Prosocial and Active Learning (PAL) Classrooms Evaluation

Final Summative Report

Ryan Williams, Bo Zhu, Max Pardo, Crystal Aguilera, Tara Zuber

October 2024

American Institutes for Research® | AIR.ORG

AIR®
Advancing Evidence.
Improving Lives.

Acknowledgments

This work was funded by a 2018 U.S. Department of Education, Office of Elementary and Secondary Education, Education Innovation and Research Early-Phase grant (U411C180114) to the eMINTS National Center and the Prosocial Development and Education Research Lab at the University of Missouri.

Contents

Exhibits

Abstract

Prosocial and Active Learning (PAL) Classrooms is a year-long teacher professional development program designed to increase students' prosocial behavior and engagement in 5th grade mathematics and science classrooms that use active, team-based lessons by altering the way that teachers interact with students using research-based strategies. The project is funded by a 6-year Education Innovation and Research grant and implemented by the eMINTS National Center and the Prosocial Development and Education Research Lab (ProsocialEd Lab) at the University of Missouri. The American Institutes for Research® (AIR®) as an independent evaluator, has completed an implementation and impact study of PAL Classrooms. The evaluation involves a mixed-methods study to assess the implementation of the program and a multicohort, school-level randomized controlled trial (RCT) to examine the impact of the program.

The evaluation took place in 41 districts in three states—Arkansas, Kansas, and Missouri. For the impact study, AIR randomly assigned two cohorts (2021–22 and 2022–23) of elementary schools to receive PAL Classrooms immediately (treatment) or conduct treatment as usual and receive PAL Classrooms a year later (control). A total of 41 schools (21 treatment and 20 control), with 65 teachers and their 1,399 students, participated in the RCT.

This final report summarizes the PAL Classrooms program and AIR's evaluation methods; findings on the extent to which PAL Classrooms' key components were implemented with fidelity; and the impact of PAL Classrooms on teacher outcomes (i.e., student-reported use of strategies, instructional quality), proximal student outcomes (i.e., prosocial behavior, engagement, perceived classroom climate, collaboration, teacher–student relationship), and student math and science achievement, .

Results of implementation analyses indicated that PAL Classrooms was generally implemented with fidelity despite occurring during the COVID-19 pandemic The program delivered key professional development activities and teachers generally participated as intended. In addition, teachers implemented strategies that promoted student prosocial behavior. Results were consistent across both cohorts.

Results of impact analyses indicated that teachers assigned to PAL Classrooms had higher levels of overall instructional quality, student engagement, emotional support, and instructional support, measured by classroom observation. However, most of the impact estimates were not statistically significant, except for the estimate of impact on emotional support (p = .05), with

an effect size of 0.58. In contrast, impacts on teachers' use of strategies such as praise or induction measured by student report were smaller, but still positive.

Results reveal that PAL Classrooms had statistically significantly positive impact on proximal student outcomes measured by self-report surveys. Students assigned to PAL Classrooms reported higher levels of prosocial behavior in their peers and teachers, with effect sizes of 0.23 and 0.20 standard deviations (*SD*), respectively, compared with students in control schools. Analyses of students' self-report of prosocial behavior, engagement, classroom climate, and student–teacher relationships also yielded positive results, although the impact estimates were not statistically significant. Impacts on student-reported collaboration and more distal student achievement in math and science were less conclusive, with the impact estimates being smaller and not statistically significant. Additionally, the impact results on student achievement did not vary by student background characteristics (i.e., free or reduced-price lunch eligibility, or receipt of academic services under an individualized education plan).

Background

Increasing students' prosocial behavior is important given the issue of worsening student behavior (National Center for Education Statistics, 2022), academic engagement (Salmela-Aro et al., 2021). Students' prosocial behavior is linked to higher achievement and engagement (e.g., Caprara et al., 2000; Galindo & Fuller, 2010; Miles & Stipek, 2006), especially among high-poverty students (Bierman et al., 2009; Griffith, 2002; Hoglund & Leadbeater, 2004; Wentzel, 1993) and is necessary for career readiness. The rising generation of professions, particularly in the science, technology, engineering, and mathematics (STEM) field, calls for greater prosocial skills such as honesty, collaborating with others, being understanding and helpful, being pleasant, and displaying a good-natured, cooperative attitude on the job (Miró-Pérez, 2020).

Problem-based learning (PBL), in which students work together in teams, provides an ideal context for promoting student prosocial behavior. PBL facilitates retention of content, critical thinking, and enhanced problem solving, and it promotes engagement and more positive attitudes (Dochy et al., 2003; Hung et al., 2008; Roseth et al., 2008).

Prosocial and Active Learning (PAL) Classrooms is a research-based teacher professional development (PD) program, funded by a 6-year Education Innovation and Research grant, to help teachers increase student prosocial behavior while engaging students in collaborative PBL with the use of technology. The program provides teachers with over 60 hours of PD and six in-class coaching visits focused on interweaving prosocial education and PBL. Teachers learn to use three strategies to increase student prosocial behavior: (a) establishing positive teacher–

student relationships, (b) praising students' prosocial behavior, and (c) using inductive rather than power-assertive discipline.

The American Institutes for Research® (AIR®), the independent evaluator of PAL Classrooms, has completed an implementation and impact study of the program. This final summative report begins by briefly describing PAL Classrooms program, followed by a summary of implementation evaluation findings. Next, the report describes impact study methods and presents impact evaluation findings. The presentation of impact evaluation findings is designed to provide all the information necessary for a What Works Clearinghouse evidence review. Last, the report concludes with our interpretation of the key findings alongside implications and study limitations.

