rppl | Annenberg Institute BROWN UNIVERSITY

July 2026

Designing Effective Curriculum-Based Professional Learning

Insights from a Cross-Study Portfolio

Technical Appendix

This is the technical appendix to RPPL's Designing Curriculum-Based Professional Learning.
Click here to read the full report.

Appendix

This appendix is organized by study. Use the links below to navigate to the corresponding study appendix.

Anticipating Common Misconceptions: Enhancing Equitable Math Instruction Through Guided Adaptation of HQIM

Achievement Network (ANet)

Brandi Phillips | Director Math Content Learning, Achievement Network
Patrick Poehailos | Senior Analyst, Achievement Network
Nadia Salman | Director of Development, Achievement Network

Exploring the Impact of Thinking Routines and Professional Learning Delivery Models on Teacher and Student Outcomes in Mathematics Education

Throughline Learning

Malika Ali | Chief Innovation Officer, Throughline Learning

Unlocking the Potential of EL Education's Math Approach to Create Equitable, Capable, and Courageous Math Communities for All

EL Education, Math Culture Lab (MCL), and National Implementation Research Network (NIRN)
Megan Bookhout | Research Scientist, National Implementation Research Network
Ximena Franco-Jenkins | Senior Research Scientist & Co-Director, National Implementation Research Network
Linda Grein | Founder and CEO, Math Culture Lab by EduQuate
Deepa Patel | Founder and Principal Consultant, DPA Strategies LLC
Jordan Templeton | Chief Operating & Program Officer, Math Culture Lab by EduQuate

Formative Feedback Study: Enhancing Observation-Based Feedback for Coaching Conversations

University of Virginia, Teach For America, and Annenberg Institute at Brown University
Claire Allen-Platt | Research Scientist, Annenberg Institute at Brown University
Emily Bosch | PhD Candidate in the Dept. of Curriculum and Instruction, University of Virginia School of Education and Human Development
Julie Cohen | Professor, University of Virginia School of Education and Human Development
Robin Greatrex | Senior Managing Director, Learning & Insights, Teach For America
Brendon Krall | Project Director, Annenberg Institute at Brown University
John Papay | Director of Annenberg Institute at Brown University
Hallie Parten | Data Scientist, University of Virginia School of Education and Human Development
Grant Van Eaton | Senior Researcher, Teach For America

Study Overview

Achievement Network (ANet) partnered with Warner County Public Schools (pseudonym) to test whether professional learning services grounded in misconception data, paired with explicit guidance on adaptation, could strengthen teacher practice and improve outcomes. ANet used school-level random assignment to assign 6th-grade math teachers to one of two conditions, intervention or control. All teachers (treatment and control) participated in two PL sessions. The key difference between treatment and control was that the treatment group received guidance on how to make adaptations to the enVision Florida B.E.S.T. curriculum in response to student misconceptions, while the control group only received PL focused on student misconceptions.

Prior to the first PL session, the ANet team, leveraging distractor rationale information from their assessment database as well as district-specific data, identified common student misconceptions aligned to Florida's B.E.S.T. Math standard 6.AR.2.2. The ANet team then co-developed the first PL session with the district partner to be aligned to the HQIM and local standards.

In the first PL session, all teachers (intervention and control) received PL centered on anticipating and interpreting common student misconceptions, specifically on Florida's B.E.S.T. Math standard 6.AR.2.2. The intervention group received the same misconceptions PL plus an added layer: targeted guidance on how to make rigorous, instructionally sound adaptations to the enVision Florida B.E.S.T. curriculum in response to those misconceptions. The PL design was intended for teachers to leave the PL with plans that they could implement in an upcoming lesson. This included an overview of general practices about how to understand misconception data and adapt lessons, as well as specific examples and resources. The PL was also designed to encourage teachers to consider their students' learning preferences and cultural funds of knowledge as they made adaptations to their instructional materials.

For the second PL session, all teachers received PL after students took the next ANet interim assessment. The PL session focused on unpacking student performance on the misconception to identify what adjustments they made to their instruction to accommodate the misconception content, to understand what adjustments worked, and to identify what further support may be needed for students.

ANet then collected and analyzed data to answer the following research questions:

1. To what extent does providing teachers with PL that supports them to make guided adaptations of their HQIM, based on their student misconception data, improve teachers' perception of their HQIM, compared to teachers who receive PL based only on understanding student misconceptions?
2. To what extent do teachers who receive the PL on guided adaptations of HQIM make more adjustments to the HQIM that enhance learning opportunities for students than those who do not?
3. To what extent do students of teachers who receive the PL on guided adaptations of HQIM perform better on the targeted math content (i.e., standards) than the students of teachers who only received PL about understanding student misconceptions?

Data and Methods

Data

For the study, 27 middle schools were randomly assigned to either the treatment or control conditions, with 36 6th-grade teachers in the treatment group and 35 6th-grade teachers in the control group. Across conditions, teachers taught a total of 3,314 students.

In order to answer the above research questions, ANet drew on teacher surveys, focus groups conducted with teachers, student performance data, and student administrative data.

- **Teacher surveys** were administered at the end of the October PL session, with questions focused on their satisfaction with the PL, their perceptions of their curriculum, and actions they may take as a result of the PL. In total, 55 teachers responded to the fall survey.
- **Focus groups** were conducted at the end of the January PL session. ANet staff asked attendees about their experience teaching the relevant standard and any shifts they may have made in response to the original professional learning. In total, 19 teachers from the control group participated across five focus group sessions, and 20 teachers from the treatment group participated across four focus group sessions.
- **Student performance data** on four items related to standard 6.AR.2.2. was analyzed from the winter benchmark assessment to capture student math performance. In total, researchers looked at data from 3,314 students across both study conditions. This included 1,620 students in the intervention group and 1,694 students in the control group whose assessment data were included in the final analysis.
- **Student administrative data**, including demographics, school, prior student B.E.S.T. performance, and classroom advanced/standard status collected by the district, were provided and merged with student performance data.

Methods

Methods are detailed here for each research question.

RQ1: To what extent does providing teachers PL that supports them to make guided adaptations of HQIM, based on their student misconception data, improve teachers' perception of their HQIM, compared to teachers who receive PL based only on understanding student misconceptions?

The ANet team analyzed teacher responses to the October survey administration to test for significant differences in teachers' perceptions of their HQIM by treatment condition. The comparative analysis focused on responses to a set of items measuring curriculum experience provided by NIRN, RAND, and CSAIL. The specific measures include the Acceptability of Intervention Measure (AIM), which assesses teachers' perception of whether a given practice (i.e., the math curriculum) is agreeable and satisfactory. The Feasibility of Intervention Measure (FIM) measures the extent to which a practice can be successfully used or delivered within a given context. The Intervention Appropriateness Measure (IAM) assesses the perceived fit, relevance, or compatibility of a practice (e.g., a math curriculum). Items from RAND's American Instructional Resources Survey (AIRS) included questions about curriculum challenge, engagement, and usability from the 2020 AIRS (Doan et al., 2020), as well as items measuring teacher perceptions that the materials meet the needs of diverse student groups. The ANet team also included items measuring teachers' perceptions of whether the materials meet the needs of diverse student groups.

