# A Measurement Hat Trick: Evidence from a Survey and Two Observation Protocols about Instructional Change after Intensive Professional Development

Sandra L. Laursen
U. Colorado Boulder

Tim Archie
U. Colorado Boulder

Timothy J. Weston
U. Colorado Boulder

Charles N. Hayward
U. Colorado Boulder

Stan Yoshinobu
U. Toronto

*Professional development about teaching is an important lever for change toward evidence-based instructional practices. Yet few studies in higher education offer robust evidence about whether and how instructors' teaching practices actually change as a result of professional development. We studied instructors who attended a four-day intensive workshop on inquiry-based learning (IBL) in college mathematics and who provided data about their teaching before and after. Of 293 instructors who attended, 136 provided complete survey data and 15 video-recorded their teaching, which we coded using two observation protocols. Comparing pre-workshop and follow-up teaching data, we find significant and well-corroborated changes toward IBL methods across all three methods, with effect sizes near 1, generally viewed as large in education research. The findings demonstrate that well-designed professional development can have a substantial impact on instruction even within the first year of implementation.*

*Keywords:* IBL, professional development, TAMI, RTOP, measurement of teaching

Teaching-related professional development (TPD) is widely seen as important to encourage the uptake of evidence-based instructional practices (EBIPs) by college STEM instructors. TPD is attractive as a lever for change because research has demonstrated that EBIPs can make a big difference in students' experiences and success, and tested classroom resources are widely available to instructors in all STEM disciplines. Proponents argue that TPD is impactful because most college instructors have ample room to improve: while deeply grounded in their discipline, most have little formal knowledge of learning and teaching. Moreover, the effects of TPD will be amplified as each instructor teaches many courses and thousands of students across a career.

However, these arguments are not currently well backed by research showing that TPD does have the desired effect on instructor practice. At present, professional developers and funders can point to examples of impactful TPD as proof of concept (e.g., Khatri et al., 2013; Laursen et al., 2019), but the formal literature relies much on case studies of small programs: large, high-quality studies are scarce. Exceptions include recent work on TPD in mathematics (Archie et al., 2022) and studies in several science fields (Bathgate et al., 2019; Borda et al., 2020; Chasteen & Chattergoon, 2020; Ebert-May et al., 2011, 2015; Manduca et al., 2017) that measure instructors' implementation of active approaches after TPD, and the barriers and supports that shape this.

Building knowledge about TPD in higher education has been slow and challenging. Simply measuring teaching and change in teaching is difficult (AAAS, 2013); common tools such as surveys and observations both offer pitfalls. Moreover, studies of TPD outcomes cannot be done as experiments: researchers must find large, well-designed, and stable TPD programs to serve as study sites to provide adequate sample sizes and a consistent, high-quality intervention. TPD is neither required nor routinely offered by institutions of higher education, unlike K-12 settings. Finally, teachers need time to learn and enact new ideas and practices, then more time to become proficient enough to improve student learning, a reality that further extends the research timeline.

Thus, it is no surprise that rigorous studies examining the outcomes of TPD are needed. Our study of workshops on inquiry-based learning (IBL) in college mathematics addresses many of these challenges. The workshops reached nearly 300 instructors, yielding samples large enough for quantitative analyses. Built on a decade of prior work, the workshop model was informed by learning from K12 and college TPD experience, evaluation feedback, and education research on professional development. Workshops were led by teams of skillful IBL educators who worked from a common general structure and schedule, drew from and contributed to a shared set of workshop and course materials, and engaged in TPD leadership training. Together these features ensured sound program design and consistent, high-quality TPD delivery.

To address the measurement challenges, we used three instruments—a survey and two observation protocols—and compared the results. Funding constraints meant we had to measure implementation just one year after workshop participation, likely before teachers had mastered new approaches. We thus chose two descriptive, behavior-focused measures (one survey, one observation protocol), positing that new IBL users would first adjust how they use class time—e.g., using less lecture and more of the student-centered activities that are common in established IBL classrooms (Laursen et al., 2014; Hayward et al., 2016). Even if instructors were not yet skilled in facilitating these activities, we could detect initial changes in their choices about the use of class time. To address quality of instruction, we chose a second observation protocol that offers an evaluative perspective via holistic ratings of each lesson. With thorough video sampling we could reliably characterize instruction for the course as a whole, not just in selected episodes (Weston et al., 2021). In this analysis, we consider: *What do three independent measures of teaching tell us about whether and how instruction changed after professional development?*

## Context for the Study

To situate the study, we provide brief descriptions of inquiry-based learning in mathematics, in the general form that was practiced by workshop leaders and taught in workshops, and of the IBL workshop design. For more detail, we refer readers to the references cited below.