Program Overview

The PAL Classrooms program is designed to help teachers learn strategies to promote students' prosocial behavior, which is expected to increase engagement and create a positive classroom climate. As students develop more positive relationships with each other, enjoy working as a team, and become successful problem solvers, they should learn more.

The planned program includes two key components:

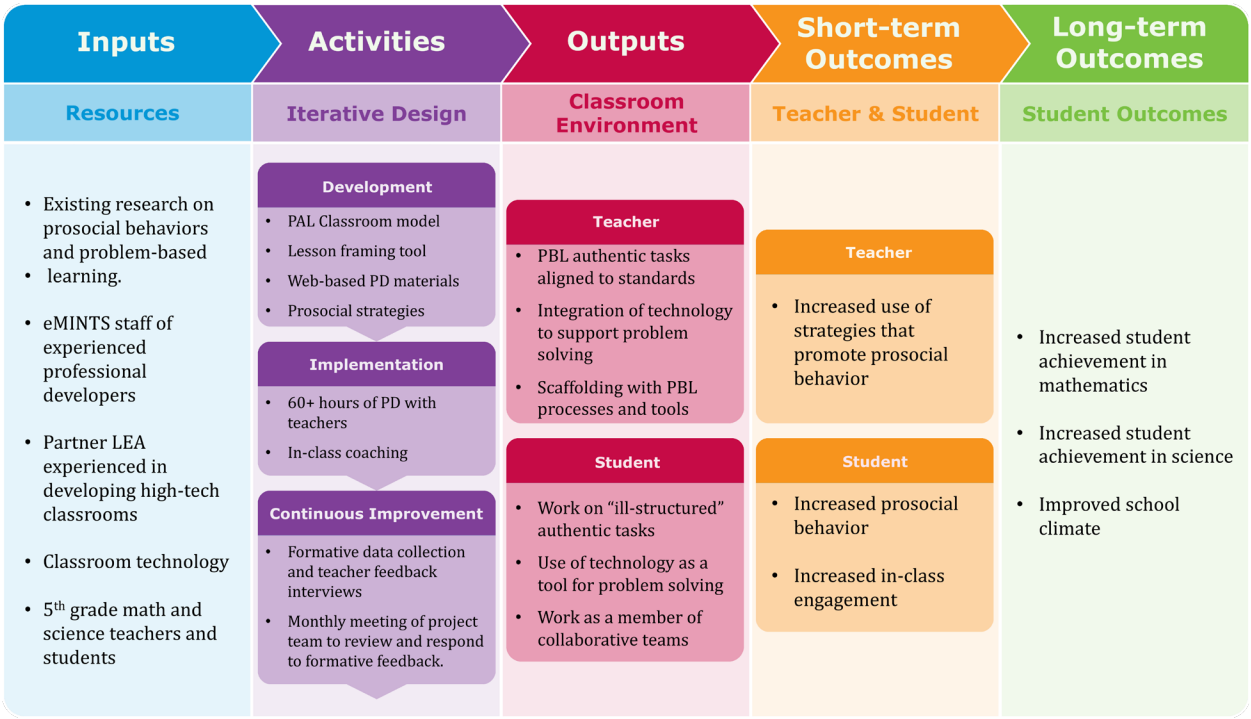
- Teachers' professional development (PD) and continuous monitoring, which includes participating in 60 hours of summer PD over 10 days, six in-class coaching visits throughout the academic year, and reports of developing project-based learning lessons. The continuous monitoring is indicated by feedback opportunities provided at the pilot sessions for programmatic improvement and the monthly virtual meetings held by the project.
- Teacher implementation (also called Teacher Prosocial Strategies), indicating teachers' use of the strategies aimed at promoting student prosocial behavior, including establishing positive teacher–student relationships, praising students' prosocial behavior, and using inductive rather than power-assertive discipline.

The PAL Classrooms program has expected short-term outcomes at the teacher/classroom and student levels. On the teacher/classroom level, short-term outcomes include: (a) implementation of problem-based learning (PBL) and prosocial behavior strategies in the classroom, and (b) effective student–teacher interactions (including instruction support, emotional support, classroom organization, and student engagement). On the student level, short-term outcomes include: (a) prosocial behavior and (b) classroom engagement.

The PAL Classrooms program has expected long-term outcomes at the teacher/classroom and student level. On the classroom level, a long-term outcome is improved classroom climate, as measured by student self-reports. On the student level, the program intends to improve academic achievement in math and science.

The logic model for the PAL Classroom is shown in Exhibit 1.

Exhibit 1. Prosocial and Active Learning (PAL) Classrooms Logic Model



Note. LEA = local education agency; PD = professional development; PBL = problem-based learning.

Setting

This multisite, cluster randomized trial (RCT), took place in 41 public elementary schools across 41 districts in three states—Arkansas, Kansas, and Missouri, across two cohorts. All participating schools were Title I schools and 90% were rural. Both Cohorts were similar in terms of their urbanicity, gender, and racial make-up. However, they differed in two ways: (1) 38% of treatment schools in Cohort 1 and 50% in Cohort 2 implemented Positive Behavioral Interventions and Supports (PBIS) at their school. (2) 61% of students from treatment schools in Cohort 1 qualified for FRPL, while 78% of treatment students in Cohort 2 qualified. Training was delivered in a standardized format across sites.

The study recruited 41 elementary schools in 41 districts a year prior to the start of implementation (2020–21 for Cohort 1 and 2021–22 for Cohort 2). Participating schools may

have had previous experience with eMINTS, but individual teachers who participated in eMINTS comprehensive professional development programming since 2016 and teachers who were part of the development phase of PAL Classrooms in years 2019–2020 and 2020–2021 were ineligible for participation in this evaluation. Most eligible schools had at least two Grade 5 teachers who were self-contained or subject-area teachers (math or science). Five schools in Cohort 1 and four in Cohort 2 only had one eligible teacher who taught both math and science. None of those teachers participated in PAL Classrooms training prior to 2020–21.