In the quantitative comparison of curriculum perception items by control or treatment condition, the ANet team controlled for participants' years of experience with the curriculum and their overall satisfaction with the PL.

RQ2: To what extent do teachers who receive the PL on guided adaptations of HQIM make more adjustments to the HQIM that enhance learning opportunities for students than those who do not?

The ANet team created a coding scheme for the focus group transcripts by drafting it through ChatGPT and revising it through consultation with ANet's PL development expert and Warner County Public Schools math team members. This coding schema was then applied to the focus group transcripts by the ANet research team following calibration. Any quotes with codes identified as "shifts teachers made to their curriculum, instruction, or language use" were then further classified by the ANet PL development expert for whether they would a) enhance or b) detract from student learning. The frequency of these different shifts was quantitatively compared across the intervention and control groups.

RQ3: To what extent do students of teachers who receive the PL on guided adaptations of HQIM perform better on the targeted math content than the students of teachers who only received PL about understanding student misconceptions?

Student performance on the four items aligned with the standard was analyzed, both on individual questions and in aggregate. Performance was modeled against the intervention and control groups using a mixed effects model that accounted first for classroom-level performance and then subsequently for student prior FAST performance. The models were then also rerun for separate student demographic subgroups to see if effects persisted within different demographics.

Limitations

- The study was conducted in a shifting instructional landscape. The district was simultaneously implementing new state standards, had recently adopted a new math curriculum, and had a new assessment system that required the research team to engage in ongoing calibration.
- The study was completed on a compressed timeline due to district needs that compressed the IRB submission and custom PL design phases.

Findings

Findings are detailed here for each research question.

RQ1: To what extent does providing teachers PL that supports them to make guided adaptations of HQIM, based on their student misconception data, improve teachers' perception of their HQIM, compared to teachers who receive PL based only on understanding student misconceptions?

Teachers who received the treatment PL on guided adaptations of HQIM rated their curriculum more favorably than teachers in the control PL across feasibility (95% vs. 86%), appropriateness (94% vs. 91%), and domains including differentiation, cultural relevance, and ELL support, in comparison with teachers who received only PL on understanding student misconceptions. While these raw differences suggest the PL supports teacher perceptions of HQIM, after controlling for teachers' years of experience with the curriculum and overall PL satisfaction, the effect was not statistically significant. This may reflect an underpowered analysis given small survey sample sizes (n=22 intervention; n=29 control).

RQ2: To what extent do teachers who receive the PL on guided adaptations of the HQIM make more adjustments to HQIM that enhance learning opportunities for students than those who do not?

The study explored how teachers adapted HQIM using focus groups with treatment and control teachers. Researchers qualitatively coded this data to identify the number of adaptations made and the extent to which the researchers expected those adaptations to enhance learning. Across all teachers who participated in the focus groups (n=39 teachers), teachers in the treatment group reported 19 curriculum modifications, while teachers in the control group reported 11 curriculum modifications. Among the reported curriculum modifications, 37%

of those from the treatment group were coded to be learning-enhancing changes, compared to 27% of the modifications reported by the control group. These raw differences suggest that teachers in the treatment group made more adaptations in their work, and that a higher percentage of their adaptations were learning-enhancing.

As examples of learning-enhancing modifications, teachers described adding in support scaffolds, with one example being, “Especially for [students classified as] ELs, and some of your ESE kids, they need that extra step… So we created visual manipulatives like drawing circles. We laminated mats and had them draw circles and erase them… it just gave us another tool.” Teachers also described engaging in planning forward, proactively adjusting upcoming lessons based on anticipated errors. For example, one teacher described, “Making them understand, 'You've been solving one-step equations for a long time… let me show you what this looked like in 1st grade and 2nd grade.'" Teachers also reported making shifts in their language and instruction that aligned with the specific guidance provided in the PL. As an example of language shifts, one teacher described how they moved towards more specific phrasing, stating, “Being way more specific about how you want to phrase things really carries… [I am] finding better ways to present that information and still spark that visual in their own head.” An instructional shift example was, “It really made me implement more manipulative work than I probably would have used on my own… Instead of just ‘isolate your variable and undo it this way’.”

However, these differences in coded statements from focus groups were not statistically significant (p = .30) after controlling for teachers' years of experience with the curriculum and overall PL satisfaction.

RQ3: To what extent do students of teachers who receive the PL on guided adaptations of HQIM perform better on the targeted math content than the students of teachers who only received PL about understanding student misconceptions?

Student performance was assessed on four assessment items aligned with the math standard that the PL was focused on. On average, scores on the standards-aligned items were higher for students whose teachers received the treatment PL. Students whose teachers participated in the control PL scored 2.2 out of 4 points on average, while students whose teachers participated in the treatment PL scored 2.5 points on average, a small statistically significant effect of 0.085 SD (p=.013). Raw differences were also significant when modeled to account for school-level clustering and student prior math performance. When models were broken out and re-run by student racial/ethnic subgroups, the overall impact was only statistically significant for Black students. For Hispanic/Latine students and White students, the effect was directionally positive, but not statistically significant.

Study Overview

Throughline Learning partnered with a district to test whether providing 2nd through 5th-grade teachers using Eureka Math HQIM with PL on math thinking routines delivered through monthly PLCs and embedded 1:1 coaching improved teacher practice and students' confidence, mindset, and success within the Eureka Math curriculum.

Throughline Learning used teacher-level random assignment to assign 2nd through 5th-grade math teachers to one of two conditions, intervention or control. Teachers in the intervention group received professional learning community (PLC) sessions and eight 1:1 coaching sessions following the PLC sessions. Teachers in the control group did not receive any PL from Throughline. In the intervention group, teachers participated in eight monthly PLC sessions, with each session lasting 90 minutes. In the pilot phase, these were conducted virtually. Teachers in the coaching group received eight one-on-one coaching sessions, each lasting one hour. These coaching sessions were designed to occur monthly, following the PLC meetings, to support implementation. In total, teachers in this treatment group received approximately 20 hours of direct professional learning support (12 hours of PLC time and eight hours of 1:1 coaching) over the course of the school year (October through May). The PL scope and sequence centered on:

1. A structured thinking routine and 7 step process for implementing with multistep Eureka Math's word problems
2. Annotated graphic organizer mapping the cognitive steps
3. Analysis of student work using a problem-solving rubric validated by SCALE from Stanford University.

Additionally, teachers were coached to select multi-step word problems throughout the year so students would have many opportunities to practice the routine. PL was directly mapped into Eureka Math word problems, ensuring instructional practices were embedded within HQIM rather than layered on top of it. The study took place over two years. In the pilot year, Throughline worked with two teachers at one treatment school, then, in the second year, Throughline worked with 32 teachers across eight treatment schools.