Inquiry-based learning (IBL) is a form of instruction that has become increasingly common as an approach to active and collaborative learning in college mathematics. Laursen and Rasmussen (2019) identified four pillars, or core principles that IBL shares with other forms of inquiry-based mathematics education (IBME). In effective IBL classrooms, students engage deeply with meaningful mathematics and make sense of these ideas in collaborative work with their peers, which may take a variety of forms. Instructors support these student experiences by inquiring into student thinking—drawing out student ideas and helping them toward further sense-making—and by making design and facilitation choices that involve all in rigorous mathematical learning and identity-building. The "big tent" philosophy of IBL recognizes that instructors will implement these principles in different ways, depending on their setting, adjusting to suit their course material, class size, student audience, and personal preferences.

Consistent with other research on EBIPs (Freeman et al., 2014), IBL offers cognitive and affective benefits and supports student persistence and success in later courses (Kogan & Laursen, 2014; Laursen et al., 2014; Laursen & Rasmussen, 2019, & references cited therein). On average, IBL approaches were found to close equity gaps in student experience compared to lecture-heavy approaches (Laursen et al., 2014), again consistent with general findings in the literature (Theobald et al., 2020). Of course, IBME is no silver bullet (Ernest, Reinholz & Shah, 2019; Johnson et al., 2020; Reinholz et al., 2022): instructors must take proactive steps to develop self-awareness and empathy and to select pedagogical approaches that foster mathematical identity and enhance classroom climate (Dewsbury & Brame, 2019).

The design of the IBL workshops was purposeful, research-aligned, and iteratively refined (Yoshinobu et al., 2022). Briefly, four interwoven strands of workshop activity sought to foster IBL-supportive knowledge, skills, and beliefs and prepare instructors to implement IBL in their own teaching contexts. Video lesson study sessions helped instructors develop a mental model for an IBL classroom, while literature-to-practice sessions used provocative readings and discussion to help instructors internalize the four pillars and reflect on their own beliefs. "Nuts and bolts" sessions shared facilitation and task-design tactics to build instructor skills, and course planning sessions provided planning exercises and personal work time to help instructors consider how to prepare their own IBL course. Through an active e-mail mentoring program, peers and facilitators provided follow-up support—resources, ideas, trouble-shooting, encouragement—as instructors implemented their IBL course after the workshop (Hayward & Laursen, 2018). Deliberate attention to equitable instruction and assessment practices was increasingly incorporated into the workshops across the 2016-2019 period that we studied.

## Study Methods

The study design was approved by the university's IRB and all instructors gave informed consent. Students in video-recorded classes were given information explaining that the focus of the study was their instructor, not them. They could move out of camera view if they wished.

### Study Population and Survey Sample

The study examined participants in 11 IBL workshops conducted in 2016-2019 (Archie, Hayward & Laursen, 2021). Online workshops in 2020 were omitted from outcomes analysis. They were surveyed at two times: All 293 answered the pre-workshop survey, and 199 (68%) the follow-up survey about 17 months post-workshop. Of these, 136 (68%) identified a target course before the workshop and then described implementing IBL (or not) in their target course on the follow-up survey; they form our sample to study instructional change.

Overall, the implementing respondents were very similar to the workshop attendees. Half were women. Most (87%) were US citizens, nationals or residents; of these, 80% were white and 5% were Asian. About 3% described themselves as Hispanic or Latino/a/X. Participation was well balanced among people in tenured (30%), untenured (30%), and non-tenurable (30%) positions, from departments granting four-year (42%), masters (23%), and PhD (21%) degrees, and some from two-year (13%) colleges. About 23% taught at a minority-serving institution. Overall, they were modestly more diverse than the U.S. mathematics professoriate as a whole, and they came from institutions that broadly represent teaching settings in U.S. higher education.