Fidelity of Implementation Study

Study Description

This section summarizes the findings pertaining to Prosocial and Active Learning (PAL) Classrooms implementation. The implementation study focused on two key components of the PAL Classrooms program: teacher professional development (PD)and teacher prosocial strategies. We developed four fidelity indicators for the PD and three indicators for the prosocial strategies component. These two components and their indicators address the core elements described in the logic model. The PD indicators focused primarily on adherence (e.g., teachers completing a certain number of hours of an activity) while the teacher prosocial strategies indicators focused on the quality of implementation.

To examine implementation, we used multiple data sources, including program records (i.e., attendance data, coaching logs, submitted lessons) and student survey. We analyzed these data descriptively to assess the extent to which program components were implemented as intended.

The implementation study assessed fidelity for both cohorts of teachers and schools in the treatment group. The study sample included nine teachers from eight schools in Cohort 1 and 21 teachers from 13 schools in cohort 2.[1] We measured and assessed fidelity separately by cohort, using the same set of indicators and thresholds to ensure comparability across cohorts.[2]

To assess fidelity, we calculated a teacher-level score for each indicator measuring the extent to which teachers engaged with or completed the corresponding activity. Then, we determined whether individual teachers received an *Adequate* rating for each program component based on the indicator-level score. Last, we determined programwide fidelity based on whether 80%

[1] One teacher from each cohort refused to participate after random assignment.
[2] In the time between the two cohorts, we updated the fidelity matrix and recalculated Cohort 1 so that the cohort scores would be comparable. Particularly, we changed the threshold for Exhibit 2, indicator 2 (coaching sessions) from four sessions to six sessions and removed two indicators (i.e., educational technology and feedback surveys) from the fidelity rubric.

or more of the study sample received the *Adequate* rating. More information on how we assessed fidelity at various levels is provided in the appendix.

Implementation Study Research Questions for the Study

The implementation study addressed five RQs focused on fidelity of implementation:

1. Do schools and teachers assigned to PAL Classrooms implement it with fidelity?
    a. Do PAL Classrooms staff deliver activities to teachers as planned?
    b. Do teachers participate in PAL Classrooms activities as intended?
    c. Do teachers incorporate PAL Classrooms strategies in the classroom as intended?
2. How do implementation and teacher experience vary across cohorts?

In the following sections, we present fidelity indicators and data sources for teacher professional development and prosocial strategies components.

Teacher Professional Development

This component includes the following four indicators:

- face-to-face PD;
- coaching visits with feedback exchange;
- PAL Classrooms book study for teachers; and
- creation of problem-based learning (PBL) lesson.

To examine fidelity of implementation, we measured the extent to which teachers engaged with or completed each of these items. For a teacher to receive an *Adequate* rating on this component, they had to meet the fidelity threshold on three of the four measures. Exhibit 2 details each indicator, including the method of measurement and the definition of fidelity.

Exhibit 2. Implementation of Teacher Professional Development (PD)

Indicator	Unit of measurement	Method of measurement	Indicator scoring at unit level	Fidelity Threshold for Adequacy
(1) Face-to-face PD	Teacher	Attendance data	Number of days that sessions were provided. Range: 0–8	1 = Teachers attended at least 7 sessions. 0 = Teachers attended fewer than 7 sessions.
(2) Coaching visits	Teacher	Coaching logs	Number of sessions attended by teacher possible. Range: 0–6	1 = Teachers received at least 4 visits. 0 = Teachers received fewer than 4 visits.

Indicator	Unit of measurement	Method of measurement	Indicator scoring at unit level	Fidelity Threshold for Adequacy
			Note. One teacher received 7 sessions.	
(3) PAL Classrooms book study for teachers	Teacher	Attendance data	Number of hours completed. Range: 0–5	1 = Teachers completed at least 5 hours. 0 = Teachers attended fewer than 5 hours.
(4) Creation of problem-based learning lesson (PBL)	Teacher	Lesson submission	Number of lessons created by the teacher. Range: 0–1	1 = The teacher created at least 1 PBL lesson. 0 = The teacher did not create a PBL lesson.
Key Component 1 Total Score **Professional Development Institute**				Percentage of teachers who scored *Adequate* on the indicators. ***Adequate* = At least 80% of teachers score *Adequate* on 3 of the 4 indicators.**

Note. After the pilot years, the program model shifted to a more informal and conversational process that did not lend itself to measurement. These informal conversations were to occur during the training sessions and coaching visits. Therefore, we did not create any fidelity indicators specifically pertaining to continuous improvement. Rather, Indicators 1 and 2 in the Exhibit 2 can be considered a proxy for continuous monitoring because those indicators concern the moments when teachers had opportunities to share feedback.

Teacher Prosocial Strategies

This component includes the following three indicators:

- teacher use of student praise;
- teacher use of induction; and
- teacher use of discipline.

To examine fidelity of implementation, we used data collected through a student survey that rated teachers' use of prosocial strategies. For a teacher to receive an *Adequate* rating on this component, they had to meet fidelity on two of the three measures. Exhibit 3 details each indicator, including the method of measurement and the definition of fidelity.