Throughline collected and analyzed data to answer the following research questions:

1. What is the effect of combined 1:1 embedded coaching and professional learning services on teacher outcomes?
2. What is the effect of a thinking routine (within a culturally responsive instructional framework) on student confidence, mindset, and success within the Eureka Math curriculum?

Data and Methods

Data

For this study, schools were randomly assigned to either the treatment or control conditions, with one treatment school (n=2 teachers) and three control schools (n=10 teachers) in the pilot year and eight treatment schools (n=32 teachers) and three control schools (n=19 teachers) in the second year.
In order to answer the above research questions, Throughline drew on teacher surveys, teacher self-assessment rubrics, student critical thinking problem-solving rubrics, student experience surveys, and student performance data. All data were collected in both year one and year two.

- **Teacher surveys** were administered at the beginning and end of the year, and included questions about teacher mindsets and confidence in teaching Eureka Math.

- **A competency rubric self-assessment** was administered monthly, where teachers self-assessed the implementation of the seven-step thinking routines protocol.
- **A critical thinking problem-solving rubric** was administered monthly, where teachers assessed students' problem-solving progress. The five-domain problem-solving rubric was used to track how students' mathematical reasoning progressed across three performance tasks (SCALE). The five domains were: 1) identifies what is known, 2) defines the problem, 3) generates possible solution strategies, 4) applies problem-solving steps, and 5) evaluates solutions.
- **A student experience survey** was administered at the beginning and the end of the year about academic mindset, engagement, and sense of belonging within the math class
- **Student performance** on the STAR, MAP, AIMSweb, or iReady Math interim assessments was analyzed to capture student math performance.

Methods

Methods are detailed here for each research question in the pilot-year analyses.

RQ1: What is the effect of combined 1:1 embedded coaching and professional learning services on teacher outcomes?

To identify the effect of receiving 1:1 embedded coaching and PLC services on teacher outcomes, the Throughline team conducted a descriptive analysis of teacher responses to the end-of-year educator mindset and professional learning survey from the pilot year. The team examined raw differences between teachers in the treatment group and those in the control group.

RQ2: What is the effect of a set of two math thinking routines (within a culturally responsive instructional framework) on student confidence, mindset, and success within the Eureka Math curriculum?

In order to identify the effect of having a teacher who received the 1:1 embedded coaching and PLC services on student outcomes, the Throughline team descriptively analyzed differences in student growth on the STAR Math interim assessment, as well as their problem-solving rubric scores, by whether their teacher was in the treatment or control group in the pilot year. The team also plans to analyze student experience survey results, along with all other data collected in the second year.

Limitations

- In year one, the pilot year, the sample size was small due to challenges with recruitment. While in year two, 30 treatment educators and 30 control educators were recruited, some control teachers dropped out of the project for various reasons, so the comparison group was smaller than the treatment group. Ideally, the same treatment and control teachers would have been followed over two years so that proximal and distal measures could be examined given that assessment scores tend to be expected to show more growth after the problem solving skills are mastered. In both the pilot year and year two, only one year of data was collected because the project included a different set of teachers each year. In year two, survey response rate from control teachers was low for both timepoints, so their results have been excluded from reporting due to low statistical power.

- School leaders prioritized educators who they felt needed the intervention so baseline educator mindset and student assessment scores were lower for the treatment group at the beginning of the study. Still, a higher percentage of treatment teachers showed growth in mindset scores.

Findings

Findings are reported by Year 1 and Year 2.

RQ1: What is the effect of combined 1:1 embedded coaching and professional learning services on teacher outcomes?

Year 1

Teachers in the treatment group received a combination of PLC and 1:1 Embedded Coaching supports. There were descriptive differences in teacher confidence and willingness to deprivatize their practice that suggest this PL structure was supportive for teacher confidence and openness. Specifically:

- Sharing Artifacts: 100% of treatment teachers felt "to a great extent" comfortable sharing artifacts from their classrooms, compared to only 20% of control teachers.
- Vulnerability: 100% of treatment teachers felt comfortable "sharing something I am wondering about in my own teaching," compared to 0% of the control group.
- Sharing Ideas: 50% of treatment teachers felt comfortable sharing ideas they were "unsure about," whereas 0% of the control group felt comfortable doing so.

Teachers in the treatment group also reported that they received feedback more frequently and were prompted to reflect on their own practice more often than those in the control group.

Year 2

Teachers in the treatment group received a combination of PLC and 1:1 Embedded Coaching supports. There were descriptive differences in teacher confidence and willingness to deprivatize their practice that suggest this PL structure was supportive for teacher confidence and openness. Specifically:

- Strategy Implementation: 100% of treatment teachers implemented an instructional strategy based on PL
- Reflection Prompting: 89% of treatment teachers reported they were often or very often prompted to reflect on their practice during the year.
- Reflecting on Teaching Practices: 89% of treatment teachers grew their teaching practices.
- Student Perseverance: 80% of teachers reported their students developed consistent or independent strategies to persevere through challenges (beginning of year: 6%; end of year: 86%).
- Sharing Thinking: 79% of teachers reported their students developed consistent or independent use of structure/strategies to share their thinking processes (beginning of year: 13%; end of year: 92%)
- Complex Problems: 61% of teachers reported that students developed consistent or independent opportunities to solve complex problems (beginning of year: 39%; end of year: 100%).

Teachers in the treatment group also reported growth in their confidence in implementing Eureka effectively, and in their comfort sharing live observations or recorded instruction at the end of year. Treatment teachers who participated in more PL sessions rated their confidence implementing Eureka, student problem solving skills, and student perseverance at end-of-year significantly higher than treatment teachers who participated in fewer sessions (+0.25 higher average confidence rating per session attended; +0.13 higher average PS Skills rating per session attended; 0.16 higher average perseverance rating per session attended; $p<0.04$ for all measures). Control teachers did not participate in PL sessions.

RQ2: What is the effect of a set of a math thinking routine (within a culturally responsive instructional framework) on student confidence, mindset, and success within the Eureka Math curriculum?

Year 1

Students whose teachers received the Throughline Learning PL had slightly higher increases in their median scores on the Aimsweb math interim assessment than students whose teachers did not. 2nd-grade treatment students increased their median score by 37%, compared to the 34% increase in the control classroom, and in the 5th-grade class, treatment students increased by 15%, compared to the 14% of the control classroom.

Students whose teachers received PL on thinking routines saw growth in their mathematical reasoning. During year two, all eight schools showed growth from baseline to mid-year on their rubric scores. Overall, each domain increased from Task 1 and Task 2. The final domain of "Evaluates Solutions" showed the most growth, with a 40% increase in scores, followed by the domain "Defines what is known," with an increase of 30%.