### Survey Methods

**Survey instrument.** The survey included closed- and open-ended items previously described (Hayward & Laursen, 2014; Archie, Hayward & Laursen, 2021; Archie et al., 2022; Hayward et al., 2016). Some items were used to provide formative feedback to workshop leaders, to characterize workshop features, and to measure participants' knowledge, skills and attitudes. Participants self-reported personal and professional demographic information. Responses across time points were matched via unique identifiers. Here we focus on items used to describe teaching strategies at both times, from the TAMI-IS (Toolkit for Assessing Mathematics Instruction-Instructor Survey; Hayward, Weston & Laursen, 2018). These descriptive, behavior-focused items ask "*what did you do*?" rather than "*how well did you do it?* Taking the course as a whole, instructors estimated the frequency of use for each of 11 teaching practices often seen in college math courses (e.g., group work, whole class discussion, formal lecture, short explanations, student presentations), coded using the scale: 0 = never, 1 = once or twice during the term, 2 = about once a month, 3 = about twice a month, 4 = weekly, 5 = more than once a week, or 6 = every class. Open-ended items probed patterns in practices and special events and elicited text descriptions of 'lecture,' 'presentations,' and 'group work' as practiced in courses.

**Survey analysis**. We created a composite variable to measure instructors' overall use of IBL methods. This variable, IBL frequency score, includes five of the 11 practices classified as 'core IBL' practices because they measure all variations of IBL that were emphasized or de-emphasized in workshops and was calculated as follows: *IBL frequency = student group work + student presentation + class discussion - lecture - instructor problem-solving*. IBL frequency scores were calculated for pre-workshop and follow-up time points; scores could range from -12 to +18. Lower scores indicate less frequent use of IBL teaching practices and more frequent use of lecture and instructor problem-solving, while higher scores indicate more frequent use of IBL methods and less frequent use of lecture and instructor problem-solving. We conducted paired samples t-tests to measure change in IBL frequency between pre-workshop and follow-up times for instructors who did (n=15) and did not (n=136) participate in classroom observation.

**Observation Methods**

**Observation sample.** The observation sample was a subset of the survey sample: 15 workshop participants who registered early and volunteered to collect video data in their target course in the term before their summer workshop, and for the same or similar course the next academic year. Their personal and institutional characteristics are broadly similar to the survey sample. Instructors placed a video camera in their classroom several times a term and completed the survey at the end of the same term. We asked them to record about 1/3 of class time. In all, we coded 278 class sessions, averaging 9 classes per course and totaling over 300 hr class time.

**Observation instruments.** We applied two observation protocols that emphasize different dimensions (Hora & Ferrare, 2013a). The TAMI-OP (TAMI Observation Protocol) is a segmented, descriptive (behavior-focused) protocol that aligns with the TAMI-IS (Hayward, Weston & Laursen, 2018). It is adapted from the TDOP (Hora & Ferrare, 2013b) and COPUS protocols (Smith, Jones, Gilbert, & Wieman, 2013) to reflect practices we saw in college mathematics. Twenty activity codes for instructor and student behaviors were coded within each 2-minute segment. Fourteen end-of-class holistic items, designed by our team, query features of classroom atmosphere and student-teacher interaction (Laursen et al., 2011; Appendix 2).

The RTOP (Reformed Teaching Observation Protocol) is a holistic, evaluative protocol developed for K12 science instruction and increasingly used in higher education (Sawada et al., 2002). Raters score the lesson on 25 items in 5 categories: lesson design, teaching of content and procedural knowledge, communication, and student-teacher relationship. For each lesson, items are scored 0-4 to describe how often each feature is seen: "never occurred" to "very descriptive" of the lesson. Summed RTOP scores of 0-100 indicate the degree to which a course is "reformed" toward student-centered, evidence-based instruction emphasizing inquiry.