Exhibit 3. Implementation of Teacher Prosocial Strategies

Indicator	Unit of measurement	Method of measurement	Indicator scoring at unit level	Indicator scoring at sample level
(1) Teacher use of student praise	Teacher	Student survey	Scale score for praise items	1 = The student survey scale score was 2 or higher. 0 = The student survey scale score was less than 2.
(2) Teacher use of induction	Teacher	Student survey	Scale score for induction items	1 = The student survey scale score was 2 or higher. 0 = The student survey scale score was less than 2
(3) Teacher use of discipline	Teacher	Student survey	Scale score for discipline items	1 = The student survey scale score was 3 or lower. 0 = The student survey scale score was greater than 3.
Key Component 2 Total Score **Teacher Coaching**				Percentage of teachers who score *Adequate* on the indicators. ***Adequate* = At least 80% of teachers score *Adequate* on 2 of the 3 indicators.**

Fidelity Findings

In this section, we present results on fidelity of implementation for the two key program components (i.e., teacher professional development and teacher use of prosocial strategies).

Overall, the program achieved fidelity on each component for both Cohorts 1 and 2 (RQ 1). Among those with available implementation data, 89% of teachers in Cohort 1 and 95% of teachers in Cohort 2 reached the *Adequate* fidelity threshold at the component level for both the *professional development* and *use of prosocial strategies* components (Exhibit 4).

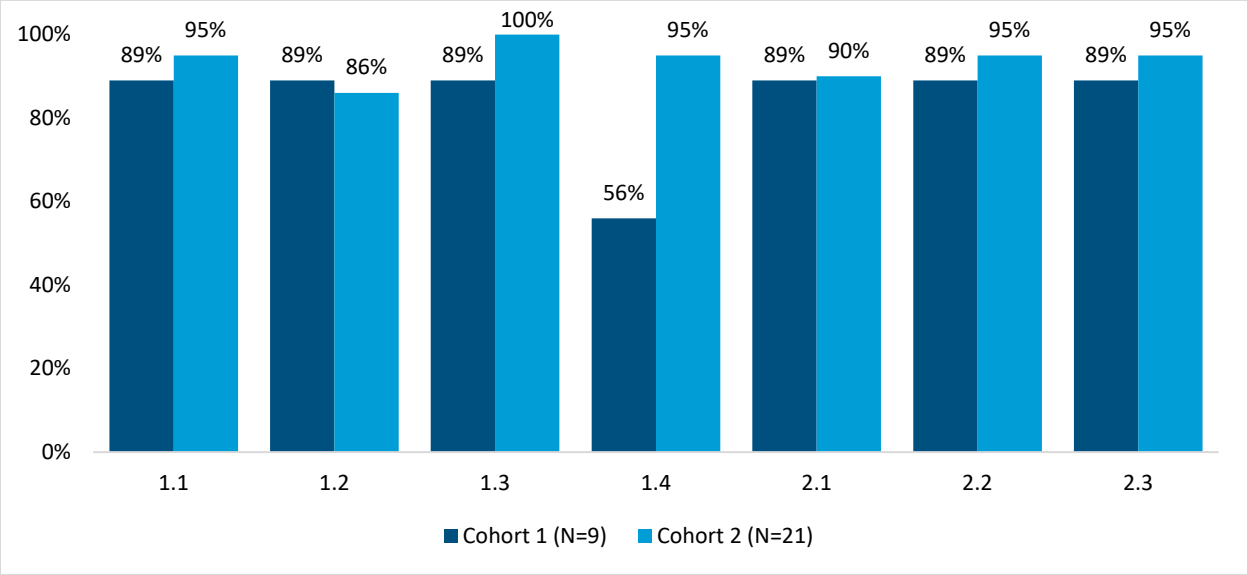
PAL Classrooms delivered key activities (RQ 1a) and teachers generally participated as intended (RQ 1b). Each year, the program hosted nine professional development sessions, six coaching visits, and 5 hours of book study. According to various program records (i.e., attendance, coaching log, lesson plans), the majority of teachers across cohorts (86% to 100%) participated in these activities. In addition, 95% of teachers in Cohort 2 created PBL lessons. However, only 56% of teachers in Cohort 1 engaged in this activity (Exhibit 5). **PBL lesson creation was the only indicator for which teachers in Cohorts 1 and 2 differed in the level of fidelity (RQ 2).** For the remaining indicators, teachers in both cohorts consistently reached the *Adequate* fidelity threshold.

Teachers in both cohorts implemented strategies that promoted student prosocial behavior (RQ 1c). According to student surveys, the majority of teachers across cohorts (89% to 95%) consistently used praise and induction practices in the classroom. On the other hand, teachers' use of discipline was also reported high, with 89% in Cohort 1 and 95% in Cohort 2 as reported by students (Exhibit 5).

Exhibit 4. Results on Fidelity of Implementation for Cohorts 1 and 2

Key components, number of indicators, units, and threshold				Cohort 1 results (2021–22 school year)			Cohort 2 results (2022–23 school Year)		
Key component	Total # of measurable indicators	Unit of implementation	Sample-level threshold for fidelity of implementation	Number of units in which component was implemented	Number of units in which fidelity of component was measured	Achieved fidelity score and whether program met sample-level threshold	Number of units in which component was implemented	Number of units in which fidelity of component was measured	Achieved fidelity score and whether program met sample-level threshold
Teacher professional development	4	Teacher	At least 80% of teachers score *Adequate* on 3 of the 4 indicators	9 teachers	9 teachers	89% of teachers; met program fidelity = yes	21 teachers	21 teachers	95% of teachers; met program fidelity = yes
Teacher prosocial strategies	3	Teacher	80% of schools with *Adequate* implementation	9 teachers	9 teachers	89% of teachers; met program fidelity = yes	21 teachers	21 teachers	95% of teacher; met program fidelity = yes

Exhibit 5. Indicator Level Fidelity for Cohorts 1 and 2



Note. Indicator 1.1: Face-to-face professional development; Indicator 1.2: Coaching visits; Indicator 1.3: PAL Classrooms book study for teachers; Indicator 1.4: Creation of problem-based learning lesson; Indicator 2.1: Teacher use of student praise; Indicator 2.2: Teacher use of induction; and Indicator 2.3: Teacher use of discipline.

