



Does Increased Agency Improve the Effectiveness of Self-Directed Professional Learning for Educators?

Dorottya Demszky
Stanford University

Heather C. Hill
Harvard University

Eric S. Taylor
Harvard University

Ashlee Kupor
Stanford University

Deepak Varuvel Dennison
Cornell University

Chris Piech
Stanford University

The role of teacher agency in professional learning has been the subject of several qualitative studies but has not yet been tested in an experimental setting. To provide causal evidence of the impact of teacher agency on the effectiveness of professional learning, we conducted a preregistered randomized controlled trial in an online computer science course with volunteer instructors who teach students worldwide. All instructors (N=583) received automated feedback on their instruction throughout the course, with half randomly assigned to have choice over the feedback topic. While choice over feedback topic alone did not significantly impact instructors' engagement with feedback or measured changes in their instruction, it led to improved student attendance---an effect that was strongest for instructors who actively engaged with additional professional learning resources, including training modules and teaching simulations. For this motivated subset of instructors, having choice over feedback had significant positive impacts on both their instruction and student outcomes compared to the control group. These findings suggest that agency in professional learning may be most effective when combined with instructors' intrinsic motivation to pursue self-directed improvement. Our study paves the way for further empirical investigations into when and how agency can be effectively integrated into professional learning systems.

VERSION: March 2025

Suggested citation: Demszky, Dorottya, Heather C. Hill, Eric S. Taylor, Ashlee Kupor, Deepak Varuvel Dennison, and Chris Piech. (2025). Does Increased Agency Improve the Effectiveness of Self-Directed Professional Learning for Educators?. (EdWorkingPaper: 25 -1162). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/04kc-7085>

Does Increased Agency Improve the Effectiveness of Self-Directed Professional Learning for Educators?

Dorottya Demszky^{1*}, Heather C. Hill², Eric Taylor², Ashlee Kupor³
Deepak Varuvel Dennison⁴, Chris Piech³

¹Graduate School of Education, Stanford University

²Graduate School of Education, Harvard University

³Computer Science Department, Stanford University

⁴School of Information Science, Cornell University

Abstract

The role of teacher agency in professional learning has been the subject of several qualitative studies but has not yet been tested in an experimental setting. To provide causal evidence of the impact of teacher agency on the effectiveness of professional learning, we conducted a preregistered randomized controlled trial in an online computer science course with volunteer instructors who teach students worldwide. All instructors ($N=583$) received automated feedback on their instruction throughout the course, with half randomly assigned to have choice over the feedback topic. While choice over feedback topic alone did not significantly impact instructors' engagement with feedback or measured changes in their instruction, it led to improved student attendance—an effect that was strongest for instructors who actively engaged with additional professional learning resources, including training modules and teaching simulations. For this motivated subset of instructors, having choice over feedback had significant positive impacts on both their instruction and student outcomes compared to the control group. These findings suggest that agency in professional learning may be most effective when combined with instructors' intrinsic motivation to pursue self-directed improvement. Our study paves the way for further empirical investigations into when and how agency can be effectively integrated into professional learning systems.

Keywords: instructor agency; automated feedback; self-directed professional learning; randomized controlled trial

1 Introduction

Teacher agency in choosing or shaping professional learning opportunities is thought to promote teacher learning and changes in practice (Calvert, 2016; Diaz-Maggioli, 2004; Lieberman & Pointer Mace, 2008; Smith, 2017, *inter alia*). Advantages of allowing teacher choice and input when it comes to professional learning content include the potential for better alignment with teacher needs, enhanced engagement, and increased follow-through as teachers more deliberately change their practice (M. Kennedy, 2016). Weaving teacher agency into learning opportunities has been endorsed by many leading scholars of teacher learning (Carter Andrews & Richmond, 2019; Zeichner, 2019), and has also been embedded in common coaching and professional learning community protocols, as when teachers choose their own focus of improvement or generate their own solutions to problems of practice.

*Corresponding author. Email: ddemszky@stanford.edu

However, many professional learning offerings lack teacher input into and choice over content. Across all forms of professional learning, teachers report having only partial control over these opportunities (Doan et al., 2021). Even in professional learning communities—typically school-based teams of teachers working together to meet their students’ needs—teacher choice over learning content is present only about half the time (Zuo, Doan, & Kaufman, 2023). Furthermore, teacher agency often conflicts with prevalent approaches that “diagnose” classroom problems and apply prescribed “solutions” through teacher professional learning (Biesta, Priestley, & Robinson, 2015)—in fact, this remains the dominant form of instructional improvement in many Western nations today.

Remarkably, despite the widespread belief that teacher agency matters in professional learning, there exists little empirical evidence on the topic. The evidence that does exist consists largely of illustrative cases or teacher self-reports (Brodie, 2021; Martin, Kragler, Quatroche, & Bauserman, 2019; Philpott & Oates, 2017), rather than carefully controlled comparative or randomized studies. To address this gap, we conducted a preregistered randomized controlled trial¹ examining whether giving instructors agency over the type of feedback they receive about recorded lessons impacts their engagement with the feedback, their instructional practice, and student outcomes. Specifically, our study sought to answer the following research questions:

1. Does providing instructors choice over feedback impact their engagement with the feedback, their perception of the feedback, or their teaching practice?
2. Does choice over feedback for instructors impact their students’ outcomes?
3. How do treatment effects vary by instructor demographics and whether the instructor engages with self-directed professional learning beyond automated feedback (i.e., training modules, teaching simulations)?

We conducted this randomized controlled trial within Code in Place, a free online introductory programming course with volunteer instructors who teach students worldwide. All instructors ($N=588$) received automated feedback on their teaching based on natural language processing analysis of their section recordings. We randomly assigned instructors to the treatment or control group. Those in the treatment group chose which feedback topics they would receive throughout the course; control instructors were randomly assigned feedback topics to match the distribution and sequence chosen by the treatment group. Thus, in expectation, treatment and control instructors differed only on whether they *chose* the feedback topics they received, as both groups received feedback on the same topics. By offering the first causal evidence on the impact of agency, our study informs theory and practice related to the design of effective teacher professional learning systems.

2 Related Work

2.1 Teacher Agency

Most broadly, *teacher agency* can be defined as “teachers’ capacity to make choices, take principled action, and enact change” (Anderson, 2010, p. 541); it often refers to control over various aspects of job-related tasks, including addressing student needs and choosing curriculum materials (Priestley, Biesta, Philippou, & Robinson, 2015). Teacher agency is often contrasted with top-down approaches to the conduct of teaching tasks and instructional improvement, for instance when states mandate teaching standards or districts mandate curriculum materials. As this implies, teacher agency is partly an individual phenomenon, but it is also interactive with the system within which teachers work (Molla & Nolan, 2020).

¹Preregistration included at <https://www.socialscienceregistry.org/trials/12746>

2.2 Teacher Agency in Professional Learning

Clarke and Hollingsworth (2002) introduced teacher agency into debates about teacher learning and professional learning. In particular, they note that, in contrast to literature that frames teacher learning as training, “The key shift is one of agency: from programs that *change* teachers to teachers as *active learners* shaping their professional growth” (p. 948, italics added). Many argue that teacher agency—whether operationalized as choice to engage in specific professional learning offerings or opportunities to shape the topics of professional learnings—is fundamental to teacher learning (Lieberman & Pointer Mace, 2008; Smith, 2017, inter alia).

Arguments for teacher agency in shaping or choosing professional learning often rest on theories of adult learning (Knowles, 1984; Merriam et al., 2001), which argue that, in contrast to children, adults see themselves as agentic and thus best direct their own learning. In this view, motivation to learn is key—adult learning is typically voluntary and will not occur without the full engagement of the learner. Together with the observation that most adults have reservoirs of experience to draw on while learning, this view motivates several broad forms of professional learning, including teacher reflection (Schön, 1983), action research (Morales, 2016), teacher study groups (Stanley, 2011, inter alia), and teacher professional communities (Stoll, Bolam, McMahon, Wallace, & Thomas, 2006). More recently, scholars have studied teacher agency within professional learning in its own right: As Vähäsantanen, Hökkä, Paloniemi, Herranen, and Eteläpelto (2017) report, “professional agency and supportive social affordances for its enactment are essential to the processes of work-related learning and organisational development” (p. 514).

Molla and Nolan (2020) suggest two types of teacher agency particularly relevant to teacher professional growth. In the first, inquisitive agency, teachers choose their own professional learning experiences. This corresponds to teacher choice about which and how much professional learning to attend; it extends, as in this study, to choosing areas for feedback and growth. In the second, deliberative agency, teachers engage in active reflection and refinement of their practices. This corresponds to taking up opportunities to learn as presented, in this case using professional learning material and feedback to drive one’s own improvement.

2.3 Empirical Studies of Teacher Agency in Professional Learning

Despite strong theoretical warrants for studying the role of teacher agency in professional learning, relatively few empirical studies on this topic exist. Case studies have explored how shifts in education systems towards more prescriptive approaches to school improvement (e.g., through high-quality curriculum materials) have impacted teachers’ perceived control over their professional learning (Lloyd & Davis, 2018; Mohammad Nezhad & Stolz, 2024). For instance, interviews with Australian teachers conducted by Mohammad Nezhad and Stolz (2024) indicate that many felt a lack of voice in typical, school-directed professional learning experiences. When choice is not allowed, these authors argue, teachers may limit their active engagement in the professional learning, experience the stifling of professional culture, and curtail their changes in practice.