**Observation analysis**. Using data from the TAMI-OP and RTOP instruments, two global measures were used to assess change in teaching:

1. the RTOP-Sum (RTOP), a composite of 25 observational survey questions. Here ratings are summed over all questions and averaged over observed classes in a course.
2. the Proportion of Non-Didactic lecture classes (PND) across a term of observation using the TAMI-OP. PND provides an estimate of the proportion of classes, for one instructor, that use non-lecture teaching methods.

The analytical method for TAMI-OP is based on Latent Profile Analysis (LPA), a statistical clustering method applied to a large observation data set (790 observations of 74 teachers) from this and other projects (Weston, Hayward & Laursen, 2023). The LPA yielded four reliable class types (Didactic Lecture, DL; Interactive Lecture and Review, IL; Student Presentation, SP; Group Work, GW) based on the proportions of various teaching activities present in an individual class session. From the LPA results, we created a simple dependent measure to characterize a teacher's course based on the proportion of all observed classes for each instructor (over an academic term) that were *not* classified as didactic lecture (PND, proportion non-didactic). A paired-sample t-test was used to assess pre-post change in the observational indicator. We used the Cohen's D effect size for the analyses, which is the difference in pre/post means divided by the standard deviation of the difference.

## Results

We analyzed the data sets for pre- to post-workshop changes for all three measures and summarize the results in Table 1.

*Table 1: Pre- to post-workshop changes in instruction using three measures from survey (S) & observation (O)*

| S/O | Measure | N | Mean, pre | Mean, follow-up | Effect size | t | df | p |
|---|---|---|---|---|---|---|---|---|
| S | IBL frequency from TAMI-IS | 136 | 0.03 | 6.34 | 0.88 | 10.31 | 135 | < 0.001 |
| S | IBL frequency from TAMI-IS, observation subset only | 15 | 2.27 | 8.80 | 0.96 | 3.71 | 14 | 0.002 |
| O | PND from TAMI-OP | 15 | 0.58 | 0.81 | 0.80 | 3.13 | 14 | 0.004 |
| O | RTOP-Sum | 15 | 37.89 | 55.06 | 1.17 | 4.58 | 14 | <0.001 |

*The range for each measure is IBL frequency -12 to +18; PND 0-1; RTOP-Sum 0-100.*

### Instructional Change Based on Surveys

Figure 1 shows the distribution of IBL frequency scores at pre-workshop and follow-up, for both the full survey sample and the observation subgroup. A paired samples t-test showed a statistically significant increase in IBL frequency from pre-workshop to follow-up (Table 1).

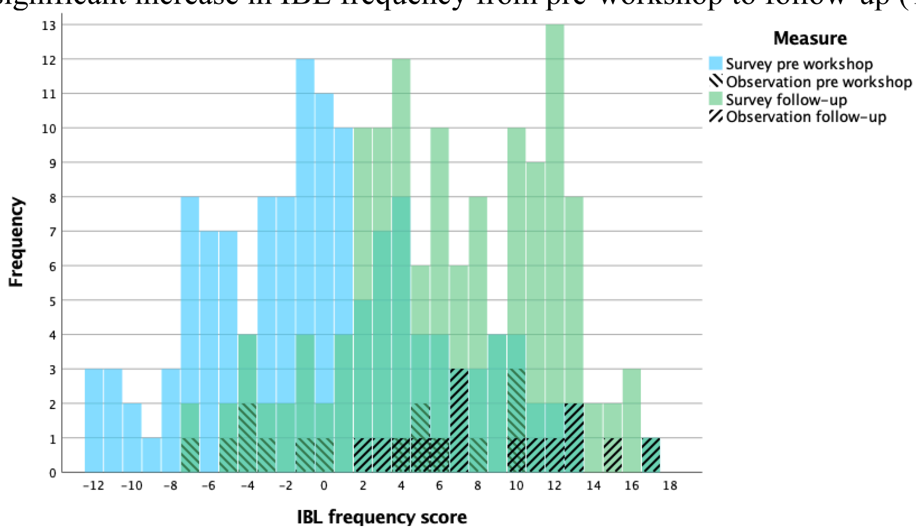


*Figure 1. IBL frequency score from surveys, for full sample and observation subgroup at two times.*

For video study participants, the change in IBL frequency score from pre-workshop to follow-up was also statistically significant, and comparable in magnitude to the average shift among all survey respondents (see Table 1). Video study participants used slightly more IBL core practices both before the workshop and subsequent to it, but these differences between the main group and video subgroup were not significant.