Implementation Evaluation Limitations

The implementation study had limitations, particularly pertaining to the collection of classroom observation data. The Classroom Assessment Scoring System (CLASS) observations posed several complications, including video quality, the time interval between two videos, and incomplete video submission.[3]

For video quality, teachers did not always place the Swivl devices in optimal locations or did not angle their camera for a good view of both the teacher and students. Lighting conditions also complicated the observations. In a few cases, the Swivl device did not turn properly, suggesting an issue with the device, its placement, or the connection between the device and the teacher's microphone dongle. These quality issues undermined the observers' ability to score classrooms.

The second challenge is associated with the time interval between teachers' two video recordings. Teachers were expected to video record instruction at two time points; namely, when they were less familiar or confident with PAL Classrooms and after they had more experience implementing the program. Unfortunately, teachers did not always follow the guidance and properly space out the timing of their video recordings. Rather, the two time

[3] Some of these complications resulted from the technical difficulties that teachers experienced when they used the Swivl device, such as access to the Swivl app.

points overlapped significantly and, in some cases, took place only days apart.[4] As a result, we were unable to compare teachers' instructional practice between the intended two time points. Instead, we constructed a single set of data for each cohort to measure instruction overall.

Last, video submission was incomplete. Teachers were asked to submit two videos of sufficient length for coding two cycles per video, or 30 minutes in total. Only a subset of teachers submitted both videos as instructed. Some teachers did not submit videos of sufficient length. In some cases, the videos were long enough to still score two cycles, but in other cases the length of the videos was inadequate. In addition, control teachers in Cohort 2 only submitted the first round of videos while treatment teachers submitted two rounds of data. Therefore, we only used the first round of scores for Cohort 2 teachers.

Impact Study

Study Description

The impact study used a school-level randomized controlled trial (RCT) in 41 elementary schools in 41 districts across two cohorts (2021–22 and 2022–23) and three states (Arkansas, Kansas, and Missouri) to examine effects of full-scale implementation of PAL Classrooms on teacher practice and student outcomes. Each district had one school participating in the study.

AIR randomly assigned 21 schools to treatment (Prosocial and Active Learning [PAL] Classrooms) and 20 schools to a waitlist control (business as usual). The randomization used four blocks, collectively defined by cohort and locale: (a) Cohort 1 schools in Missouri, (b) Cohort 1 schools in Arkansas and Kansas, (c) Cohort 2 schools in Missouri, and (d) Cohort 2 schools in Arkansas. Teachers in schools assigned to treatment group implemented PAL Classrooms immediately (2021–22 for Cohort 1 and 2022–23 for Cohort 2) while teachers in schools assigned to control group conducted their treatment as usual without implementation of PAL Classrooms. Teachers in schools assigned to control schools implemented PAL Classrooms a year later.

The impact study examined both teacher and student outcomes. Teacher outcomes included student-reported teachers' use of strategies and instructional quality, which were measured by student surveys and classroom observations, respectively. Student outcomes included prosocial

[4] One possible reason was the delay in the first round of observation data collection. Due to the extended school recruitment timeline resulting from COVID-19, PAL Classrooms began later than expected. After the program began, teachers were given their Swivl devices to record their classroom observation data, but many experienced difficulties using them at the beginning. The facilitators then needed to dedicate time during one of the November sessions to teach teachers how to use the Swivl robots. This led to a delay in the first round of data collection.

behaviors, engagement, classroom climate, collaboration, teacher–student relationship, and academic achievement in math and science, which were measured using student surveys and extant administrative data.

Impact Research Questions for the Study

The impact study addressed the following 13 *confirmatory* research questions (RQs) focused on teacher outcomes (RQs 1–6) and student outcomes (RQs 7–13):[5]

1. What is the effect of one year's experience in PAL Classrooms on fifth-grade teachers' use of strategies that promote students' prosocial behavior compared with fifth-grade teachers' use of strategies in business-as-usual condition classrooms?
2. What is the effect of one year's experience in PAL Classrooms on fifth-grade teachers' overall instructional quality compared with fifth-grade teachers' instructional quality in business-as-usual condition classrooms?
3. What is the effect of one year's experience in PAL Classrooms on the quality of emotional support in fifth-grade mathematics and science classrooms compared with the quality of emotional support in fifth-grade mathematics and science classrooms in business-as-usual condition classrooms?
4. What is the effect of one year's experience in PAL Classrooms on the quality of classroom organization in fifth-grade mathematics and science classrooms compared with the quality of classroom organization in fifth-grade mathematics and science classrooms in business-as-usual condition classrooms?
5. What is the effect of one year's experience in PAL Classrooms on the quality of instructional support in fifth-grade mathematics and science classrooms compared with the quality of instructional support in fifth-grade mathematics and science classrooms in business-as-usual condition classrooms?
6. What is the effect of one year's experience in PAL Classrooms on the quality of student engagement in fifth-grade mathematics and science classrooms compared with the quality of student engagement of fifth-grade mathematics and science classrooms in business-as-usual condition classrooms?
7. What is the effect of one year's experience in PAL Classrooms on the prosocial behavior of fifth-grade students and teachers compared with fifth-grade students' and teachers' prosocial behavior in business-as-usual condition classrooms?