Researchers have applied various methods, including case studies, randomized experiments and interviews to investigate the effectiveness of mandated vs voluntary teacher professional learning. A review of 28 studies by M. M. Kennedy (2016) suggests that mandatory teacher professional learning does not effect changes in instruction or student outcomes, while Lynch, Hill, Gonzalez, and Pollard (2019) finds no difference in effectiveness between mandated vs voluntary professional learning. Case studies and interviews often note that teachers who are engaged in professional learning frequently choose actively to do so (Philpott & Oates, 2017), but these investigations rarely compare learning situations chosen by teachers to learning situations not chosen by teachers.

2.4 Instructional Feedback

The teaching profession has long recognized that regular, formative feedback is vital for professional growth (Hattie & Timperley, 2007; Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009; Hill, 2009). Yet, questions remain about how to optimize such feedback systems. Theory suggests that giving teachers choice about which aspects of their practice to focus on when receiving feedback could improve engagement and learning (Clarke & Hollingsworth, 2002; Molla & Nolan, 2020). However, empirical evidence on the impact of agency in feedback systems is limited, likely due to the small scale of most of these studies (Kraft, Blazar, & Hogan, 2018) and to the challenges relating to variations in agency within a single feedback model.

Computational approaches, particularly those leveraging natural language processing, create new possibilities for studying how teacher agency affects the uptake and effectiveness of feedback. Because they automatically analyse multiple aspects of instruction simultaneously—such as questioning patterns (Kelly, Olney, Donnelly, Nystrand, & D’Mello, 2018; Jensen et al., 2020), dialogic teaching strategies (Suresh et al., 2021a), and responsiveness to student contributions (Demszky et al., 2021)—feedback tools can offer teachers choice over their focus areas while maintaining consistency in measurement. The scalability of computational methods also allows for larger-scale experimental studies in this area.

Early implementations of automated feedback systems show the promise of these tools. Automated feedback has been found to improve teacher practice in targeted areas, such as increasing student talk time (Wang, Miller, & Cortina, 2013; Demszky, Wang, Geraghty, & Yu, 2024), uptake of student contributions (Demszky, Liu, Hill, Jurafsky, & Piech, 2023; Demszky & Liu, 2023), use of focusing questions (Demszky, Liu, Hill, Sanghi, & Chung, 2024) and other dialogic practices (Jacobs et al., 2022). Many of these studies were conducted in online contexts, where digital platforms facilitate the recording and analyses of classroom interactions. Furthermore, the feedback tools used in these studies, including LENA, M-Powering Teachers, the Talk Moves application, and TeachFX, foster teacher agency by design, as the feedback is descriptive rather than evaluative, and because teachers engage with it on their own time as a way to self-reflect. At the same time, these feedback systems have so far not allowed for the user to customize the type of feedback they receive, and hence the utility of such customization remains unexplored. This study seeks to fill the gap by building and testing the impact of a feedback system that allows for such customization.

3 Study Background

We conducted the study during the spring of 2023 as part of Code in Place, a free, online, 6-week-long introductory programming course. Anyone could apply to serve as a volunteer section leader (henceforth, instructor) for the course by submitting a programming exercise and a 5-minute video of themselves teaching; course organizers selected instructors based on this application. Students applied by completing several lessons and assignments. Each instructor was assigned to 12.1 ± 2.1 students.² Once per week (between Wednesday and Friday), instructors held a 1-hour session for students in their group to cover material and answer questions related to lectures and assignments from the course. The materials were prepared by course organizers and were uniform across instructors. Sessions took place on Zoom and were recorded and automatically transcribed by Zoom’s built-in transcription service. Instructors received automated feedback based on their transcripts;

²Students were assigned to instructors prerandomization, using the following process: (1) instructors selected their preferred time slots; (2) students chose available time slots; (3) within each time slot, students were assigned to sections randomly, with one exception—instructors and section leaders were sorted by age, so older students were assigned to older instructors, and underage students (<18) were never paired with adult [18+] instructors.

half of the instructors, as described below, had the option to choose among different types of feedback. The study was approved under institutional IRB.

3.1 Participants

Our participant sample consisted of all adult (18+) instructors in Code in Place ($N=583$). Table 1 shows the demographics of our analytic sample, based on information that Code in Place collected during the instructor application process. In terms of gender, 66% of instructors identified as male, 32% as female, 1% as nonbinary, and 1% as other or "prefer not to say." The instructors ranged in age from 18 to 75 years, with an average of roughly 30 years old. They were located in 70 unique countries, with about 48% in the United States; three quarters (75%) were first-time instructors for Code in Place in 2023. Based on their open-ended responses about their background, the majority were young professionals working in the technology industry and had limited prior teaching experience. Their motivation to teach came from wanting to help beginners overcome their fears, be part of a supportive global community, and "pay forward" the education they once received. Many also saw teaching as a way to improve their communication, leadership, and coding skills.

Our analytic sample also included 8,254 students who were taught by these instructors. Students were more balanced in terms of gender than instructors, with 52% identifying as female, 45% as male, 1% as nonbinary, and 2% as other or "prefer not to say." Students were on average 31 years old, and they were located in 145 unique countries, with about 28% in the United States.

3.2 Automated Feedback to Instructors

All instructors received automated feedback on their instruction during the weekend following each session. The feedback was generated via a three-step process: First, we used Zoom to record and automatically transcribe the session; next, we analysed the transcripts using a set of natural language processing models; then, we used the results of these analyses to generate feedback for instructors. We describe the latter two steps below.

Transcript analysis. We developed and applied measures to identify three instructor moves in session transcripts: GETTING IDEAS ON THE TABLE (e.g., "Who would like to share their solution?"), BUILDING ON IDEAS (e.g., "Can you explain why you used a 'for' loop?"), ORIENTING STUDENTS TO ONE ANOTHER (e.g., "Bryan and Jen used a similar approach—do you see how?"). These moves were inspired by the Accountable Talk framework (O'Connor, Michaels, & Chapin, 2015), which proposes these moves to be ones that facilitate students' active participation in learning. Our choice for these moves, henceforth referred to as *feedback topics*, was also motivated by the fact that the teacher education team for Code in Place in 2023 decided to create training modules for these moves. By creating measures that correspond to these modules, we hoped to create a consistent and more holistic professional learning experience for instructors that built on their initial training.

Prior work has employed the Accountable Talk framework to develop automated measures based on K–12 transcripts (Suresh et al., 2021b; Jacobs et al., 2022). Kupor, Morgan, and Demszky (2023) leveraged transcripts from a prior Code in Place course (from 2021) to develop such measures for the Code in Place domain. They annotated 2,000 instructor moves for five talk moves (eliciting, revoicing, adding on, probing, and connecting students' ideas) and fine-tuned RoBERTa (Liu, 2019) models to create classifiers for these moves. The feedback topics in this study correspond to models by Kupor et al. (2023) in the following way: GETTING IDEAS ON THE TABLE corresponds to "eliciting," BUILDING ON IDEAS corresponds to "revoicing," "probing" and "adding on," and ORIENTING STUDENTS TO ONE ANOTHER corresponds to "connecting."

Variable	Mean/%	SD
A. Instructor Characteristics		
Number of instructors	583	
Female	31.7%	
Age	30.153	12.076
First time Code in Place instructor	74.6%	
In United States	47.9%	
In India	14.1%	
In Great Britain	3.9%	
In Canada	3.6%	
In Bangladesh	3.4%	
In other country	27.1%	
B. Student Characteristics		
Number of students	8,254	
Female	51.9%	
Age	31.357	10.091
In United States	27.5%	
In Bangladesh	8.0%	
In India	7.9%	
In China	5.6%	
In Canada	4.3%	
In Great Britain	3.9%	
In Turkiye	3.2%	
In other country	39.6%	

Table 1: Descriptive statistics of the analytic sample. We group countries into the “other” category if less than 3% of instructors/students were located in that country.

We additionally used GPT-3 (specifically text-davinci-003³) to generate experimental insights into a summary of what happened during the class as well as specific moments when the instructor or student exhibited curiosity. This feedback was separate from feedback on talk moves, and we only provided these insights to a subset of instructors during their last 2 weeks in the course. The supplement includes the prompt we used to generate these insights.

Displaying feedback to instructors. Similar to prior work on automated feedback to educators (Demszky et al., 2023; Jacobs et al., 2022), our goal was to generate nonjudgmental, concise, and actionable feedback to instructors that would encourage self-reflection as a mechanism for instructional improvement. Feedback was private to each instructor, and it always focused on a single topic at a time. Instructors received feedback on

³This was the best performing cost effective GPT model available at the time of the study (spring 2023).