In the pilot year, for all grades, overall rubric performance rose across the year. However, there was heterogeneity by grade level in how much scores improved. For overall domain scores, 2nd-grade students in the intervention group improved their total scores by 24% and 5th-grade students improved by 43%. For example, 2nd-grade students who had teachers in the treatment group showed a 67% increase in their ability to identify what is known in word problems, indicating a stronger capacity to organize relevant information. 5th grade students showed a 96% increase in their ability to define the problem, reflecting a substantial growth in articulating the core question and what must be solved. Year two findings, expected July 2026, will provide a more complete picture with a larger sample across eight schools.

To further explore whether growth on the problem-solving rubric domains was associated with students' broader performance, researchers conducted a correlational analysis between students' spring Aimsweb scores and final rubric performance. In particular, there was a moderate correlation between Mental Computation Fluency and the rubric domain "Evaluates solutions" ($r=.46$), as well as additional moderate correlations between Concepts & Applications and rubric domains tied to later-stage reasoning (e.g., Applies problem-solving skills and Evaluates solutions). At the same time, correlations between overall Math Score and rubric domains were generally low, and showed low correlations between "Identifies what is known" and "Defines problem" and end-of-year assessment scores, suggesting that gains in early-stage problem deconstruction may not immediately translate into higher standardized assessment performance.

Year 2

Students whose teachers received the Throughline Learning PL had slightly higher increases in their overall scale scores from beginning to end of year on the iReady and MAP math assessment compared to students whose teachers did not, but the increase was not significant (+0.93 SD; $p=0.61$).

Students whose teachers received PL on thinking routines saw growth in their mathematical reasoning. During year two, on average, students in all eight intervention schools showed significant growth of nearly 50% (49% growth, $p<0.0001$) from baseline to mid-year in their mathematical reasoning rubric scores, with increases across all assessed domains. Overall, each domain increased from Task 1 to Task 4. The final domain of "Evaluates Solutions" showed the most growth, with a 65% increase in scores ($p=0.1$), followed by the domain "Applies problem solving steps," with an increase of 21% ($p=0.04$).

In year two, for all grades, overall rubric performance rose significantly across the year ($p<0.001$). However, there was heterogeneity by grade level in how much scores improved. For example, 4th-grade students who had teachers in the treatment group showed a significant, 29% increase in their ability to identify what is known in word problems, indicating a stronger capacity to organize relevant information. 2nd grade students showed a 130% increase in their ability to evaluate solutions, reflecting a substantial growth in reviewing their thinking process.

Qualitative data suggest that being in a classroom led by a teacher who received the PL on thinking routines resulted in positive shifts in students' mathematical identity. Data from student-written reflections ("I used to think... Now I think...") revealed a shift from fixed mindsets to growth mindsets and metacognition. For instance, students moved from believing they had to do math "in their head" to understanding they "have to show my work." This illustrates a shift from abstract thinking to concrete thinking. Additionally, early learnings suggest that students shifted from thinking they were "bad at math" or unable to solve problems alone to feeling "better/smarter" and empowered to solve problems independently. The study suggests that thinking routines successfully help students—particularly in urban settings—decode complex math problems and improve their math identity.

Eureka Task Growth by Grade and Rubric Domain

Grade	Composite score	Identifies what is known	Defines the problem	Generates solution strategies	Applies problem-solving steps	Evaluates solutions
All Grades	49%***	35%	45%	38%	21%**	65%*
Grade 2	97.25%***	67%	96%	88%	11%**	130%°
Grade 3	39.03%***	32%	54%	31%	24%	47%
Grade 4	41.00%***	29%**	11%	15%**	19%**	56%*
Grade 5	37.19%***	20%**	25%	31%**	29%	58%°

*** = $p<0.0001$; ** = $p<0.05$; * = $p<0.1$; ° = $p<0.11$

Study Overview

Math Culture Lab (MCL) partnered with a public charter middle school in Washington, D.C., to examine whether a professional learning model emphasizing 1:1 coaching, peer collaboration, and strong teacher–student relationships alongside HQIM led to growth in middle school students' mathematical identities, sense of belonging, and academic achievement.

At the intervention school, the principal, one School-based Math Culture Lead (SMCL), four general education teachers, and four special education teachers attended a three-hour Vision Institute to establish shared goals and a common understanding of strong mathematics teaching and learning. Then, MCL provided monthly 1.5-day onsite support and twice-monthly one-on-one coaching for the SMCL, focused on clarifying leadership roles, building coaching capacity, collecting and analyzing learning walk data, and planning next steps for coaching and professional learning. Additionally, MCL and the SMCL engaged teachers in regular professional learning, facilitated at least monthly and responsive to learning walk data. These sessions focused on strengthening teachers' pedagogical skill, content knowledge, and their ability to cultivate positive student math identities and strong teacher–student relationships. Teachers also participated in ongoing peer collaboration and received one-on-one coaching from the SMCL designed to strengthen instructional practice and support effective use of high-quality instructional materials.

MCL collected and analyzed data to answer the following research questions:

1. Across the academic year, is there growth in the outcomes of
    a. Teachers' confidence in establishing a positive classroom environment with strong teacher–student relationships?
    b. Use of equitable math instructional practices, as evidenced by integrity to the Illustrative Mathematics/Desmos Math 6-8 curriculum? and
    c. Students' math identity, sense of belonging in math, and academic achievement, particularly for students who are Black, Latine, or those experiencing poverty?
2. Does implementation (e.g., acceptability, feasibility, appropriateness of the curriculum) predict teacher and/or student outcomes?
3. Are associations between implementation and student outcomes moderated by teacher–student relationships?

Data and Methods

Researchers used a mixed-methods case study design in a single school, with eight participating teachers and no comparison group. In order to answer the above research questions, MCL drew on teacher surveys, student surveys, semi-structured interviews with staff, listening sessions with students, classroom observations, and student performance data.

- **Teacher surveys** were administered to assess teacher implementation confidence, math instruction, culturally responsive teaching self-efficacy, and relational self-efficacy. In total, six teachers completed both the fall baseline and spring follow-up surveys.
- **Student surveys** were administered to measure students' positive math identity, belonging, engagement, growth mindset, and perceptions of cultural pedagogy. These students were

assigned to four teachers in the 5th through 8th grades participating in this project. Of those, 258 students initiated the survey and completed at least one construct, and 247 completed the survey during the fall of 2024. A total of 321 students completed the survey during the spring of 2025.