[5] RQ 7a and 8a are exploratory questions. Student-level administrative data could not be linked to student survey responses, so moderator analyses using administrative data were only conducted on the academic achievement outcomes that appeared in that data.

8. What is the effect of one year's experience in PAL Classrooms on the engagement of fifth-grade students compared with fifth-grade students' engagement in business-as-usual condition classrooms?
9. What is the effect of one year's experience in PAL Classrooms on fifth-grade classroom climate compared with fifth-grade mathematics and science classroom climate in business-as-usual condition classrooms?
10. What is the effect of one year's experience in PAL Classrooms on the collaboration of fifth-grade students compared with fifth-grade students' collaboration in business-as-usual condition classrooms?
11. What is the effect of one year's experience in PAL Classrooms on 5th grade students' perceived relationship with teachers compared with 5th grade students' perceived relationship with teachers in business-as-usual condition classrooms?
12. What is the effect of one year's experience in PAL Classrooms on the mathematics achievement of fifth-grade students compared with fifth-grade students' achievement in business-as-usual classrooms?
    a. Does the effect of PAL Classrooms on the mathematics achievement of fifth-grade students compared with a business-as-usual condition vary by student background characteristics, as measured by free or reduced-price lunch (FRPL) eligibility and students receiving services under an individualized education plan (IEP)?
13. What is the effect of one year's experience in PAL Classrooms on the science achievement of fifth-grade students compared with fifth-grade students' achievement in business-as-usual condition classrooms?
    a. Does the effect of PAL Classrooms on the science achievement of fifth-grade students compared with a business-as-usual condition vary by student background characteristics, as measured by FRPL eligibility and students receiving services under an IEP?

Intervention Condition

Two cohorts of teachers participated in the PAL Classrooms training. Training for Cohort 1 was led by a lead and assistant facilitator.[6, 7] In Cohort 2, the same two facilitators led training for a

[6] The lead facilitator is a 45-year-old female with a doctorate in educational leadership from the University of Missouri who has spent the past 13 years providing professional development to educators as an eMINTS instructional specialist and program support coordinator.
[7] The assistant facilitator is a 40-year-old woman with a master's degree in educational technology from the University of Missouri who began as an instructional specialist with eMINTS National Center in 2021. She previously worked as a public school teacher for 15 years.

group of teachers while a third new assistant facilitator joined the team.[8] Cohort 2 was divided into two groups because of its larger sample size. The first group (Cohort 2a) was based in Rolla, Missouri, and led by the lead facilitator from Cohort 1; the second group (Cohort 2b) was based in Kansas City, Missouri, and led by the assistant facilitator from Cohort 1.

Cohort 1 was held between August 2021 and April 2022, and Cohort 2 was held between September 2022 and March 2023. The PAL Classroom program consisted of nine 6-hour sessions in total; three were virtual and six were in-person. All sessions were 6 hours regardless of the format. The orientations were held in the summer before the school year began. The remaining seven sessions were held roughly every month over those 6 months and were held during designated school hours. The schedule was the same for both cohorts.

Control Condition

Teachers in the schools assigned to the control group conducted business as usual without implementation of PAL Classrooms during the year that treatment-group teachers implemented PAL Classrooms. Students in the control group were expected to receive standard fifth-grade math and science curriculum and instruction.

Teachers in the control group had an opportunity to implement PAL Classrooms a year later (2022–23 for Cohort 1 and 2023–24 for Cohort 2). Diffusion of treatment was unlikely due to the school-level randomization design.

Study Participants

After teachers agreed to participate in the study, AIR randomly assigned 21 schools (eight in Cohort 1 and 13 in Cohort 2) to the treatment condition and 20 schools (eight in Cohort 1 and 12 in Cohort 2) to the control condition. At the time of random assignment, the teacher sample included 32 treatment teachers and 33 control teachers. The analytic sample of teacher outcomes included 27 (84%) treatment teachers and 26 (81%) control teachers and did not include any teacher joiners.[9]

After random assignment, the study team obtained parent consent and student assent when students in the participating schools began fifth grade (fall 2021 for Cohort 1 and fall 2022 for Cohort 2). There were no restrictions on the eligibility of individual students in the study. Participation in the study was voluntary, and participants could withdraw at any time without penalty. The analytic sample of student outcomes included all eligible students in study schools,

[8] The new facilitator, who joined Cohort 2, is a 32-year-old woman with a Master of Science in education from the University of Missouri. She began as an eMINTS instructional specialist in 2019; she's also served as an educator for 9 years.
[9] Teacher outcome analyses had multiple analytic samples because the number of teachers available for analysis differed by outcomes. The counts reported in the narrative indicate the overall numbers of teachers, by treatment condition, available for teacher outcome analyses. The outcome-specific sample sizes can be found in Exhibit 3.

with 661 treatment students and 734 control students.[10] The analytic sample included student joiners, defined as those who joined the participating schools after random assignment. Given the nature of the PAL Classrooms program, its relatively low visibility, and the study design (i.e., school level randomization), the risk of bias due to student joiners is low (What Works Clearinghouse, 2022).

Sample Alignment With Those Served by the Program

The evaluation sample included all of the schools that were offered the intervention over the duration of the evaluation.

Design and Measures

Independence of the Impact Evaluation

AIR's impact evaluation of the PAL Classrooms was independent. Particularly, AIR independently conducted all key evaluation activities, including randomization, data analyses, and reporting of findings. AIR, eMINTS and the ProsocialEd Lab collaborated on data collection, with eMINTS emailing survey invitation links to participants, and with AIR securely collecting and storing all survey responses and administrative records.

Preregistration of the Study Design

The study was preregistered on the Registry of Efficacy and Effectiveness Studies (REES; https://sreereg.icpsr.umich.edu/sreereg/home?msg=published#registry-5360).