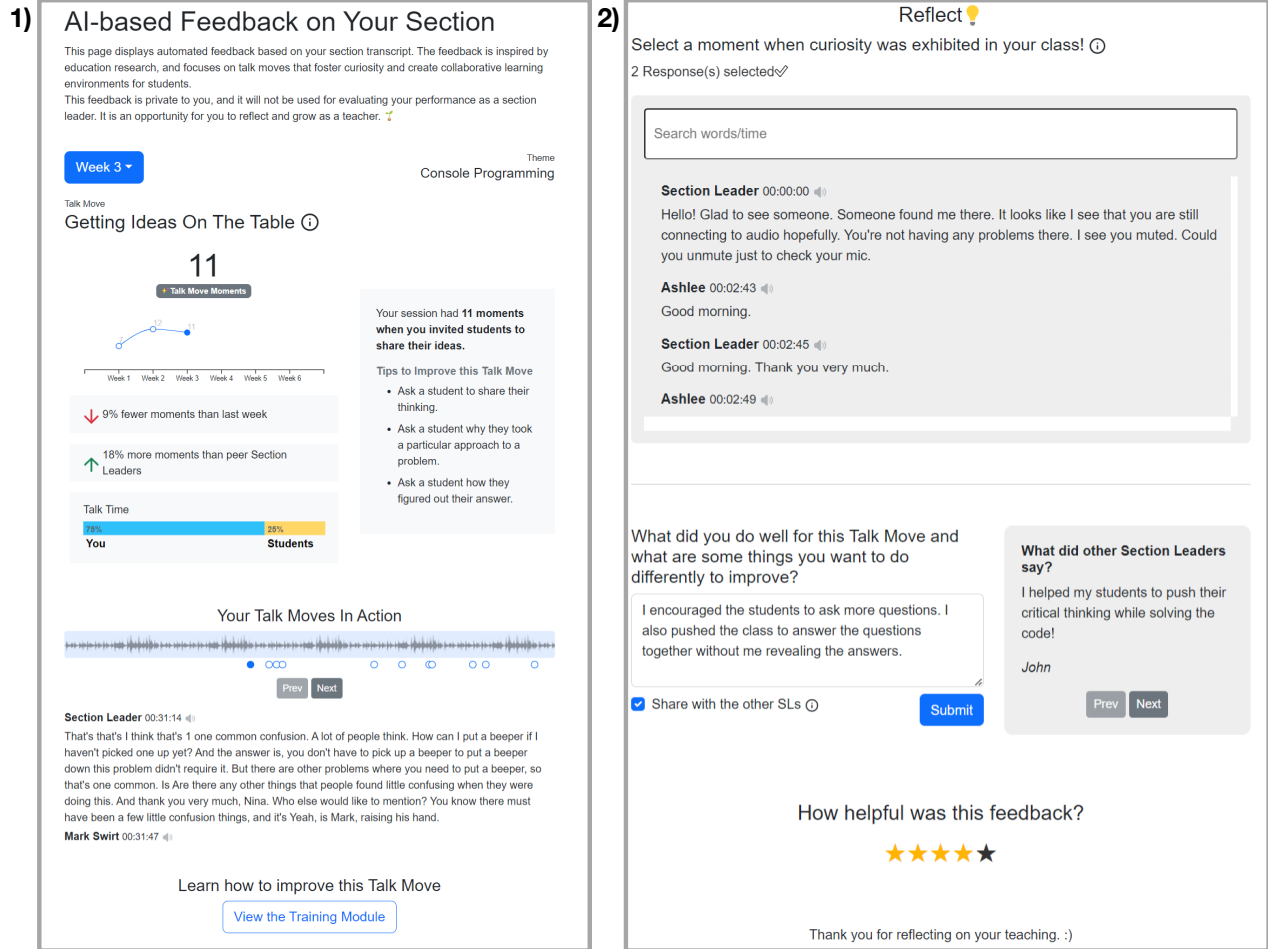


Figure 1: Screenshot of the feedback page, with a focus on the GETTING IDEAS ON THE TABLE talk move.

the same topic for 2 consecutive weeks. The topic assignment criteria by experimental condition are described in detail below.

The feedback included the following components (see Figure 1): an introduction to the feedback; the week and theme for the session; summary statistics for a given talk move; relative change in the frequency of that talk move compared to prior weeks; the talk ratio between the instructor and students; specific instances of the given talk move in the instructor's transcript; and a link to a relevant training module. The page additionally included the instructor's full transcript and a box they could use to search and identify moments for reflection, as well as a reflection question: "What did you do well for this Talk Move, and what are some things you want to do differently to improve?" Instructors could opt in to share their responses to the reflection question with other instructors and could view their reflections. Finally, instructors could rate the feedback on a scale of 1 to 5 stars.

4 Randomized Controlled Trial

We randomized instructors once they were accepted to teach in the course, before the course began. Half of the them got a choice of feedback topic. Instructors were asked to make this choice on the Code in Place website (Figure 2), as an action item on their precourse checklist. The choice involved feedback on three talk move topics (GETTING IDEAS ON THE TABLE, BUILDING ON IDEAS, ORIENTING STUDENTS TO ONE ANOTHER), which they could select for pairs of weeks (Weeks 1–2, 3–4, or 5–6⁴). Instructors also toggled between two options: They could enable experimental GPT-based feedback for the last 2 weeks, and they could enable seeing how their metrics compared with those of other instructors for all weeks. Below each talk move, we displayed a short definition and an example to help inform their choices. Before the course began, we sent instructors up to three email nudges to make a choice. About 80% of treatment group instructors made a choice of feedback topic.

The control group did not get to choose feedback topics. Instead these instructors were randomly assigned to feedback under the constraint that the distribution and sequence of feedback patterns in the control group was the same as the distribution in the treatment group. For example, 36% of the treatment group chose the following pattern: 2 weeks on GETTING IDEAS ON THE TABLE, 2 weeks on BUILDING ON IDEAS, 2 weeks of experimental feedback and comparison of metrics to instructors. Thus, 36% of the control group was assigned to that same pattern. As such, the only difference between treatment and control, in expectation, is whether the instructor chose their pattern of feedback or was assigned their pattern of feedback. The 20% of treatment group instructors who did not choose feedback were assigned feedback with the same weighted random assignment method as the control group, and remained part of the treatment group.

4.1 Emails About Feedback

At the end of each week, when all feedback was ready, we released it to all instructors at once, both by displaying it to them on the course platform and by emailing them that their feedback was ready. The email was short and did not contain any of the feedback itself—it merely included a link to the feedback page. However, the content did differ slightly based on condition. As illustrated in Figure 3, in order to reinforce the effect of the treatment, treatment group instructors were reminded in the email (both in the subject line and content) that they had *chosen* the focus of the feedback.





4.2 Measures of Outcomes





Following our preregistration, we measured four key types of outcomes to evaluate the impact of giving instructors choice over their feedback. We chose these measures to capture both immediate instructor responses to having agency (engagement, perception) and downstream effects on teaching practice and student outcomes.




4.2.1 Engagement With Automated Feedback

Instructor-level measures of engagement with automated feedback are based on their engagement with the feedback page.

⁴We had thought that the course would only be 5 weeks long; hence, the choice interface only had Week 5 listed for the third box. When we realized the course would be 6 weeks long, we applied their choices for Week 5 to Week 6 as well.

-  HOME
-  TRAINING
-  LOUNGE
-  SECTION

-  STUDENT
-  CODE
-  LEARN
-  FORUM

-  STORIES
-  EVENTS
-  ABOUT

Configure Talk Moves Feedback

AI-based Feedback on Your Section

We plan to provide automated feedback on the transcript of your section. This feedback is private to you, and it will not be used for evaluation of your performance as a section leader. The feedback is a reflection opportunity for you and we hope it will support your professional development.

We invite you to choose which aspect of your teaching you improve throughout Code in Place via AI-based feedback! The feedback is inspired by education research, and focuses on talk moves that foster curiosity and create collaborative learning environments for students.

Here are a few examples of each talk move:

1. Getting Ideas on the Table – What are students thinking? ^
 - How did you figure that out?
 - What information do you know? What are you trying to find out?
 - What have you tried so far? What happened?
 - What do you know definitely won't work? Why?
2. Building on Students' Ideas – What do they mean by that? or Oh that gives me an idea too! v
3. Orienting Students to One Another – What do other students think about that? v

Drag and drop your preference for each week

Feel free to choose a specific talk move more than once if it's something you really want to focus on.

<p>⋮</p> <p>Getting ideas on the table</p> <p>You'll get feedback on your questions that get students thinking and sharing with peers.</p>	<p>⋮</p> <p>Building on students' ideas</p> <p>You'll get feedback on how you give voice to and build on students' ideas.</p>	<p>⋮</p> <p>Orienting students to one another</p> <p>You'll get feedback on how you get students to listen to and build on each other's ideas.</p>
---	--	---

Week 1 & 2	Week 3 & 4	Week 5
Drag and Drop	Drag and Drop	Drag and Drop

Enable Experimental Feedback Feature for Week 5

We're experimenting with providing feedback using generative AI on other aspects of your instruction. We are still working on specifying this feedback, and it is not robustly tested like the feedback on talk moves; but if you would like to try it out and help us evaluate it, please feel free to enable this feature.

Would you like to compare?

Every Section Leader will receive personalized feedback describing their use of the selected talk moves. If you would also like to see a comparison of your use of these talk moves to other section leaders, please toggle this button to "on."

Submit

 1404 ONLINE

Figure 2: Screenshot of the Configure Talk Moves Feedback page.