- **Semi-structured interviews** were conducted with two coaches and the middle school principal. The semi-structured interview protocols developed by the NIRN and the MCL team consisted of 14 and 11 questions, respectively. Each interview lasted 45-60 minutes. The interviews were recorded and transcribed.
- **Listening sessions** were held with students to gather student voices on their math experience. Student listening sessions with students in 5th, 6th, 7th, and 8th grades at the middle school (n=24, 6 students per grade). Each of the four listening sessions involved students from a single grade. Students were randomly selected from those who had completed the student survey in the fall, with stratification by class section, gender, and race/ethnicity where possible to obtain a representative sample. 54% of the participating students identified as male, 42% as female, and 4% as nonbinary. 50% of the students identified as Hispanic, 25% as Black, 12.5% as White, and 12.5% as more than one race. English Language Learners (ELLs) comprised 42% of the student population, while 58% were classified as non-ELLs.
- **Classroom observations** were conducted to assess implementation integrity and instructional shifts. A total of 28 classroom observations were completed across four cycles: September (n = 8), October (n = 6), November (n = 6), and April/May (n = 8).
- **Student math performance** was assessed using the District of Columbia Comprehensive Assessments of Progress in Education (DC CAPE) mathematics scores, the general assessment administered each spring to students in grades 3–8. It measures student proficiency relative to grade-level academic standards. The study drew on baseline scores from spring 2024 and post-intervention scores from spring 2025 (n=333 students). Second, interim benchmark achievement was assessed using the NWEA MAP Growth Mathematics assessment at the beginning, middle, and end of the school year. NWEA MAP Growth reports scores as equal-interval measures of academic achievement, which allow for tracking growth over time, independent of grade level. NWEA MAP also provides percentile ranks to indicate how a student's performance compares to a nationally normed sample of same-grade peers. The study drew on beginning-of-year, middle-of-year, and end-of-year MAP scores (n=342 students).

Methods

Methods are detailed here for each research question.

RQ1: Across the academic year, is there growth in the outcomes of teachers' confidence in establishing a positive classroom environment with strong teacher–student relationships; use of equitable math instructional practices as evidenced by integrity to the IM/Desmos Math 6-8 curriculum; and students' math identity, sense of belonging in math, and academic achievement, particularly for students who are Black, Latine, or those experiencing poverty?

Descriptive quantitative analyses were used to examine changes over time in teacher perceptions of teacher–student relationships and student outcomes throughout the academic year. Specifically, dependent t-tests and linear regression models, including latent change score analyses, were used to identify significant differences between time points and among student groups. To assess the impact of demographic predictors, students' reports of classroom cultural elements, and math beliefs and experiences, a multiple regression model was run using Bayesian estimation in Blimp 3.0. Because teacher and grade were fully confounded (i.e., each grade had a unique teacher pair), and teachers were not modeled as a separate level due to the small number of clusters (four), the grade dummy variables necessarily capture both grade and teacher differences. These effects cannot be disentangled in the present design. Other predictors included race/ethnicity, English learner status, school belonging, perception of teachers' cultural respect, math self-efficacy, growth mindset, engagement, enjoyment, and the prior year's DC CAPE scale score. Based on school-reported descriptive data indicating that 7th-grade students demonstrated the greatest raw growth in the outcome, 7th grade was selected as the reference group in regression analyses. This parameterization allowed for direct tests of whether other grade levels differed significantly from 7th grade after adjusting for demographic and psychosocial covariates. Framing the 7th grade as the comparison group enabled evaluation of whether the observed descriptive advantage persisted under statistical controls. White students were designated as the racial reference group to enable a direct examination of whether Black and Hispanic students exhibited significantly different outcomes after adjusting for covariates.

Qualitative findings were incorporated to provide insights into the perceived impact of the professional learning program on these outcomes.

RQ2: Does implementation (e.g., acceptability, feasibility, appropriateness of the curriculum) predict teacher and/or student outcomes?

Bivariate correlations were used to examine relations between implementation constructs and teacher outcomes at each time point. The small number of teachers, paired for each grade such that the teacher was confounded with the grade (n=4 teacher groups), prevented analyses linking implementation to student outcomes. Qualitative findings complemented the statistical analysis by providing a deeper understanding of the implementation process, strategies employed, and their impact on outcomes.

RQ3: Are associations between implementation and student outcomes moderated by teacher–student relationships?

Ultimately, the participating school's final classroom setup this year prevented a quantitative exploration of this question through cross-level moderation analyses. Teachers were paired, and each pair was then assigned to all math sections in a single grade. Therefore, the two teachers' impacts on their students could not be disaggregated, and averaging their scores resulted in an n of 4 for the teacher indicators, which is insufficient to answer a moderation question. However, qualitative findings enhanced understanding by exploring the impact of the program's implementation on teachers' confidence in building strong relationships.

Limitations

- The study drew on a small sample size.
- The distribution of students across racial/ethnic categories was uneven, making some subgroups too small to meaningfully examine
- The study spanned a one-year timeframe, which is likely insufficient for meaningful programmatic or instructional changes to take hold or for their effects to become observable
- The survey and observation instruments may not have been sensitive enough to detect subtle variation or incremental change.
- Teachers were paired and co-taught all students in the grade, so teacher effects could not be distinguished from developmental effects or grade-cohort effects.
- Only one school was included; the lack of a comparison group makes it difficult to interpret change over time, and generalizability may be limited.

Findings

Findings are detailed here for each research question.

RQ1: Across the academic year, is there growth in the outcomes of teachers' confidence in establishing a positive classroom environment with strong teacher–student relationships; use of equitable math instructional practices as evidenced by integrity to the IM/Desmos Math 6-8 curriculum; and students' math identity, sense of belonging in math, and academic achievement, particularly for students who are Black, Latine, or those experiencing poverty?

There were no significant differences in average teacher-reported perceptions of confidence in teacher–student relationships across time points; overall, their levels of confidence remained largely the same. Six out of eight teachers completed the teacher survey at both time points, with the seven teachers who responded in the fall averaging 8.29/10 (range: 7.33 – 8.67). For the seven teachers (including one different teacher) who responded in the spring, their average was 7.76 (range: 6.67 - 9.00). While teachers' confidence remained high throughout the year, learning walk data captured by the MCL coach and SMCL showed little change in instructional practice over the school year, likely explaining the student outcome results.

There were mismatches between student perceptions of relationships and teacher perceptions. For example, MCL facilitated data review sessions with math coaches that looked across three distinct sources: learning walk data, teacher perception surveys, and student surveys. The MCL coach shared this triangulated data with the SMCL to explicitly demonstrate where the student perception was not matching what the teachers believed was working. This process moved the conversation from subjective feelings of confidence to objective gaps in student belonging. Overall, this approach built both MCL coaches' and the school staff's data literacy, including the types of data that might be possible to collect, how to collect that data, how to analyze that data, and ultimately, ways to leverage that data in making strategic improvements to the culture of math at the school.

Qualitative evidence suggested that, while a vision for equitable and ambitious mathematics instruction may have been introduced or discussed at the school, it was not yet consistently enacted across classrooms. For example, one teacher did not consistently use the HQIM, and some teachers continued to express beliefs that step-by-step direct instruction was the most appropriate approach for their students. As the SMCL described in their interview, some teachers believed students could not engage with grade-level standards until perceived gaps from earlier grades were addressed. This reflects both persistent beliefs about student capability and uncertainty about how to enact problem-based instruction in practice.