Half the main sample, and all 15 video participants, self-described as implementing at least "some IBL methods;" one third of the main sample and half of video participants described their implementation as a "full" IBL course. These are lower limits to the overall implementation rate, as 32% of all workshop attendees did not report on implementation. Their target courses served first-years to seniors; math majors (28%), mixed STEM majors (42%), non-STEM majors (13%) and pre-service teachers (4%). About 90% implemented IBL in a class of 35 students or fewer.

**Instructional Change Based on Observations.**

In general we saw gains in the average scores for both observation measures from pre to follow-up. Comparing these using paired samples t-tests, the changes in both the RTOP-Sum and the PND were statistically significant. For both measures, negative or no change was observed only for small numbers of instructors who had initial high scores. Figure 2 illustrates pre-post changes in PND for the 15 instructors. Each observed class session fits a latent profile (one colored cell) and PND for each course is based on the set of such profiles (one row of cells).

We also used LPA to examine changes in teaching approach. In post-TPD classes, instructors used less continuous lecture and more interactive lecture and group work, but also made less use of student presentations. Participants also adjusted the sequencing of activities. For example, changes in the timing of lecture and group work suggest a more structured and strategic approach to group work, with more consistent use of short lectures to summarize or signpost.

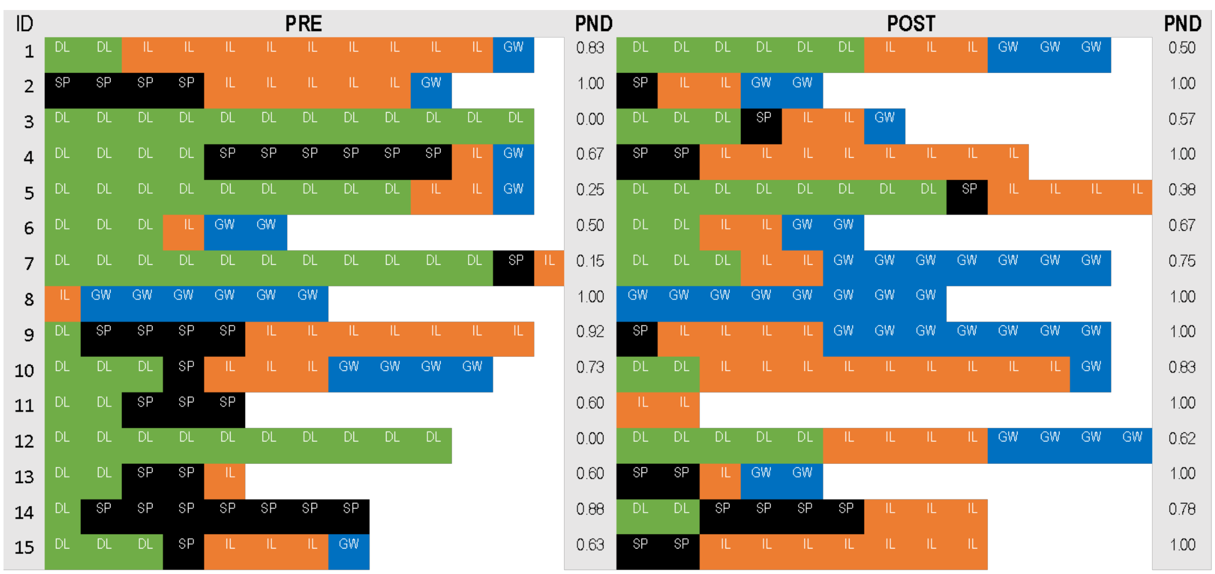


*Figure 2. Latent profiles and PND for 15 observed courses for the same instructors (IDs 1-15) at pre-workshop and follow-up ("post") times. Cells represent 1 observed class fitting a profile from LPA. DL (green) = didactic lecture; SP (black) = student presentation; IL (orange) = interactive lecture & review; GW (blue) = group work.*

## Discussion

This study identifies significant changes in the teaching practices of instructors after PD, using a rigorous approach to measurement. We found significant changes using three distinct measures. The effect sizes are impressively large (0.8 to 1.2) for an educational intervention. We used well characterized tools and large samples, and drew on different methods (survey vs. observation) with different foci (behavioral vs. evaluative). The TAMI-IS and TAMI-OP offer evidence that instructors' classroom practice shifted toward inquiry-based methods after TPD, and the RTOP-Sum offers evidence that post-TPD classes offered more student-centered, engaging and encouraging environments that were better aligned with evidence-based practice.