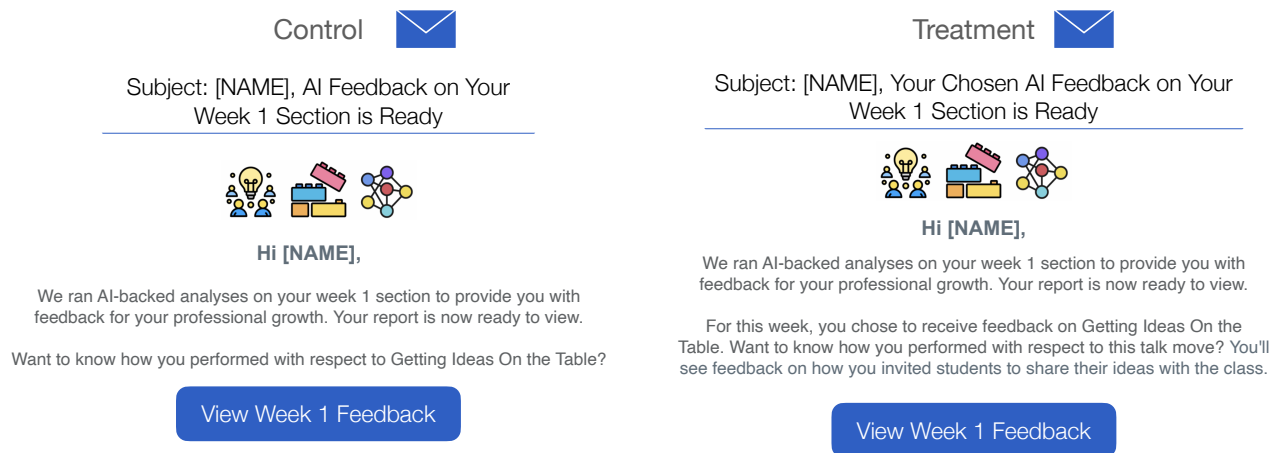


Figure 3: Example email received by control and treatment group instructors when the feedback was ready. The treatment group email emphasized that the feedback focus was chosen by the instructor.

- **Ever Viewed:** Whether instructors ever viewed their feedback *before their subsequent session* (binary). We also tracked the number of times they viewed the feedback, but the results were similar to Ever Viewed—hence, we use this binary measure.
- **Seconds Spent:** Total seconds spent viewing feedback across weeks.

4.2.2 Perception of Feedback

Instructor-level measures of perception of feedback are based on a post-course survey about the feedback they received. Survey questions are included in Appendix B. While the relatively low response rate of 33% limits the conclusions we can draw from these analyses, we focus on the following items:

- **Net Promoter Score (NPS):** 1–10 rating of likelihood of recommending the feedback tool.
- **Overall Perception:** Aggregated items from the final instructor survey measuring perceptions of feedback utility and satisfaction. As explained in the preregistration, a factor analysis showed a single dominant factor explaining most variance; hence, we mean-aggregated the items.

4.2.3 Changes in Instructional Practice

We measured changes along the three talk moves that the automated feedback was targeting. We calculated the hourly rate of each talk move by dividing the frequency of the talk move for a given session by the session duration in minutes, and then multiplying that number by 60. Finally, we standardized the talk move rates within each talk move (mean = 0, standard deviation = 1) to account for differences in talk move frequencies (e.g., GETTING IDEAS ON THE TABLE is about 8 times as frequent on average as ORIENTING STUDENTS TO ONE ANOTHER). This standardization enabled us to combine all talk moves into a single model as outcomes, as explained in the next section.

Since the treatment (i.e., making a feedback choice) was delivered *before* the course began, it could have impacted baseline, prefeedback instructional practices. Thus, we created two separate outcome variables at the instructor-week-talk move level:

- **Week 1 Talk Move Rate:** The standardized talk move rate(s) within the first session, across all talk moves. This measure captures discourse practices after treatment, but before instructors received any feedback.
- **Week 2+ Talk Move Rate:** The standardized talk move rate(s) within the second through sixth sessions, across all talk moves. This measure captures discourse practices after instructors received their first feedback. To improve precision, in models that use this outcome we controlled for talk move rates in the first session.

4.2.4 Student Outcomes

The course did not have any mandatory assignments. Attending sessions and completing assignments were the key indicators of student success in the course, and we used these student-level measures as outcome features.

- **Number of Sessions Attended:** Number of sessions attended by students between Week 2 and Week 6. We excluded attendance at the first session—while the first session was after random assignment, students did not interact with instructors until showing up (or not) for this session; thus, attendance at the first session could not have been affected by treatment.
- **Number of Assignments Completed:** The total number of assignments completed by summing completion rates across the six course assignments (usually one assignment per week).

4.3 Variables for Subgroup Analysis

We completed a subgroup analysis to determine whether instructor characteristics interacted with choice of topic to affect study outcomes. Based on our preregistration, we considered demographic characteristics of the instructor (gender, age, location, and whether they were returning as instructors in Code in Place) as well as behavioural characteristics. For behavioural characteristics, we considered whether the instructor engaged with two other forms of self-directed professional learning on the platform, both available *after* randomization but prior to the first session:

- **Training Modules About Talk Moves:** The four optional training modules included interactive videos and reflection questions related to each of the three talk moves (GETTING IDEAS ON THE TABLE, BUILDING ON IDEAS, ORIENTING STUDENTS TO ONE ANOTHER), as well as a module synthesizing all three. We used a binary measure indicating whether the instructor completed any of these modules. (Using the number of completed modules did not change our results.) Overall, 43% of instructors completed at least one training module.
- **GPTEach:** GPTEach (Markel, Opferman, Landay, & Piech, 2023) is an LLM-powered chat-based training tool that allowed instructors to practice engaging with simulated students. Created via GPT-3, the students had diverse backgrounds and familiarity with course material (programming), and the instructor was asked to facilitate office hours with these simulated students. We used a binary measure indicating whether the instructor accessed GPTEach. (Using the number of times they accessed GPTEach did not change our results.) Overall, 23% of instructors accessed GPTEach at least once during the course.

Since these professional learning tools were available to instructors postrandomization, the treatment (choice) may have influenced whether they engaged with the tools. However, using a t-test, we found no statistically

	Control Mean	Treatment Mean	p Value	N
Female	0.31	0.32	0.75	583
In United States	0.49	0.47	0.55	583
Age	30.61	29.72	0.37	583
Returning Instructor	0.25	0.25	0.98	583
Female	0.31	0.32	0.75	583
Number of Transcripts	5.75	5.74	0.81	583
Proportion of Female Students	0.52	0.52	0.681	567
Proportion of Students in United States	0.31	0.26	0.004	567
Mean Student Age	31.39	30.91	0.535	567

Table 2: A randomization check shows that instructor characteristics did not differ significantly by condition. The only exception was the proportion of students in the United States, which was significantly higher for section leaders in the control group. However, since we controlled for this variable in all analyses, this difference should not impact results.

significant difference by condition for completing any training module ($t = 0.022$, $p = 0.983$) or accessing GPTeach ($t = 0.092$, $p = 0.927$). This suggests no observable correlation between these moderator variables (engagement with professional learning tools) and our key independent variable (condition). However, these moderators could still have influenced the relationship between the condition and the dependent variables (outcomes) listed above.

4.4 Validating Randomization

To verify whether our randomization created groups that were balanced on observable variables, we evaluated whether the demographics of instructors in the treatment and control groups differed statistically. As Table 2 shows, we found no statistically significant differences in instructor demographics by condition. To examine whether the treatment and control conditions suffered from differential attrition, we also conducted an attrition analysis by calculating the number of transcripts available for each instructor. Attrition in our data occurred when instructors missed a session or a session was cancelled, leading to a missed transcript. We found no statistically significant differences in the number of recordings per instructor by condition. Finally, when comparing the section demographics (gender, age, and location of students), we found that instructors in the control group had a significantly higher proportion of students in the United States. However, since we controlled for this variable, this difference should not affect our results.

4.5 Regression Analysis

We conducted a preregistered intent-to-treat analysis using ordinary least squares regression. The analysis compared instructors by condition rather than by compliance (i.e., whether they made a choice over feedback), since the latter could introduce selection bias: Instructors who made a choice may have had different characteristics (e.g., more time or motivation for self-directed learning) than those who did not.

4.5.1 RQ1: Impact on Instruction Engagement with and Perception of Feedback

For instructor-level outcomes, we fit the following specification:

$$Y_i = \delta T_i + X_i \beta + \varepsilon_i \quad (1)$$

where Y_i refers to the outcome measure for instructor i ; T_i is a binary treatment indicator (1 if instructor i was assigned to choose their feedback type); X_i is a vector of instructor and student covariates; and ε_i indicates the residuals. The instructor covariates include age, gender (binary female indicator), location (binary U.S. indicator), and whether they were a returning Code in Place instructor. Student covariates, averaged at the section level, include age, gender composition, and location.

4.5.2 RQ1: Impact on Talk Moves

When using talk moves as outcomes, we pooled together all moves into one estimation sample. By pooling all talk moves together, we preserved balance in the experiment and avoided selection bias. Consider the alternative, where we would not pool but instead condition on feedback type (i.e., limiting the estimation sample to instructors who all received feedback on the same topic the same week). In that alternative, the treatment instructors would be a self-selected subset of the full treatment group, because they chose the feedback topic, but control instructors would be a randomly selected subset of the full control group. The treatment and control would no be longer balanced by random assignment, raising the potential for bias.

Before pooling, we standardized each talk move outcome (mean = 0, standard deviation = 1, using the full sample), so the outcome would be in standard deviation units. By pooling together, we expected to gain more power at the expense of some interpretability. Since each instructor had three talk move outcomes for each week, the specification becomes:

$$Y_{iwm} = \delta T_i + X_i \beta + \pi_w + \theta_m + \varepsilon_{iwm} \quad (2)$$

where Y_{iwm} refers to the standardized rate for instructor i in week w and for move m ; T_i is a binary treatment indicator; X_i is a vector of instructor and student covariates, same as above; π_w represents week fixed effects; θ_m represents talk move fixed effects (getting ideas, building on ideas, orienting students); and ε_{iwm} indicates the residuals. For models examining changes in instructional practice in Weeks 2–6, we additionally included the Week 1 (baseline) rate of each of the three talk moves as controls in X_i . The point estimates are similar if we omit the Week 1 controls (see Appendix C). We clustered standard errors at the instructor level.