MCL's program approach may rely too heavily on the School Math Culture Lead (SMCL), particularly those new to the role. The SMCL is a school-based staff member or team charged with leading the transformation of mathematics culture and instruction at the school. The Math Culture Lab coach works directly with the SMCL to build their leadership and math instructional expertise. The SMCL is then responsible for coaching and supporting teachers. This study revealed that it is unrealistic to expect SMCLs, particularly those new to their position (i.e., in their first or second years), to independently lead the work and coach teachers while they are still building their own knowledge and skills. The MCL coach reported that the SMCL needed time to build their math team: "[SMCL] is still in the process of creating a coherent team, and it's January, so that's a little late in the process to be doing that, but that's where he is." While the SMCL was trying to build a coherent team, they were also responsible for coaching individual teachers and attending to changing mindsets. The MCL coach shared that "teachers that are new are not as familiar with IM and are not as willing to utilize it as it is written" and "there's a general fixed mindset belief about students' capacities." These dual responsibilities seemed to be overly burdensome and pulled the SMCL in multiple directions, likely interfering with both their ability to transform culture and support individual teachers with deepening their instructional capacity.

Ideally, coaching is led by the SMCL. However, the Math Culture Lab learned that when the SMCL is new, coaching must be led by a Math Culture Lab coach initially. This structure allows the SMCL to build their own instructional and coaching competencies while teachers develop their own pedagogy that demonstrates vision-aligned practices.

School leaders emphasized walkthroughs with feedback as an important lever for improving math accessibility, noting ongoing work to refine feedback protocols. Leaders also identified implementation challenges that stood in the way of advancing outcomes, including teacher fixed mindsets, skill gaps, unfamiliarity with IM, inconsistent expectations and follow-through, limited resources for special populations, and a largely new teaching staff still learning the school's math vision and MCL's approach. Despite these challenges, leaders highlighted the value of MCL's PL, resources, and guidance in building coaching and departmental capacity, and described a gradual shift toward more problem- and inquiry-based instruction that is expected to take several years to fully take hold.

There was no overall growth in student math identity and achievement. Table 1 summarizes the direction of the change, with notes on how outcomes varied by student subgroups. Differences by demographic groups, as well as differences by grade, emerged across outcomes. For example, a decrease in belonging was largely driven by the 7th-grade students, whereas 5th-, 6th-, and 8th-grade students showed no significant change. Race and ethnicity predicted differences only in the trajectory of math self-efficacy, such that Black and Hispanic students showed a decline over the year as compared to their White peers. The demographic makeup of the school was predominantly Hispanic; therefore, examining trajectories of change with an interaction for all different racial/ethnic groups by grade was not possible, as there were not a sufficient number of students who identified as Black, White, and another race in each grade to serve as separate comparison groups. Exploratory regression analyses examining change in math self-efficacy within each of the two largest groups (Hispanic students and Black students), however, showed that differences were varied; for example, among Hispanic students, 6th-grade students showed the greatest decline, whereas among Black students, 6th-grade students had the most positive outcome, and were predicted to increase in their math self-efficacy.

Table 1. Results by Student Outcome

Outcome	Change	Notes
Math Engagement	↓	English language learners reported greater decrease in engagement than non-English Language Learners
Math Enjoyment	↓	No differences by student subgroups
Growth Mindset	=	No differences by student subgroups
Math Self-efficacy	=	Students who identify as Black and Hispanic showed a decline in math self-efficacy
Belonging in Math	↓	Decline in belonging was largely driven by 7th grade
Math Achievement	=	7th-grade students showed gains in achievement; Students who identify as Black and Hispanic demonstrated declines relative to their White peers; and English Language Learners declined as compared to non-English Language Learners

Note: "↓" indicates a decline in outcomes; "=" indicates no change in outcomes

Student academic achievement

Achievement was assessed in two ways. First, year-end summative achievement was assessed using the DC CAPE mathematics scores Second, interim benchmark achievement was assessed using the NWEA MAP mathematics assessment at the beginning, middle, and end of the school year.

DC CAPE scores fall into five performance levels, for which thresholds are scaled by grade. Table 2 shows the levels, labels, threshold scores, and number of students at the school who achieved each performance level in both spring 2024 (pre-intervention) and spring 2025 (post-intervention). There were no significant differences in the percentage of students who scored at Levels 4 and 5 from 2024 to 2025. Additionally, the 2024 average overall score was 720.66, while the 2025 average overall score was 719.74.

In regression models predicting DC CAPE scores as a function of time and demographics, growth differed by student groups. Compared to White students, Hispanic students ($\beta = -.195$, 95% CrI [−.345, −.046]) and Black students ($\beta = -.285$, 95% CrI [−.423, −.146]) demonstrated lower growth, whereas students of another race ($\beta = -.053$, 95% CrI [−.151, .048]) did not differ. Compared to 7th graders, students in 5th ($\beta = -.346$, 95% CrI [−.444, −.243]), 6th ($\beta = -.266$, 95% CrI [−.362, −.170]), and 8th grade ($\beta = -.351$, 95% CrI [−.446, −.253]) showed lower growth. English Language Learners ($\beta = -.138$, 95% CrI [−.234, −.043]) also exhibited lower growth than non-English Language Learners. The interval for female students ($\beta = -.019$, 95% CrI [−.094, .058]) included zero, indicating no difference from male students.

Greater math self-efficacy ($\beta = .184$, 95% CrI [.079, .289]) and growth mindset ($\beta = .089$, 95% CrI [.005, .168]) were associated with greater academic growth. In contrast, belonging ($\beta = -.008$, 95% CrI [−.092, .075]), cultural respect ($\beta = -.007$, 95% CrI [−.087, .075]), 1-item perceived cultural respect ($\beta = .035$, 95% CrI [−.042, .112]), engagement ($\beta = .023$, 95% CrI [−.075, .117]), and enjoyment ($\beta = -.032$, 95% CrI [−.132, .071]) all had intervals that included zero, suggesting no associations with academic growth. Prior-year achievement was the strongest predictor ($\beta = .467$, 95% CrI [.377, .551]), indicating substantial stability in academic performance over time.

For NWEA, at each time point (Beginning of Year, BOY; Middle of Year, MOY; End of Year, EOY), percentile ranks indicate how a student's performance compares to a nationally normed sample of same-grade peers (e.g., a percentile rank of 60 means the student performed as well as or better than 60% of students in the national comparison group). Across the full sample, math achievement scores increased from 31.53 to 32.06 to 34.48 across the three time points.

RQ2: Does implementation (e.g., acceptability, feasibility, appropriateness of the curriculum) predict teacher and/or student outcomes?

Survey data showed that there were no significant differences in teachers' perceptions of the acceptability, appropriateness, and feasibility of the implementation of the Illustrative Math curriculum from the baseline fall survey administration to the spring survey administration. Indicators of implementation (acceptability, appropriateness, and feasibility) were significantly positively correlated with each other. Given the small sample size of teachers, none significantly predicted other outcomes. However, positive trends were observed between implementation indicators and quality of professional development, teacher knowledge, self-efficacy for math instruction, and teacher–student relationships.

RQ3: Are associations between implementation and student outcomes moderated by teacher–student relationships?