Survey responses from the subset of 15 instructors who were observed reflect a significant pre/post change similar to that seen in the larger survey sample. This indicates that such changes in teaching are not outliers, due to self-selection for video observation, but typical of the workshop's impact on instructors' teaching. Robust time sampling ensured that the coded observations were not "dog and pony shows" showing only the most active teaching days.

Other survey data enable us to attribute these changes to the workshop. Following the workshop, we observed robust and sustained growth in participants' knowledge and skills to use IBL; their motivation to use IBL and belief in its effectiveness remained strong and stable over time. In this sample and a larger group of IBL workshop participants, such growth in knowledge and skills, or "IBL capacity," was a good predictor of participants' later implementation of IBL (Archie et al., 2022). That is, changes in teaching behaviors are linked to the nature and strength of the gains that participants reported and sustained. Thus, the observed changes can be ascribed to participants learning and applying ideas they gained at the workshop.

Other studies of substantial TPD interventions for college instructors offer context for interpreting these results. Two studies, like ours, compared individuals to themselves before and after a TPD intervention. Chasteen and Chattergoon (2020) used the PIPS survey (Walter et al., 2016) to measure pre/post change in teaching practices of early-career physics faculty after a multi-day workshop, with an effect size of 0.8. In contrast, Ebert-May et al. (2011) reported no significant changes in RTOP scores after their TPD program for early-career biologists.

Two more studies compared teaching practices of TPD participants to samples of people who did not undergo TPD, and provided the data needed to compute effect sizes for mean differences between treated and untreated groups. After refining their biology TPD workshop, Ebert-May et al. (2015) measured differences in RTOP scores for participants vs. non-participants with an effect size of 0.68. Manduca et al. (2017) also used RTOP scores to compare treated (TPD) and untreated groups of Earth science instructors, with an effect size of 1.06. Both types of studies suggest that the changes seen here are impressive but not implausible—in line with what others have measured for STEM instructors who have taken part in thoughtfully designed, well supported, multi-day programs offered by skilled faculty developers in their discipline.

## Conclusions

We report a "hat trick" of TPD measurement: three independent measurements of instructional change after a well-designed TPD intervention. The changes are large, statistically significant and well corroborated, and can be ascribed to the workshop. We propose that the workshop has this substantial impact due to its strategic approach (Kennedy, 2016), providing instructors with a set of principles and tools to achieve a desired teaching goal, thus empowering them to adapt their practices in ways sensitive to their own context and preferences (Yoshinobu et al., 2022).

## Acknowledgments

The authors thank the enthusiastic facilitators who led the workshops and supported participants as they planned and implemented their IBL courses. This work was funded by the National Science Foundation under awards DUE-1525058 and DUE-1525077.

## References

American Association for the Advancement of Science (AAAS) (2013). *Describing & measuring undergraduate STEM teaching practices: A report from a national meeting on the measurement of undergraduate science, technology, engineering and mathematics (STEM) teaching*. Retrieved from https://cns.utexas.edu/images/CNS/Deans_Office/ResearchFacilities/strategic_initiatives/measuring-stem-teaching-practices.pdf

Archie, T., Hayward, C., & Laursen, S. (2021, July). *Collaborative Research: PROfessional Development and Uptake through Collaborative Teams (PRODUCT) Supporting Inquiry Based Learning in Undergraduate Mathematics. Cumulative Evaluation Report: 2016-2020*. [Report to AIBL] Boulder, CO: Ethnography & Evaluation Research, U. Colorado Boulder. https://www.colorado.edu/eer/content/archie-cumulative-quant-report-product-workshops-july2021

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.