Design

The study used a block randomized controlled trial design, with school as the unit of assignment. AIR randomly assigned two cohorts of elementary schools within blocks to receive PAL Classrooms immediately (treatment) or conduct treatment as usual and receive PAL Classrooms a year later (control). The randomization used 1:1 assignment probability and four blocks, collectively defined by cohort and state:[11] (a) Cohort 1 schools in Missouri, (b) Cohort 1 schools in Arkansas and Kansas, (c) Cohort 2 schools in Missouri, and (d) Cohort 2 schools in Arkansas.

For Cohort 1, randomization occurred in July 2021. Schools and teachers consented to participate in the study in August 2021. The initial PD sessions began in early August, while the implementation of PAL Classrooms strategies in participating teachers' classrooms began in

[10] Student outcome analyses had multiple analytic samples because the number of students available for analysis differed by outcomes. The counts reported in the narrative indicate the overall numbers of students, by treatment condition, available for student outcome analyses. The outcome-specific sample sizes can be found in Exhibit 4.
[11] Because most of the participating schools were in small rural single-school districts, blocking on school district was not a feasible design for either cohort.

September 2021 and ended in April 2022. Parents of study students provided consent in September 2021. Baseline survey administration with teachers and students took place between September and November 2021 and the postintervention administration occurred between April and May 2022.[12] There were no systematic differences in the timing of study milestones by condition.

For Cohort 2, randomization occurred in July 2022. Schools and teachers consented to participate in the study in August 2022. The initial training sessions began in early August, while the implementation of PAL Classrooms began in September 2022 and ended in April 2023. Parents of study students provided consent in September 2022. Baseline survey administration with teachers and students took place September to December 2022 and postintervention administration occurred between April and May 2023. There were no systematic differences in the timing of study milestones by condition.

Measures

This section presents the confirmatory outcomes with their corresponding research questions (RQs), measures, and timing of data collection. We describe each individual measure used in the study; all measures were administered with consistent procedures across treatment and control conditions and time points (pre and post).

Teacher Outcome Measures

Student-Report of Teacher Prosocial Behavior Strategies. We developed a 4-item measure of *perceived teacher praise* when students behaved prosocially which was derived from research on growth mindset (Dweck, 2014). The items differentiated between person-focused praise (e.g., "Tell you that you are a nice or good person") and act-focused praise (e.g., "Tell you that specific things you do are nice"). In addition, we adapted items from the Parenting Styles and Discipline Questionnaire (PSDQ; Robinson et al., 1995) to create a 5-item *induction discipline* measure to assess students' overall rating of their teachers' use of induction (e.g., "Explain the reasons for behavior expectations before you start an activity"). For all items, students rate on a 4-point Likert-type frequency scale ranging from 1 (almost none of the times) to 4 (almost every time). The internal consistency estimates of the teacher praise and induction scales were 0.72 and 0.80, respectively, for Cohort 1; and 0.69 and 0.79, respectively, for Cohort 2.

Teacher Instructional Quality. To measure teacher instructional quality, we used the Classroom Assessment Scoring System–Upper Elementary (CLASS–UE; Pianta et al., 2012), an observation tool that evaluates student–teacher interactions in Grades 4 to 6. Using the CLASS–UE, certified observers who were blind to treatment condition rated classrooms on a scale of 1 (dimension is

[12] Students provided assent at the time of survey completion.

not present) to 7 (dimension is very present) for each dimension across multiple dimensions within each of the three domains (i.e., emotional support, classroom organization, and instructional support). Observers also rated on a fourth dimension focused on student engagement. Across dimensions, the interrater agreement within one category response option ranged from 64% to 98%, with internal consistency estimates ranging from 0.87 to 0.92 (Pianta et al., 2012).

For the three domains, we produced and used the domain-level scores by averaging scores across dimensions within each domain. In addition, we used a total score across three domains and the fourth dimension to represent overall quality of instruction.

To collect the data, participating teachers recorded their classroom using Swivl recording devices and shared the recordings for study team review and coding. Each teacher was asked to share two rounds of 30 to 40 minutes of recorded instruction for observation. That amount of time would allow the observers to conduct two cycles of observation, with each cycle involving 15 to 20 minutes of active observation. The first round of videos was recorded at the start of the spring semester (e.g., January) and the second toward the end of the semester (e.g., May). However, due to technical difficulties with recording and delays in setting up the Swivl devices, the two time points overlapped significantly and, in five cases in Cohort 1, took place on the same day. Three more teachers in Cohort 1 submitted videos less than 20 days apart and none of the Cohort 1 teachers submitted videos more than 49 days apart. In Cohort 2, four teachers submitted videos less than 50 days apart. In addition, control teachers in Cohort 2 only submitted the first round of videos while treatment teachers submitted two rounds of data. Therefore, we combined all Round 1 and Round 2 scores into a single set of data for Cohort 1 teachers and used Round 1 scores only for Cohort 2 teachers. Certified CLASS-UE observers coded the recordings and were blind to treatment status.

Student Outcome Measures

We used a suite of self-reporting measures for five outcomes: prosocial behavior, classroom engagement, classroom climate, collaboration, and teacher–student relationship.[13] To collect data, students in each cohort completed these measures via an online survey twice: once in the fall (generally between September and November) and once in the spring (between April and May). For each student measure, we use Rasch rating scale analysis (e.g., Andrich, 1978; Rasch, 1980; Wright & Masters, 1982; Wright & Stone, 1979) to generate scale scores and evaluate measurement properties (e.g., unidimensionality) using the Winsteps software program

[13] Outcome domains defined in this report are specific to this study and do not necessarily reflect the corresponding outcome domains defined in the What Works Clearinghouse Study Review Protocol.