As seen in the models described above, student belonging was not associated with academic growth. The small N of teacher pairs (n=4, 1 team per grade) was insufficient for a cross-level quantitative moderation analysis. However, qualitative evidence suggests that students' math experiences depend on individual teachers' abilities to establish a positive classroom environment, strong teacher–student relationships, and their own math content knowledge and instructional skill. During listening sessions, students reported the following about their math classroom experiences:

- feeling the most engaged when teachers used interactive activities;
- feeling confident in their math abilities when teachers challenged them with hard math problems or provided support and encouragement;
- feeling respected when teachers took the time to explain concepts, provide additional help, and provide positive reinforcement; and
- experiencing more belonging and joy when teachers tailored the difficulty of math concepts to each student's understanding.

Implementation data show that teachers varied in their comfort and skill in creating math classrooms with these attributes, which may explain why the student outcomes, on average, reveal no clear changes overall or by student subgroups.

Study Overview

The University of Virginia (UVA) and the Annenberg Institute at Brown University (AIB) partnered with Teach for America (TFA) to study the impact of different observation-based feedback formats on novice teacher development. Specifically, the study examines TFA's enterprise-wide adoption of an observation tool designed to assess and improve teacher–student interactions, the Classroom Assessment Scoring System (CLASS). Alongside qualitative interviews exploring coach and teacher experiences, the team conducted a randomized experiment to test the effectiveness of different routines for providing teachers feedback in coaching conversations.

The research team randomly assigned TFA coaches to provide early-career teachers with one of three feedback formats in coaching debriefs: 1) numeric score reports; 2) numeric score reports with comparative data; and 3) business-as-usual descriptive feedback. The study examines impacts on coach and teacher perceptions, coaching conversations, and teacher performance growth. Building on a pilot study in 2024-25 with 10 coaches (n=96 teachers), the enterprise-wide experiment randomized 129 coaches (n=4,166 teachers) to these three conditions. Researchers integrated qualitative and quantitative data, including teacher and coach surveys, CLASS observation scores, coaching transcripts, and interviews with central staff, coaches, and teachers to answer the following research questions:

1. How do TFA teachers, coaches, and national staff view the CLASS observation system and related feedback?
2. Do teachers and coaches report that numeric score reports, numeric scores with comparisons, or descriptive feedback promote better (or worse) coaching conversations?
3. Do teachers improve more (or less) after receiving numeric feedback, numeric feedback with comparisons, or descriptive feedback in coaching conversations?
4. What are the characteristics of coaching conversations with and without numeric scores (e.g., use of CLASS language, sentiment, emotional tenor, etc.)?

Data and Methods

Data

In order to answer the above research questions, the UVA/AIB team drew on teacher surveys, coach surveys, interviews conducted with TFA central office staff, coaches, and teachers, classroom observation scores, coach administrative data, and teacher administrative data. They conducted a pilot RCT study during the 2024-25 school year, an enterprise-wide RCT during the 2025-26 school year, and an interview study spanning both years. The enterprise-wide RCT and interview studies are still ongoing, with final results anticipated at the end of the school year. Initial findings are reported here.

The following data was collected during the Pilot RCT (complete) and full RCT thus far (in progress).

Pilot RCT

- **Teacher surveys**: Post-observation CLASS debrief surveys; a mid-year survey, and an end-of-year survey (n=96 responding K-12 teachers)
- **Coach surveys**: Post-observation CLASS debrief surveys; a mid-year survey, and an end-of-year survey (n=10 responding coaches)
- **CLASS observation scores** (n=289 teacher-level observation scores)

- **Teacher and coach administrative data**, including information on role and demographics, to link with the other data sources

Full RCT (*In progress; sample sizes reflect data collected as of June 2026*)

- **Teacher surveys**: Post-observation CLASS debrief surveys; a mid-year survey, and an end-of-year survey (n=4,166 K-12 teachers)
- **Coach surveys**: Post-observation CLASS debrief surveys; a mid-year survey, and an end-of-year survey (n=129 coaches)
- **CLASS observation scores**: Fall 2025 (n=7,525 cycle observations representing 3,723 teachers)
- **Teacher and coach administrative data**, including information on role and demographics, to link with the other data sources

Interview study

- Interviews with TFA central office staff (n=11), coaches (n=19), and teachers (n=5). We identified national staff members using snowball sampling, coaches using random sampling (sampling equally from each treatment condition), and teachers using convenience sampling from within sampled coach cohorts.

Methods

RQ1: How do TFA teachers, coaches, and national staff view the CLASS observation system and related feedback?

Researchers conducted 60-minute semi-structured interviews with TFA central office staff, coaches, and teachers. Interview transcripts were first qualitatively coded using a deductive approach based on TFA's stated intended uses of the CLASS tool. Subsequently, researchers refined the coding structure using an inductive approach to capture emergent themes regarding how stakeholders discuss and perceive CLASS in practice. We supplemented these interview findings with descriptive results from the surveys.

RQ2: Do teachers and coaches report that numeric score reports, numeric scores with comparisons, or descriptive feedback promote better (or worse) coaching conversations?

To answer RQ2, researchers randomly assigned coaches to the three treatment conditions within geographic hub randomization blocks. Randomization was successful; baseline characteristics were balanced across conditions. We composited survey responses related to perceptions into a single composite "usefulness" measure (one for coaches, one for teachers).

To estimate impacts on teachers' perceived usefulness of the coaching conversation, we used a two-level hierarchical linear model with geographic block fixed effects, clustering teachers within their randomly assigned coaches. To estimate impacts on coaches' perceived usefulness, we used a two-way analysis of variance (ANOVA) that incorporated geographic block fixed effects, without intermediate clustering. Consistent with our pre-registration plans (SREE IDs: 26160.1v1 and 26160.2v1), we estimated impacts overall (numeric vs. descriptive) and by treatment condition.

RQ3: Do teachers improve more (or less) after receiving numeric feedback, numeric feedback with comparisons, or descriptive feedback in coaching conversations?

Results are pending spring 2026 CLASS data collection.

RQ4: What are the characteristics of coaching conversations with and without numeric scores (e.g., use of CLASS language, sentiment, emotional tenor, etc.)?

Results are pending spring 2026 data collection.

Limitations

- Without spring 2026 data collection, the sample size is more limited.
- This study focuses on TFA corps members, early-career educators whose feedback preferences may be different from those of other teachers.

Findings

Here, we detail initial findings for our first two research questions. When spring coaching conversations and CLASS data are available, we will estimate impacts on these outcomes. All results should be treated as preliminary.

RQ1: How do TFA teachers, coaches, and national staff view the CLASS observation system and related feedback?

Stakeholders across TFA saw CLASS as a tool that establishes a common language and focuses coaching conversations. Both teachers and coaches found the coaching conversations during the enterprise-wide RCT helpful. Teachers predominantly felt "reflective" (80%), "encouraged" (79%), and "valued" (53%) after coaching conversations. Both coaches and teachers reported that the conversations were clear in purpose, highlighted teacher strengths, and were useful for identifying specific areas for growth. Teachers reported looking forward to their next observation. In Table 1, we present means and standard deviations (SD) of teacher and coach survey responses. Overall, both groups had quite positive reactions to the conversations.