4.5.3 RQ2: Impact on Student Outcomes

For student outcomes, the specification is:

$$Y_j = \delta T_{i(j)} + X_{i(j)} \beta + \varepsilon_j \quad (3)$$

where students are indexed by j , and $i(j)$ indicates the instructor of student j . In $X_{i(j)}$ we include the same instructor covariates as above as well as the demographics (age, gender, location) of the student j . We clustered standard errors at the instructor level.

	Engagement		Perception		Practice	
	(1) Ever Viewed	(2) Seconds Spent	(3) NPS	(4) Overall Perception	(5) Wk 1 Talk Move Rate (std)	(6) Wk 2+ Talk Move Rate (std)
Treatment	-0.019 (0.027)	19.713 (69.283)	0.018 (0.393)	0.084 (0.126)	0.084 (0.071)	-0.001 (0.047)
Control Mean	0.876	483.979	6.03	3.505	-0.014	-0.014
R2	0.065	0.088	0.131	0.114	0.029	0.032
Observations	567	567	193	193	1611	7686

Table 3: Standard errors are in parentheses. These models estimate the effect of choice over the feedback (treatment) on instructor outcomes. Models 1–5 are at the instructor level, and Model 6 is at the instructor-week level. All models include instructor covariates (age, is female, is returning, in United States) and student covariates averaged within section (mean age, proportion female, proportion in United States). Models 5 and 6 additionally include dummy fixed effects for whether the given talk move was eliciting, building, or orienting. Model 6 also includes controls for baseline discourse features rates for eliciting, building, and orienting, and dummy indicator variables for the week of the session, rating between 2 and 6.

Subgroup Analysis. To examine how feedback choice interacts with instructor characteristics or behaviour, we augmented the above models with an interaction term $\gamma(T_i * F_i)$, where F_i represents a given instructor feature listed in Section 4.3. The coefficient γ captures whether the treatment effect was larger when instructors had a certain characteristic or not.

5 Results

This section provides a summary of findings in response to each of our preregistered research questions.

5.1 RQ1: Does providing instructors choice over feedback impact their engagement with the feedback, their perception of the feedback, or their teaching practice?

We found that the treatment did not, on average, significantly impact instructors’ engagement with the feedback, their perception of the feedback, or their teaching practice. As shown in Table 3, across all outcome measures, the effects were statistically insignificant and relatively small in magnitude.

However, examining patterns over time reveals interesting trends. As shown in Figure 4, treatment group instructors tended to use more talk moves than control group instructors, particularly in the first 3 weeks of the course. This pattern was most pronounced for the ORIENTING STUDENTS TO ONE ANOTHER MOVE, where treatment instructors maintained consistently higher rates throughout the course. While these differences did not reach statistical significance, they suggest that having choice over feedback may have had a positive impact on some aspects of instructional practice.

5.2 RQ2: Does choice over feedback for instructors impact their students’ outcomes?

Next, we examined whether the treatment impacted student outcomes, including attendance and assignment completions. Unexpectedly, we found a small but statistically significant positive effect on student attendance.

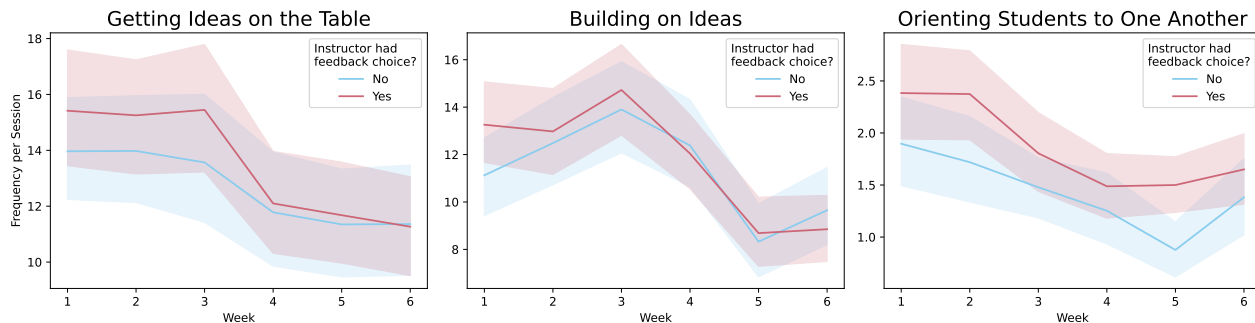


Figure 4: Talk move rates over time split by condition. Shaded bands represent 95% confidence intervals.

	(1)	(2)
	Num. Sessions	Num. Assignments
	Attended	Completed
Treatment	0.112*	0.050
	(0.048)	(0.052)
Control Mean	3.575	3.434
R2	0.042	0.013
Observations	8254	8254

Table 4: Standard errors are in parentheses. These models estimate the effect of choice over the feedback (treatment) on student outcomes, at the student level. Both models include instructor covariates (age, is female, is returning, in United States) and student covariates (age, is female, in United States). Standard errors are clustered at the instructor level. * $p < 0.05$

As shown in Table 4, students whose instructors had choice over feedback attended on average 0.112 more sessions compared to the control group ($SE = 0.048$, $p < 0.05$). This represents a meaningful increase of about 3.1% over the control mean of 3.575 sessions.

Given that students were unaware of the intervention, this effect was likely mediated through changes in instructor behaviour. While we did not detect significant changes in our measured instructional practices (Table 3), the treatment may have influenced other aspects of instruction that impacted student engagement. For assignment completions, we found no significant treatment effects—the coefficients were consistently positive but small for each assignment (ranging from 0.001 to 0.015, all $p > 0.05$).

5.3 RQ3: How do treatment effects vary by instructor demographics and whether the instructor engages with self-directed professional learning beyond automated feedback?

Finally, we turn to examining subgroup treatment effects. We did not find any notable variation in treatment effects based on instructor gender, U.S. location, or their returning instructor status (Appendix Table D). We did find, however, interesting subgroup differences based on engagement with self-directed professional learning tools. Figure 5 illustrates that students of instructors who engaged with professional learning tools, including training modules and GPTeach, had the best outcomes. As a reminder, these tools were available to instructors

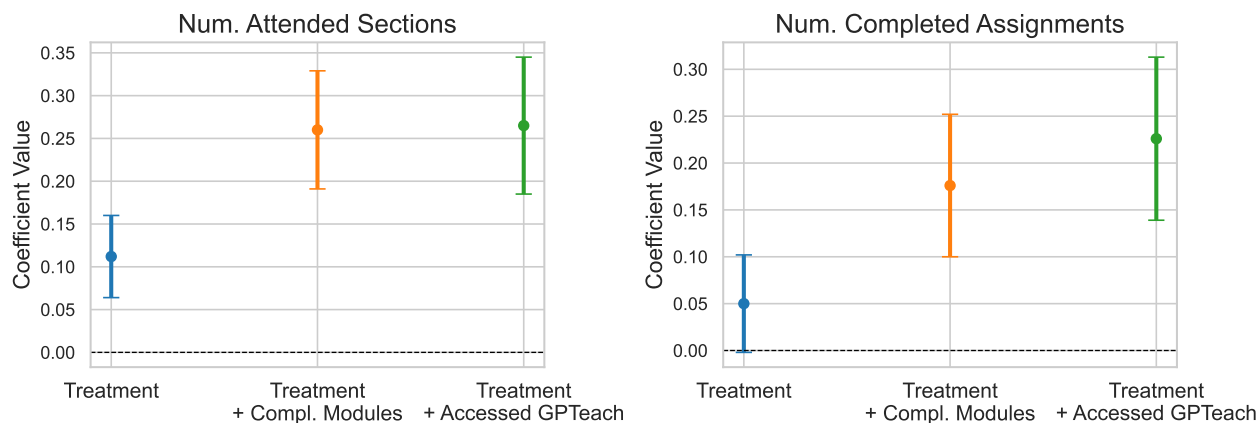


Figure 5: Treatment effects on the number of attended sessions and number of completed assignments. The lines represent the entire treatment group (values are same as in Table 4), the subset of the treatment group that also completed a training module (values are same as in Tables 5 Columns 7–8, Row (c)) and another subset of the treatment group that accessed GPTEach (values are same as in Tables 6 Columns 7–8, Row (c)), respectively. The bands represent standard errors.

postrandomization, but instructors’ likelihood to engage with them did not significantly differ by condition (Section 4.3).

As shown in the subgroup comparison in Tables 5 and 6, instructors who completed training modules used more talk moves (+0.198, $p < 0.05$ in Week 1; +0.135, $p < 0.10$ in Weeks 2+), and their students attended 0.260 more sessions ($p < 0.01$) and completed 0.176 more assignments ($p < 0.05$) than students of control group instructors who did not complete the modules. Similarly, as shown in Table 6, students whose instructors were in the treatment group and accessed GPTEach attended 0.265 more sessions ($p < 0.01$) and completed 0.226 more assignments ($p < 0.01$) than students of control group instructors who did not access GPTEach. These effects are about 2–4 times greater than those reported in Table 4 for all treatment group instructors.

