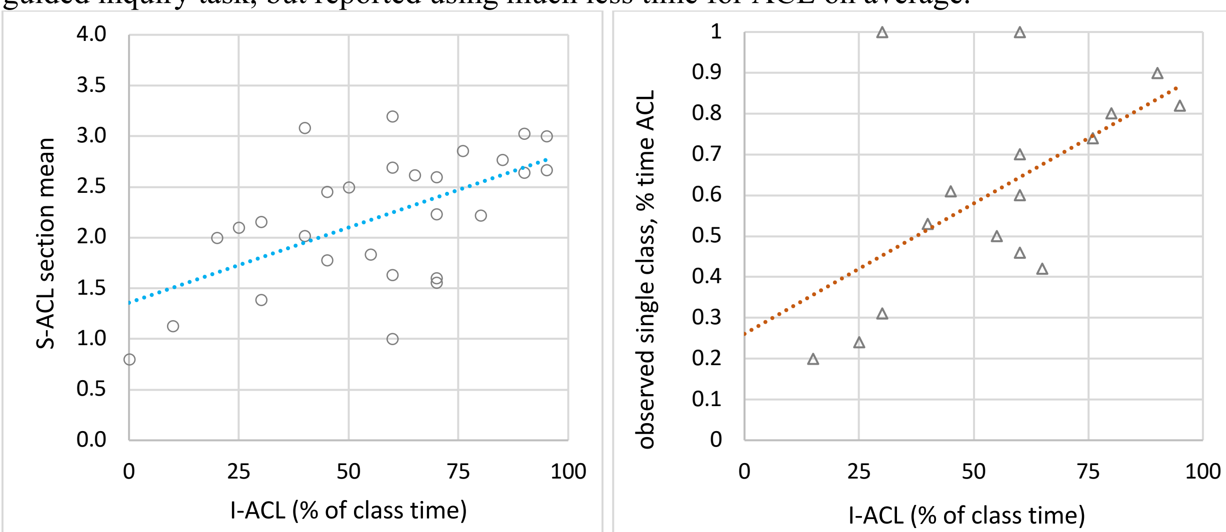

Figure 2. S-ACL section means correlate with their instructor's report, which in turn correlates with observation.

RQ 3: How can these metrics inform formative and summative program evaluation?

Figure 3 (left side) compares mean S-ACL for target courses actively involved in the project versus nontarget courses that had not yet implemented ACL in 2021. The S-ACL metric clearly distinguishes courses that had and had not implemented ACL with an effect size of 1.05.

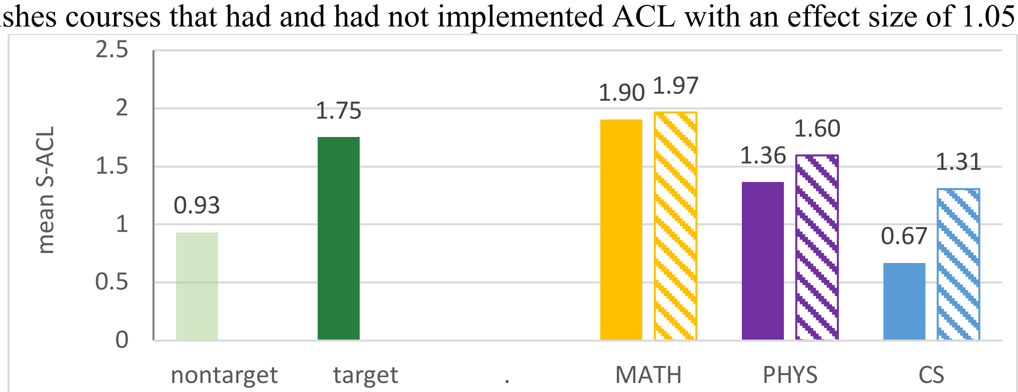

Figure 3. Project-wide and department means for S-ACL in 2021(solid), and 2023 (stripes)

Figure 3 (right side) shows department means for 2021 (solid) and 2023 (stripes). These student data corroborate the PIs' reports about implementation levels: math had already implemented ACL in target courses in 2021 and sustained that work in 2023; physics' implementation increased from 2021 to 2023; and CS was not active in 2021, but had begun to implement ACL by 2023.

Discussion

Students responded to the items in the TAMI-SS block and were able to discriminate them on the response scale provided. Nearly all students responded; we filtered out responses missing any of the items, as the composite variable S-ACL is not meaningful without answers to each item. Over 95% of students reported coming to class most or all of the time: they are observers who have experiences to report. The S-ACL scale is normally distributed, indicating that most students reported experiencing some ACL practices, with smaller proportions reporting either high or low levels of ACL practices. Analyses are ongoing, but we have found no systematic patterns of difference in response to these items by student subgroup (Apkarian, et al., 2024).

Drawing on student observations to assess implementation of ACL strategies has several affordances. There are many observers to describe the teaching experience across the full term, not just one day. The measure is simple, behavior-focused, and not evaluative. It aligns with other measures we have developed to facilitate triangulation among measures to characterize teaching, and/or to combine with other survey or non-survey measures (exams, grades, etc.). Student-provided evidence can be a golden spike linking a specific initiative to longer-term or harder-to-measure outcomes established in the literature (Brown Urban & Trochim, 2009). However, good response rates are needed to compare courses or sections or pre/post time points. Both student and instructor response rates limited the matching of data for course sections.