Qualitative interviews with 35 interest holders (national staff, coaches, teachers) revealed similar patterns. Coaches reported that CLASS was useful to focus conversations. However, they report widely varied approaches to delivering feedback; integration of CLASS data into feedback routines is uneven but increasing over time. Teacher interviews similarly reveal varying levels of familiarity with CLASS language and CLASS-aligned growth goals. Coaches reported needing more support to translate observational data into actionable steps and connect that data to concrete resources for the teacher. Recent interviews suggest that numeric score reports can serve as one such support, accelerating TFA corps members' internalization of CLASS language and increasing coaches' comfort with and use of the CLASS tool.

Survey Items	Mean	SD
Teacher Responses		
I understand the purpose of the CLASS coaching conversation	4.53	0.87
The CLASS coaching conversation was:		
useful for my development	4.48	0.92
highlighted my strengths	4.51	0.90
helped me identify specific areas for growth	4.55	0.89
I am looking forward to my next CLASS observation	4.37	0.98
Coach Responses		
I understand the purpose of the CLASS coaching conversations	4.58	0.67
The CLASS coaching conversations:		
were useful for my CMs' development	4.15	0.87
provided an opportunity to highlight CMs' strengths	4.50	0.64
helped CMs identify specific areas for growth	4.48	0.67
helped CMs identify clear next steps	4.00	0.84
Teachers felt anxious or nervous in CLASS coaching (reverse code)	3.71	1.14

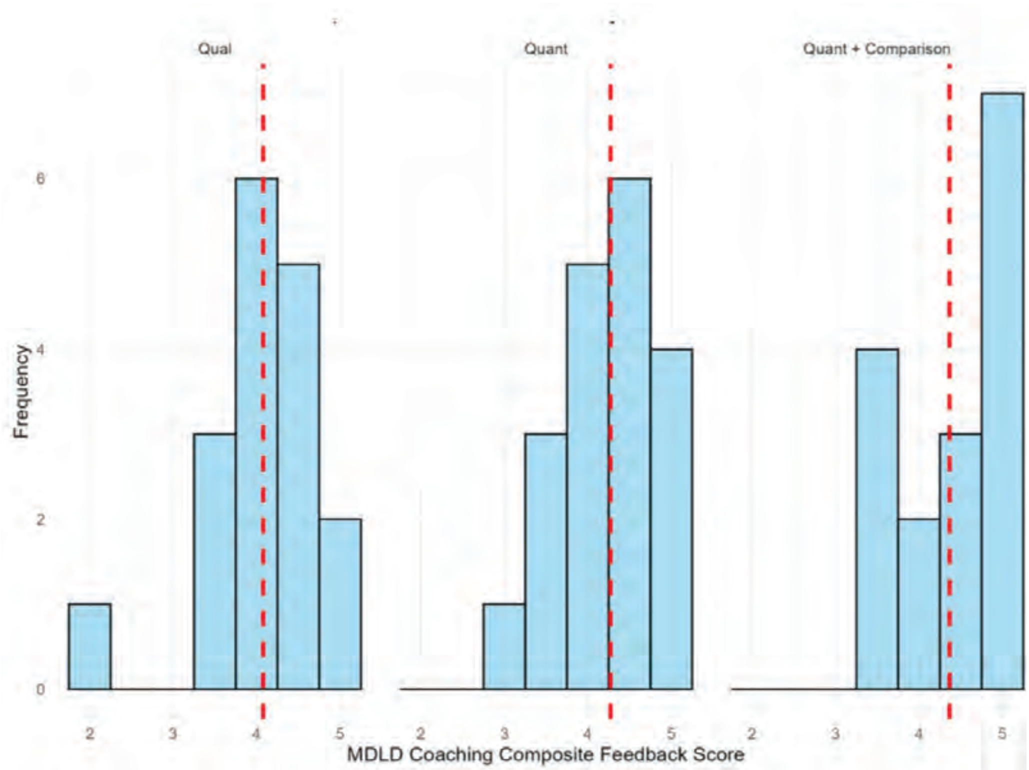
NOTE: Likert response items range from 1 to 5 as follows: 1 - Strongly Disagree, 2 - Somewhat Disagree, 3 - Undecided, 4 - Somewhat Agree, 5 - Strongly Agree.

RQ2: Do teachers and coaches report that numeric score reports, numeric scores with comparisons, or descriptive feedback promote better (or worse) coaching conversations?

Initial findings from the full-scale RCT rely on immediate post-conversation surveys that capture teachers' initial reactions to their Fall 2025 CLASS debrief sessions. Coaches much preferred the numeric format. In Figure 1, we show the distribution of usefulness ratings for coaches in the three conditions. We see strong visual evidence that coaches much preferred the numeric condition and especially the numeric format with the additional comparative information about average CLASS scores. We estimate treatment effects more formally using the models described above. We present the initial results in Table 2. These estimates suggest that coaches much preferred the numeric format, both with (effect size = 0.49) and without (effect size = 0.32) comparison data. Many more coaches reported that they understood the purpose of the conversation, that the conversations were useful for teacher development, and that the conversations provided clear areas for strength and growth when they were provided numeric reports. While these estimates are not statistically significant at traditional levels, the magnitude of the impacts is quite large.

As seen in Table 2, teachers also reported more satisfaction with coaching debriefs when they received numeric score reports with comparison data (effect size = 0.10), although the differences were not statistically significant. At minimum, it is clear that numeric feedback was not less popular than descriptive feedback. However, these findings provide suggestive evidence that teachers may have preferred quantitative feedback with comparison data. Table 2. Estimated treatment effects of receiving quantitative CLASS feedback reports (Quant) with and without comparison data compared to business-as-usual (BAU) descriptive feedback on perceived usefulness of feedback conversations, fall 2025.

Figure 1. Distribution of coach composite ratings of the usefulness of coaching conversations by treatment condition, fall 2025.



Comparison	Teacher-evel	Coach-Level
Quant (no comparison) vs. BAU	-0.005 *p*=0.95	0.32 *p*=0.54
Quant + comparison vs. BAU	0.10 *p*=0.19	0.49 *p*=0.29
Quant + comparison vs. Quant (no comparison)	0.10 *p*=0.19	0.16 *p*=0.87
N	2,241	120

Knowing that coaches and teachers value these feedback conversations and prefer them when the conversations are grounded in numeric observation data provides useful insights about instructional feedback and coaching. However, a central question is how feedback translates into instructional practice. Researchers will continue to examine these questions this spring and summer as end-of-year data arrives.

Note: These estimates are preliminary. As planned, we will analyze impacts on more distal survey data—alongside impacts on teachers' instructional practice and the nature of the coaching conversations—at the conclusion of the RCT in spring 2026.

rppl