One might ask: Could the increased coefficients in Figure 5 be associated with characteristics of instructors who use *professional learning tools*, rather than the impact of choice? Our results suggest otherwise. We find that providing instructors with choice had benefits to students above and beyond the tools: Students of treated instructors who completed modules attended more sessions (+0.147, $p < 0.05$) and completed slightly more assignments (+0.118, $p = 0.131$) than students of control group instructors who completed modules. Analogically, students of treated instructors who accessed GPTEach attended more sessions (+0.216, $p < 0.05$) than students of control group instructors who accessed GPTEach. Altogether, these results indicate that both providing choice over feedback to instructors *and* instructors’ engagement with professional learning tools had greater benefits for students than choice or engagement with the tools alone.

However, we did not observe the same added benefit of the treatment for instructors who used professional learning tools for other outcomes. Interestingly, it seems that treated instructors who engaged with these tools spent less time on the feedback than control group instructors who engaged with the tools. This trend is negative for all four outcomes (Columns 1 and 2 in Tables 5 and 6), but it is only significant for one of the values: Among instructors who accessed GPTEach, those in the treatment group were about 8% less likely to view the feedback than those in the control group ($p < 0.05$).

	Engagement		Perception		Practice		Students	
	(1) Ever Viewed	(2) Seconds Spent	(3) NPS	(4) Mean of Items	(5) Wk 1 Talk Move Rate	(6) Wk 2+ Talk Move Rate	(7) Num. Sessions Attended	(8) Num. Assn. Completed
(a) Treatment=0#Compl. Module=1	0.078* (0.037)	390.388** (94.403)	1.322* (0.602)	0.401* (0.191)	0.158 (0.105)	0.125+ (0.069)	0.113 (0.070)	0.058 (0.076)
(b) Treatment=1#Compl. Module=0	-0.007 (0.040)	143.350* (66.800)	0.740 (0.733)	0.261 (0.224)	0.114 (0.092)	0.018 (0.060)	0.081 (0.064)	-0.005 (0.071)
(c) Treatment=1#Compl. Module=1	0.041 (0.037)	240.703** (90.675)	0.796 (0.593)	0.347+ (0.193)	0.198* (0.099)	0.135+ (0.070)	0.260** (0.069)	0.176* (0.076)
(c) - (a)	-0.037	-149.685	-0.526	-0.054	0.040	0.010	0.147*	0.118
(c) - (b)	0.048	97.353	0.056	0.086	0.084	0.117	0.179**	0.181*
Control Mean	0.827	273.056	5.410	3.311	-0.070	-0.070	3.468	3.378
R2	0.074	0.113	0.155	0.138	0.033	0.026	0.043	0.013
Observations	567	567	193	193	1611	7992	8254	8254

Table 5: Treatment effects based on whether instructors completed any of the four optional training modules (moderator). Standard errors are in parentheses. Estimates are compared to the control group where the moderator variable is 0. The model specifications are the same as in Tables 3 and 4, with an additional interaction term added as specified in Section 4.5. Above we report the estimated total effect for each of the three groups relative to the control group, which did not complete any modules. The difference between point estimates was calculated with a robust Wald test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	Engagement		Perception		Practice		Students	
	(1) Ever Viewed	(2) Seconds Spent	(3) NPS	(4) Mean of Items	(5) Wk 1 Talk Move Rate	(6) Wk 2+ Talk Move Rate	(7) Num. Sessions Attended	(8) Num. Assn. Completed
(a) Treatment=0#Used GPTeach=1	0.142** (0.026)	380.527** (116.504)	0.158 (0.580)	0.036 (0.184)	0.069 (0.117)	0.222** (0.075)	0.049 (0.084)	0.091 (0.091)
(b) Treatment=1#Used GPTeach=0	0.000 (0.033)	34.228 (72.142)	-0.000 (0.484)	0.069 (0.153)	0.114 (0.083)	0.069 (0.052)	0.080 (0.054)	0.025 (0.060)
(c) Treatment=1#Used GPTeach=1	0.060 (0.039)	353.411* (147.144)	0.226 (0.617)	0.154 (0.186)	0.057 (0.116)	0.067 (0.090)	0.265** (0.080)	0.226** (0.087)
(c)-(a)	-0.082*	-27.116	0.068	0.118	-0.012	-0.155	0.216*	0.135
(c)-(b)	0.06	319.183*	0.226	0.085	-0.057	-0.002	0.185*	0.201*
Control Mean	0.853	392.078	5.918	3.470	-0.050	-0.050	3.568	3.413
R2	0.084	0.116	0.132	0.116	0.030	0.027	0.043	0.013
Observations	567	567	193	193	1611	7992	8254	8254

Table 6: Treatment effects based on whether instructors engaged with GPTeach (moderator). Standard errors are in parentheses. Estimates are compared to the control group where the moderator variable is 0. The model specifications are the same as in Tables 3 and 4, with an additional interaction term added as specified in Section 4.5. Above we report the estimated total effect for each of the three groups relative to the control group, which did not access GPTeach. The difference between point estimates was calculated with a robust Wald test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This decrease in engagement may have been driven by treated instructors who were offered a choice but decided *not* to make one: This subgroup of the treatment group ($n = 44$) was 26% less likely than the control group to check the feedback ($p < 0.001$). When we removed this subgroup, there was no longer a significant difference by condition in engagement with feedback—either overall, or for the subgroup who used GPTEach. This suggests that offering feedback choice to instructors who do not desire to make choices may harm their engagement with feedback.

6 Discussion

6.1 Theoretical Implications

Our study is among the first to offer experimental evidence on the role of agency in teacher feedback and professional learning. Prior work has relied on theory and qualitative methods, such as self-reports, interviews, or case studies, to suggest that granting teachers more autonomy over their learning fosters professional growth (Molla & Nolan, 2020; Priestley et al., 2015; Brodie, 2021; Martin et al., 2019; Philpott & Oates, 2017; Brod, Kucirkova, Shepherd, Jolles, & Molenaar, 2023). These studies have illustrated that when teachers perceive themselves as active agents in their learning, they are more likely to integrate new strategies into their practice. However, such studies could not disentangle the impact of agency on instruction from other factors, such as teachers' intrinsic motivation or teacher growth over time. To fill this gap, we conducted a randomized controlled trial to test whether providing volunteer instructors with choice over their feedback topics improves engagement with feedback, instructional practice, and student outcomes. Random assignment helped isolate the effect of agency from other factors.

Our results paint a nuanced picture. While we observed a positive trend in treated teachers' use of high-leverage talk moves, choice over feedback alone did not lead to significant changes in observed instructor behaviour. This finding aligns with prior work suggesting that while choice can be empowering, it does not guarantee meaningful engagement or skill acquisition (Lynch et al., 2019). At the same time, we found a significant and unexpected effect of agency on student attendance. Students taught by instructors who received choice over their feedback attended more sessions on average than those in the control group. This suggests that the treatment may have induced unobserved changes in instructor behaviour that influenced student engagement.

The impact on student attendance was greater for students whose instructors engaged with additional professional learning resources, including training modules and teaching simulations. Relatedly, metrics of instructional practice and student outcomes were the best among the subgroup of instructors who both received choice and engaged in additional professional learning. At the same time, we also have evidence to suggest that offering choice to instructors who are not interested in making choices may be detrimental to their engagement with professional learning. Thus, agency appears to be most beneficial when instructors are intrinsically motivated to engage in further learning. These findings align with theories of adult learning that suggest that agency is most impactful when learners have both the autonomy to make choices and the internal motivation to act upon them (Deci & Ryan, 2013; O'Brien & Reale, 2021).

6.2 Practical Implications

The lack of a widespread benefit of choice to instructors raises practical questions about the optimal integration of agency into professional learning programs. A high-agency model may be most efficient in contexts where participants are genuinely motivated and have the capacity to self-direct (O'Brien & Reale, 2021). On one hand, in elective professional learning—where instructors already have a clear interest in professional

growth—increased autonomy can enhance motivation and ownership. For instance, by giving instructors choice over feedback, programs can better align professional learning content with immediate needs, which can improve both buy-in and teaching practice (Molla & Nolan, 2020; O'Brien & Reale, 2021). On the other hand, in mandated or large-scale K–12 professional learning contexts, educators are often time-constrained or less intrinsically motivated to experiment with new practices. As a result, providing extensive choice might not pay off in terms of engagement or outcomes (Lynch et al., 2019), especially when it creates logistical complexity for administrators.

In order to keep the intervention "light-touch" and minimize the time volunteer instructors would need to make a choice, we provided minimal scaffolding for choices. However, additional scaffolding could increase instructors' motivation and deliberative engagement with choices. For example, programs where instructors have more time to engage in professional learning could include core pathways of "must-do" content for essential skills, then offer elective modules for instructors who wish to delve deeper into specific areas. Such programs could also embed self-assessment opportunities that prompt teachers to set goals or select topics based on their perceived strengths and weaknesses. These options would help maintain a coherent structure within a professional learning program while fostering a "pedagogy of choice" (O'Brien & Reale, 2021). In programs like Code in Place, where most instructors have minimal capacity to engage in professional learning, gamification (e.g., with rewards) could motivate them to be intentional about improving their teaching. A hybrid professional learning system could support educators with varying capacity as well as differentiate the degree of autonomy based on instructors' readiness or willingness to engage in self-directed learning (Brod et al., 2023).