The correlation of S-ACL with WIHIC suggests a positive relationship between active learning and students' classroom experience. This alone may be a benefit, if having more positive experiences encourages students to persist in their STEM studies. Further, ACL is linked to stronger learning, as measured by the smaller positive correlation of S-ACL to learning gains. However, in this setting, the impact of these benefits may be modest, given that the intervention focuses on recitations only (and a small group of blended lecture-recitation classes). It remains to be seen if the culture of ACL will "trickle up" to lecture sections as the PI team postulated.

Results from the CS courses offer an example of how these data may be useful in program evaluation and research. The increase in S-ACL (Figure 3, far right) reflects a strong implementation signal as CS joined the project work. Slightly weaker relationships of S-ACL to student experience and learning variables in CS may reflect this earlier stage of implementation. Students may not perceive benefits of ACL teaching until instructors' ACL implementations are less tentative, consistent and clearly communicated. Indeed, class observations in CS displayed less frequent cooperation among students. Moreover, in write-in comments, CS students expressed stronger preference for learning on their own, compared to math and physics students. Instructors may need to actively reshape norms about learning in CS to strengthen student buy-in. Norms can shift quickly as instructors adopt active facilitation practices that students notice—and likely share via the student grapevine (Turpen & Finkelstein, 2010).

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. We do not recommend using this tool to evaluate individual instructors. Rather, it is useful to indicate progress on efforts to support instructors to shift their practices toward RBIS, at an institutional scale (Laursen, et al., 2019).

Acknowledgments

The study was funded by the National Science Foundation (NSF-DUE award #1821440). The authors thank the many instructors and students who took part in the study, and the GMU PI team for their assistance in acquiring data.

References

American Association for the Advancement of Science [AAAS]. (2013). *Describing and measuring undergraduate STEM teaching practices*. Washington, DC. https://drive.google.com/file/d/1Du7sCx3g3p21C9LwgnyJvmA5E9EosEU2/view?usp=sharing

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.

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.

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 Netherlands.

Brown Urban, J., & Trochim, W. (2009). The role of evaluation in research-practice integration Working toward the "Golden Spike." *American Journal of Evaluation*, *30*(4), 538-553.

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.

Campbell, C. M., Cabrera, A. F., Ostrow Michel, J., & Patel, S. (2017). From comprehensive to singular: A latent class analysis of college teaching practices. *Research in Higher Education*, *58*, 581-604.

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.

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.

Guskey, T. R. (2000). *Evaluating professional development*. Corwin Press.

Guskey, T. R. (2002). Does it make a difference? Evaluating professional development. *Educational Leadership*, *59*(6), 45-51.

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* (pp. 727-735). San Diego, CA: Mathematical Association of America, SIGMAA on RUME.

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.

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.

Hora, M. T., & Ferrare, J. J. (2013). *A review of classroom observation techniques in postsecondary settings* (WCER Working Paper 2013-1). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Iverson, E. (2012). *2012 AAAS Measurement Workshop – Faculty Surveys White Paper*. Commissioned white paper for the AAAS Measurement of Teaching Practices Meeting. Washington DC. http://serc.carleton.edu/files/serc/commissioned_paper_faculty_sur.pdf

Koch, A. K. (2017). It's about the gateway courses: Defining and contextualizing the issue. *Improving Teaching, Learning, Equity, and Success in Gateway Courses: New Directions for Higher Education*, *180*, 11-17. Wiley.

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*, 1-12.

Laursen, S., 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*. Washington, DC: American Association for the Advancement of Science. https://www.aaas.org/sites/default/files/2019-07/levers-for-change-WEB100_2019.pdf

Laursen, S., 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) Boulder, CO: University of Colorado Boulder, Ethnography & Evaluation Research, Assessment & Evaluation Center for Inquiry-Based Learning in Mathematics.

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.

Measures of Effective Teaching (MET) Project (2012). *Asking students about teaching: Student perception surveys and their implementation*. Policy and Practice Summary. Seattle: Bill & Melinda Gates Foundation.

Neumann, A. (2014). Staking a claim on learning: What we should know about learning in higher education and why. *The Review of Higher Education*, *37*(2), 249-267.

Prince, M., & Felder, R. (2007). The many faces of inductive teaching and learning. *Journal of College Science Teaching*, *36*(5), 14-20.

Schoenfeld, A. H. (1998). Toward a theory of teaching-in-context. *Issues in Education*, *4*(1), 1-94.

Seymour, E., Wiese, D., Hunter, A. & Daffinrud, S. M. (2000, March). Creating a better mousetrap: On-line student assessment of their learning gains. Paper presented at the National Meeting of the American Chemical Society, San Francisco, CA.

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.

Turpen, C., & Finkelstein, N. D. (2010). The construction of different classroom norms during Peer Instruction: Students perceive differences. *Physical Review Special Topics—Physics Education Research*, *6*(2), 020123.

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.

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.

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*, 1-14.