Bathgate, M. E., Aragón, O. R., Cavanagh, A. J., Waterhouse, J. K., Frederick, J., & Graham, M. J. (2019). Perceived supports and evidence-based teaching in college STEM. *International Journal of STEM Education*, *6*(1), 1-14.

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.

Chasteen, S. V., & Chattergoon, R. (2020). Insights from the Physics and Astronomy New Faculty Workshop: How do new physics faculty teach? *Physical Review Physics Education Research*, *16*(2), 020164.

Dewsbury, B., & Brame, C. J. (2019). Inclusive teaching. *CBE—Life Sciences Education*, *18*(2), fe2.

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. *BioScience*, *61*(7), 550-558.

Ebert-May, D., Derting, T. L., Henkel, T. P., Middlemis Maher, J., Momsen, J. L., Arnold, B., & Passmore, H. A. (2015). Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development. *CBE—Life Sciences Education*, *14*(2), ar22.

Ernest, J. B., Reinholz, D. L., & Shah, N. (2019). Hidden competence: Women's mathematical participation in public and private classroom spaces. *Educational Studies in Mathematics*, *102*(2), 153-172

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.

Hayward, C. N., Kogan, M., & Laursen, S. L. (2016). Facilitating instructor adoption of inquiry-based learning in college mathematics. *International Journal of Research in Undergraduate Mathematics Education, 2*(1), 59-82.

Hayward, C., & Laursen, S. (2014). *Collaborative research: Research, dissemination, and faculty development of inquiry-based learning (IBL) methods in the teaching and learning of mathematics, Cumulative evaluation report: 2010-2013*. Boulder, CO: University of Colorado Boulder, Ethnography & Evaluation Research. http://www.colorado.edu/eer/research/documents/IBLCentersCumulativeWorkshopReportALL_2014.pdf

Hayward, C. N., & Laursen, S. L. (2018). Supporting instructional change in mathematics: Using social network analysis to understand online support processes following professional development workshops. *International Journal of STEM Education*, *5*(28). https://doi.org/10.1186/s40594-018-0120-9

Hayward, C. N., Weston, T. J., & Laursen, S. L. (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. http://sigmaa.maa.org/rume/RUME21.pdf

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

Hora, M. T., & Ferrare, J. J. (2013). Instructional systems of practice: A multidimensional analysis of math and science undergraduate course planning and classroom teaching. *Journal of the Learning Sciences*, *22*(2), 212-257.

Johnson, E., Andrews-Larson, C., Keene, K., Melhuish, K., Keller, R., & Fortune, N. (2020). Inquiry and gender inequity in the undergraduate mathematics classroom. *Journal for Research in Mathematics Education*, *51*(4), 504-516.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945-980.

Khatri, R., Henderson, C., Cole, R., & Froyd, J. (2013). Successful propagation of educational innovations: Viewpoints from principal investigators and program directors. In *AIP Conference Proceedings* (Vol. 1513, No. 1, pp. 218-221). American Institute of Physics.

Kogan, M., & Laursen, S. L. (2014). Assessing long-term effects of inquiry-based learning: A case study from college mathematics. *Innovative Higher Education, 39*(3), 183-199.

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.

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 of Research in Mathematics Education*, *45*(4), 406-418

Laursen, S. L., & Rasmussen, C. (2019). I on the prize: Inquiry approaches in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education, 5*(1), 129-146.

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.

Reinholz, D., Johnson, E., Andrews-Larson, C., Stone-Johnstone, A., Smith, J., Mullins, B., Fortune, N., Keene, K., & Shah, N. (2022). When active learning is inequitable: Women's participation predicts gender inequities in mathematical performance. *Journal for Research in Mathematics Education*, *53*(3), 204-226.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., et al. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics*, *102*(6), 245-253.

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.

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., ... & 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.

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.

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., Hayward, C. N., & Laursen, S. L. (2023). Measures of success: Characterizing teaching and teaching change with segmented and holistic observation data. Manuscript in review.

Yoshinobu, S., Jones, M. G., Hayward, C. N., Schumacher, C., & Laursen, S. L. (2022). A broad doorway to the big tent: A four-strand model for discipline-based academic development on inquiry-based learning. *PRIMUS*. https://doi.org/10.1080/10511970.2022.2072427