6.3 Limitations

Although our study yields valuable insights, questions remain about the generalizability of the findings. First, our study was conducted in a unique context—an online programming course with volunteer instructors who had diverse backgrounds, teaching experiences, motivations, and constraints on their time. Some may have been particularly motivated to improve their teaching and engage with optional resources, while others may not have had the interest or time to do so. The effects of agency might differ in traditional K–12 or higher education settings with trained educators who have institutional structures and incentives to participate in professional learning (Mohammad Nezhad & Stolz, 2024).

Second, the automated measures of teaching behaviours we employed, while convenient and scalable, are imperfect (Kupor et al., 2023) and cannot capture the full complexity of instructional change. Prior work on teacher agency highlights how changes in practice can be subtle, multifaceted, and best captured through multimethod triangulation (Biesta et al., 2015). The fact that we observed positive student attendance effects despite modest changes in measured instructional practices suggests there may be important, unmeasured dimensions of teacher–student interactions at play.

Third, although we hypothesized that teachers' decisions to explore optional professional learning resources were driven by their motivation, it is also possible that they were driven by other factors, such as available time or self-efficacy. To better isolate the role of different factors, future research might incorporate direct measures of teacher motivation, self-regulatory capacity, and sense of agency based on existing frameworks (Brod et al., 2023).

Finally, our intervention was relatively minimal—offering a one-time choice over feedback before the course began. It would be valuable to investigate the impact of providing more extensive and ongoing options for

instructors to customize feedback. Relatedly, the short 6-week time frame of the course may have limited opportunities for deeper, more perceptible shifts in instruction. It may therefore be valuable to test the impact of agency in a longer-term program.

6.4 Future Directions

There are several avenues to build on this work. First, replication in both voluntary and mandatory professional learning settings would clarify how contextual factors affect the impact of agency. Randomized studies that manipulate the *degree* of choice could illuminate the optimal balance between structure and autonomy in different contexts (Brod et al., 2023; O'Brien & Reale, 2021).

Second, drawing on learning analytics, one might design a professional learning system to test the effectiveness of adapting the degree of agency in real time (Chen, Mitrovic, & Mathews, 2019). Such a system could dynamically adjust learning pathways—balancing freedom with structured supports—by monitoring metrics such as time spent on feedback, use of talk moves, or self-assessment responses. This approach would also advance the broader goal of designing personalized professional learning that efficiently responds to teacher needs while maintaining quality and rigor.

Third, longer-term studies can help us understand how agency-driven improvements evolve over time. While this study captured short-term impacts of a light-touch intervention, follow-up work could explore how continued agency over a longer period impacts teaching practice and student outcomes (O'Brien & Reale, 2021).

Finally, we need to better understand the mechanisms linking teacher agency to student outcomes. Investigating *how* and *why* teacher autonomy influences attendance, assignment completion, or other measures of student behaviour will further clarify when and for whom agency is most beneficial. Altogether, these insights would fuel evidence-based and adaptive approaches to designing the future of teacher professional learning.

Acknowledgments

We appreciate John Papay's advice on experimental design. We would like to thank the Code in Place team for supporting us with the implementation of the experiment. Thanks to Jenny Osuna, Jim Malamut and Miroslav Suzara for brainstorming sessions related to the talk moves and for sharing information about the training modules. Thank you to the Carina Foundation for funding the Code in Place project.

Disclosure Statement

The authors report there are no competing interests to declare.

References

- Anderson, L. (2010). Embedded, emboldened, and (net) working for change: Support-seeking and teacher agency in urban, high-needs schools. *Harvard Educational Review*, 80(4), 541–573.
- Biesta, G., Priestley, M., & Robinson, S. (2015). The role of beliefs in teacher agency. *Teachers and teaching*, 21(6), 624–640.

- Brod, G., Kucirkova, N., Shepherd, J., Jolles, D., & Molenaar, I. (2023). Agency in educational technology: Interdisciplinary perspectives and implications for learning design. *Educational Psychology Review*, 35(1), 25.
- Brodie, K. (2021). Teacher agency in professional learning communities. *Professional development in education*, 47(4), 560–573.
- Calvert, L. (2016). The power of teacher agency. *The Learning Professional*, 37(2), 51.
- Carter Andrews, D. J., & Richmond, G. (2019). *Professional development for equity: What constitutes powerful professional learning?* (Vol. 70) (No. 5). SAGE Publications Sage CA: Los Angeles, CA.
- Chen, X., Mitrovic, A., & Mathews, M. (2019). Investigating the effect of agency on learning from worked examples, erroneous examples and problem solving. *International Journal of Artificial Intelligence in Education*, 29(3), 396–424.
- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and teacher education*, 18(8), 947–967.
- Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the united states and abroad*. National Staff Development Council.
- Deci, E. L., & Ryan, R. M. (2013). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.
- Demszky, D., & Liu, J. (2023). M-powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes. *L@S '23: Proceedings of the Tenth ACM Conference on Learning @ Scale*.
- Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*.
- Demszky, D., Liu, J., Hill, H. C., Sanghi, S., & Chung, A. (2024). Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education*, 105183.
- Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 1638–1653.
- Demszky, D., Wang, R., Geraghty, S., & Yu, C. (2024). Does feedback on talk time increase student engagement? evidence from a randomized controlled trial on a math tutoring platform. , 632–644.
- Diaz-Maggioli, G. (2004). *Teacher-centered professional development*. ASCD.
- Doan, S., Fernandez, M.-P., Grant, D., Kaufman, J. H., Setodji, C. M., Snoke, J., . . . Young, C. J. (2021). American instructional resources surveys: 2021 technical documentation and survey results. research report. rr-a134-10. *Rand Corporation*.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hill, H. C. (2009). Fixing teacher professional development. *Phi Delta Kappan*, 90(7), 470–476.
- Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112, 103631.
- Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13).

- Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D’Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7), 451–464.
- Kennedy, M. (2016). Parsing the practice of teaching. *Journal of teacher education*, 67(1), 6–17.
- Kennedy, M. M. (2016). How does professional development improve teaching? *Review of educational research*, 86(4), 945–980.
- Knowles, M. S. (1984). The adult learner: A neglected species.
- Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4), 547–588.
- Kupor, A., Morgan, C., & Demszky, D. (2023). Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint arXiv:2311.10749*.
- Lieberman, A., & Pointer Mace, D. H. (2008). Teacher learning: The key to educational reform. *Journal of teacher education*, 59(3), 226–234.
- Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.
- Lloyd, M., & Davis, J. P. (2018). Beyond performativity: A pragmatic model of teacher professional learning. *Professional development in education*, 44(1), 92–106.
- Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs stem instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, 41(3), 260–293.
- Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). Gpteach: Interactive ta training with gpt based students. *L@S '23: Proceedings of the Tenth ACM Conference on Learning @ Scale*.
- Martin, L. E., Kragler, S., Quatroche, D., & Bauserman, K. (2019). Transforming schools: The power of teachers’ input in professional development. *Journal of Educational Research and Practice*, 9(1), 179–188.
- Merriam, S. B., et al. (2001). Andragogy and self-directed learning: Pillars of adult learning theory. *New directions for adult and continuing education*, 2001(89), 3.
- Mohammad Nezhad, P., & Stolz, S. A. (2024). Unveiling teachers’ professional agency and decision-making in professional learning: the illusion of choice. *Professional Development in Education*, 1–21.
- Molla, T., & Nolan, A. (2020). Teacher agency and professional practice. *Teachers and Teaching*, 26(1), 67–87.
- Morales, M. P. E. (2016). Participatory action research (par) cum action research (ar) in teacher professional development: a literature review. *International Journal of Research in Education and Science*, 2(1), 156–165.
- O’Connor, C., Michaels, S., & Chapin, S. (2015, 04). "scaling down" to explore the role of talk in learning: From district intervention to controlled classroom study. In (p. 111-126). doi: 10.3102/978-0-935302-43-1_9
- O’Brien, E., & Reale, J. (2021). Supporting learner agency using the pedagogy of choice. *Unleashing the power of learner agency*, 73–82.
- Philpott, C., & Oates, C. (2017). Teacher agency and professional learning communities; what can learning rounds in scotland teach us? *Professional development in education*, 43(3), 318–333.
- Priestley, M., Biesta, G., Philippou, S., & Robinson, S. (2015). The teacher and the curriculum: Exploring teacher agency. *The SAGE handbook of curriculum, pedagogy and assessment*, 187–201.
- Schön, D. A. (1983). The reflective practitioner: How professionals think in action.
- Smith, K. (2017). *Teachers as self-directed learners*. Springer.

- Stanley, A. M. (2011). Professional development within collaborative teacher study groups: Pitfalls and promises. *Arts Education Policy Review*, 112(2), 71–78.
- Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change*, 7(4), 221–258.
- Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., & Sumner, T. (2021a). Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. *arXiv preprint arXiv:2105.07949*.
- Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., & Sumner, T. (2021b). Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application..
- Vähäsantanen, K., Hökkä, P., Paloniemi, S., Herranen, S., & Eteläpelto, A. (2017). Professional learning and agency in an identity coaching programme. *Professional Development in Education*, 43(4), 514–536.
- Wang, Z., Miller, K., & Cortina, K. (2013). Using the lena in teacher training: Promoting student involvement through automated feedback. *Unterrichtswissenschaft*, 41(4), 290–302.
- Zeichner, K. (2019). The importance of teacher agency and expertise in education reform and policymaking. *Revista Portuguesa de Educação*, 32(1), 5–15.
- Zuo, G., Doan, S., & Kaufman, J. H. (2023). How do teachers spend professional learning time, and does it connect to classroom practice? findings from the 2022 american instructional resources survey. american educator panels. research report. rr-a134-18. *RAND Corporation*.

A Experimental Feedback

For Weeks 5–6, instructors had the option to choose "experimental feedback," which was powered by GPT-3.5. We first split up the transcript into smaller sections due to the character limit for the model. For each section, we asked GPT to "Please summarize what happened in the following segment of the classroom session." Then, we took each summary and asked GPT to summarize them all together with "Please summarize the following summaries of classroom sessions." Additionally, we asked GPT to find moments of curiosity with the following prompt: "Please identify specific moments with the timestamp in the following transcript where the section leader or the students exhibited curiosity. For each moment, please tell me your reasoning for what they did that exhibited curiosity."

B Final Survey on AI Feedback

The final survey, administered to all instructors, included the questions below. We used two items from this survey as outcomes. For "Mean of items," we aggregated responses to items within Question 4 by converting the answer choices to a 5-point Likert scale, reversing the scale for Question 4c and taking the mean of the numeric responses. For "NPS," we used responses to Question 5.

1. How often did you engage with the AI teaching feedback?
 - (a) Not at all.
 - (b) Once or twice.
 - (c) Regularly.
2. Could you tell us why you didn't engage with the AI teaching feedback? Select all that apply.
 - (a) I didn't know about it.
 - (b) It wasn't available to me.
 - (c) I didn't have the time.
 - (d) I didn't think it would be helpful.
 - (e) Other (please explain).
3. Could you tell us why you engaged with the AI teaching feedback only once or twice? Select all that apply.
 - (a) I only learned about it later in the course.
 - (b) It wasn't available to me after each session.
 - (c) I didn't have the time.
 - (d) I didn't find it helpful.
 - (e) Other (please explain).
4. To what extent do you agree with the following about the AI teaching feedback? (Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree)
 - (a) The feedback has helped me become a better teacher.
 - (b) The feedback made me realize things about my teaching that I otherwise would not have.
 - (c) The feedback was difficult to understand.
 - (d) The feedback made me pay more attention to the teaching strategies I was using.

- (e) I tried new things in my teaching because of this feedback.
 - (f) The feedback areas (e.g., getting ideas on the table, building on student ideas, orienting students to one another) represented important aspects of good teaching.
 - (g) The feedback allowed me to improve my teaching around areas that were important to me.
 - (h) The feedback felt appropriate to my teaching strengths and weaknesses.
 - (i) The feedback aligned with my priorities for growth in my teaching.
5. How likely are you to recommend AI teaching feedback to other educators? (Scale of 1–10)
 6. How helpful was each of the following types of feedback?
 - (a) Getting Ideas on the Table.
 - (b) Building on Student Ideas.
 - (c) Orienting Students to One Another.
 - (d) Experimental (ChatGPT) Feedback.
 7. Please rank the different elements of feedback in terms of helpfulness.
 - (a) Number of talk move moments identified.
 - (b) Chart to compare the number of moments to previous weeks.
 - (c) Comparison of the number of moments to class average.
 - (d) Talk time percentage.
 - (e) Tips to improve the talk move.
 - (f) Examples from your transcript demonstrating the talk move.
 - (g) Selecting moments when curiosity was exhibited.
 - (h) Answering the reflection question.
 - (i) Seeing other section leaders' answers to the reflection question.
 - (j) Resources to improve the talk move.
 - (k) Other (please explain).
 8. Do you have any suggestions for how we could improve this feedback tool?
 9. Any other thoughts/comments?

C Table 3 Column 6 With No Week 1 Controls

	Week 2+ Talk Move Rate
Treatment	0.017 (0.047)
Control Mean	-0.014
R2	0.023
Observations	7992

Table 7: Table 3 Column (6) with no Week 1 controls.

D Heterogeneity by Instructor Demographics

	Engagement		Perception		Practice		Students	
	(1) Ever Viewed	(2) Seconds Spent	(3) NPS	(4) Mean of Items	(5) Wk 1 Talk Move Rate	(6) Wk 2+ Talk Move Rate	(7) Num. Sessions Attended	(8) Num. Assn. Completed
(a) Treatment=0#Female Instr.=1	-0.001 (0.039)	79.820 (95.513)	0.392 (0.566)	0.232 (0.189)	0.040 (0.114)	0.012 (0.074)	0.092 (0.074)	0.010 (0.081)
(b) Treatment=1#Female Instr.=0	0.004 (0.031)	52.090 (86.134)	0.140 (0.489)	0.172 (0.151)	0.104 (0.086)	0.043 (0.057)	0.147* (0.058)	0.059 (0.063)
(c) Treatment=1#Female Instr.=1	-0.069 (0.046)	28.036 (100.069)	0.136 (0.586)	0.115 (0.198)	0.079 (0.105)	-0.029 (0.072)	0.127+ (0.074)	0.041 (0.082)
(c)-(a)	-0.068	-51.784	-0.256	-0.117	0.039	-0.041	0.035	0.031
(c)-(b)	-0.073	-24.054	-0.004	-0.057	-0.025	-0.072	-0.02	-0.018
Control Mean	0.882	462.026	5.903	3.432	-0.014	-0.014	3.560	3.435
R2	0.068	0.089	0.132	0.120	0.029	0.023	0.043	0.013
Observations	567	567	193	193	1611	7992	8254	8254

Table 8: Treatment effects based on whether instructor is female. Standard errors in parentheses. Estimates are compared to the control group where the moderator variable is zero. The model specifications are the same as in Tables 3 and 4, with an additional interaction term added as specified in Section 4.5. Above we report the estimated total effect for each of the three groups, relative to nonfemales in the control group.

	Engagement		Perception		Practice		Students	
	(1) Ever Viewed	(2) Seconds Spent	(3) NPS	(4) Mean of Items	(5) Wk 1 Talk Move Rate	(6) Wk 2+ Talk Move Rate	(7) Num. Sessions Attended	(8) Num. Assn. Completed
(a) Treatment=0 # Returning Instr.=1	-0.067 (0.050)	-319.437** (116.275)	-0.792 (0.718)	-0.239 (0.195)	-0.011 (0.112)	-0.069 (0.073)	0.047 (0.080)	-0.022 (0.087)
(b) Treatment=1 # Returning Instr.=0	-0.049 (0.030)	23.512 (89.590)	-0.047 (0.444)	0.086 (0.148)	0.099 (0.082)	-0.005 (0.056)	0.111* (0.056)	0.064 (0.061)
(c) Treatment=1 # Returning Instr.=1	0.003 (0.040)	-310.959** (91.156)	-0.542 (0.610)	-0.162 (0.191)	0.031 (0.120)	0.014 (0.073)	0.159* (0.077)	-0.001 (0.085)
(c)-(a)	0.07	8.478	0.25	0.077	0.042	0.083	0.112	0.021
(c)-(b)	0.052	-334.471***	-0.495	-0.248	-0.068	0.019	0.048	-0.065
Control Mean	0.896	528.682	6.405	3.616	-0.003	-0.003	3.486	3.391
R2	0.071	0.088	0.131	0.114	0.029	0.023	0.042	0.012
Observations	567	567	193	193	1611	7992	8254	8254

Table 9: Treatment effects based on whether instructor is a returning instructor in Code in Place. Standard errors in parentheses. Estimates are compared to the control group where the moderator variable is zero. The model specifications are the same as in Tables 3 and 4, with an additional interaction term added as specified in Section 4.5. Above we report the estimated total effect for each of the three groups, relative to instructors in the control group who were not returning instructors.

	Engagement		Perception		Practice		Students	
	(1) Ever Viewed	(2) Seconds Spent	(3) NPS	(4) Mean of Items	(5) Wk 1 Talk Move Rate	(6) Wk 2+ Talk Move Rate	(7) Num. Sessions Attended	(8) Num. Assn. Completed
(a) Treatment=0#Instr. in US=1	-0.101* (0.041)	-95.381 (82.656)	-1.444* (0.650)	-0.363+ (0.212)	0.222* (0.110)	0.114 (0.070)	-0.024 (0.070)	-0.095 (0.077)
(b) Treatment=1#Instr. in US=0	0.007 (0.029)	171.193+ (90.395)	0.009 (0.460)	0.094 (0.153)	0.010 (0.092)	-0.006 (0.067)	0.064 (0.066)	0.035 (0.072)
(c) Treatment=1#Instr. in U.S.=1	-0.148** (0.043)	-239.925** (83.171)	-1.408* (0.692)	-0.298 (0.223)	0.390** (0.108)	0.156* (0.069)	0.143* (0.070)	-0.026 (0.076)
(c)-(a)	-0.047	-144.544	0.036	0.065	0.168	0.042	0.167*	0.069
(c)-(b)	-0.155***	-411.118***	-1.417*	-0.392+	0.38**	0.162*	0.079	-0.061
Control Mean	0.917	469.458	6.710	3.683	-0.087	-0.087	3.471	3.417
R2	0.067	0.096	0.131	0.115	0.031	0.023	0.043	0.013
Observations	567	567	193	193	1611	7992	8254	8254

Table 10: Treatment effects based on whether instructor is in the United States. Standard errors in parentheses. Estimates are compared to the control group where the moderator variable is zero. The model specifications are the same as in Tables 3 and 4, with an additional interaction term added as specified in Section 4.5. Above we report the estimated total effect for each of the three groups, relative to instructors in the control group who were not in the United States. The difference between point estimates are calculated with a robust Wald test. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$