(Linacre, 2015). Scale scores were initially created separately for baseline and outcome. Rating scale and item parameters were equated across the baseline and outcome administrations using anchoring. Below we present a description of each measure along with their reliability estimates from the current study sample.

Prosocial Behavioral Scale. We adapted the Prosocial Behavioral Scale **(**Bergin et al., 2011**)** to measure student and teacher sharing, helping, complimenting, encouraging, and cooperating. Students rated their own, their classmates', and their teachers' prosocial behavior on an 8-item scale (e.g., "In this class, my classmates comfort others [like cheer others up when they are down or talk over problems]") ranging from 1 (never*)* to 5 (every day). For the current study, the internal consistency estimates of the self-report, peer report, and perceived teacher Prosocial Behavioral Scale was 0.90, 0.88, and 0.92, respectively, for Cohort 1; and 0.89, 0.87, and 0.92, respectively, for Cohort 2.

Classroom Engagement. Student engagement was assessed using a 17-item scale adapted from the Classroom Engagement Inventory (Wang et al., 2014) comprising three subscales, including *affective engagement* (5 items; "In this class, I feel excited"), *cognitive engagement* (7 items; e.g., "In this class, I go back over things I don't understand"), and *behavioral engagement* (5 items; e.g., "In this class, I get really involved in class activities"). Students rated their experiences in a given class on a 4-point Likert-type frequency scale ranging from 1 (almost none of the times) to 4 (almost every time). For the current study, the internal consistency estimates of the affective engagement, cognitive engagement, and behavioral engagement subscales were 0.91, 0.80, and 0.79, respectively, for Cohort 1; and 0.90, 0.81, and 0.79, respectively, for Cohort 2.

Classroom Climate. We used a subset of items from the ED School Climate Surveys (EDSCLS, student form; Wang et al., 2016) to measure student perceived classroom climate. The EDSCLS is a suite of student survey measures adapted for students in different grade spans, including Grades 5 to 8. Students in the current study rated on a 7-item scale (e.g., "I feel like I belong") ranging from 1 (strongly disagree*)* to 4 (strongly agree). The internal consistency estimate was 0.88 for both cohorts.

Collaboration. Collaboration was assessed using a 4-item scale adapted from the Cooperative Skills subscale (Ladd et al., 2014). Students rated their collaboration with classmates on the items (e.g., "When you work with classmates on projects, rate how often your classmates cooperate with you") ranging from 1 (almost none of the times) to 4 (almost every time). For the current study, the internal consistency estimates of the scale were 0.79 and 0.82 for Cohort 1 and Cohort 2, respectively.

Perceived Teacher–Student Relationship. Students rated their perceived relationship with a given teacher using a 12-item scale adapted from the teacher–student relationship scale (Huang et al., 2018; Roeser et al., 1996). Items measured the students' perceptions of their teachers' caring, fairness, and general support (e.g., "My teacher treats all students respectfully"). Students rated on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). The internal consistency estimates of the scale based on the current sample were 0.93 and 0.94 for Cohort 1 and Cohort 2, respectively.

In addition to self-reporting measures, we used standardized state assessments to measure student math and science achievement.

Standardized State Assessments, Missouri Assessment Program (MAP). To measure student math and science achievement, we used standardized state assessments in mathematics and science. Specifically, we used the MAP tests, ACT Aspire, and Kansas Assessment Program (KAP) for students in Missouri, Arkansas, and Kansas, respectively. All three are end-of-year summative assessments administered to students in certain grades each spring: (a) students in Grades 5 to 8 take MAP math and science tests; (b) students in Grades 3 to 10 take ACT Aspire math and science tests; and (c) students in Grades 3 to 8 and Grade 10 take KAP math tests, and those in Grades 5, 8, and 11 take KAP science tests. We obtained MAP data from the Missouri Department of Elementary and Secondary Education and ACT Aspire and KAP data from participating schools immediately following the end of PAL Classrooms program implementation. Student test scale scores from the prior spring (e.g., spring 2021 for Cohort 1) were used as the baseline measure, and scores during the implementation year (e.g., spring 2022 for Cohort 1) were used as the outcome measure. Test scores were standardized and converted to z-scores using statewide grade-level means and standard deviations for each year.[14]

Outcome measures, including data collection methods, timing, and scoring procedures appear in Exhibit 6.

Exhibit 6. Confirmatory Outcome Measures

RQ #	Domain	Measure name	Data collection methods	Data collection timing[a]	Baseline	Scoring
Teacher outcomes						
1	Teacher prosocial behavior strategies	Study-developed measures	Student survey (aggregated)	Spring 2022 and spring 2023	Fall 2021 and fall 2022	Subscale Rasch Score

[14] The 2023 statewide standard deviations were not available for Arkansas at the time of analysis. We used statewide means and standard deviations from 2022 to standardize the scores of Cohort 2 students from Arkansas, instead.

RQ #	Domain	Measure name	Data collection methods	Data collection timing[a]	Baseline	Scoring
2	Overall instructional quality	Classroom Assessment Scoring System–Upper Elementary (CLASS–UE)	Observation protocol	Spring 2022 and spring 2023	Not collected	Total Score
3	Emotional support	CLASS–UE	Observation protocol	Spring 2022 and spring 2023	Not collected	Total Score
4	Classroom organization	CLASS–UE	Observation protocol	Spring 2022 and spring 2023	Not collected	Total Score
5	Instructional support	CLASS–UE	Observation protocol	Spring 2022 and spring 2023	Not collected	Total Score
6	Student engagement	CLASS–UE	Observation protocol	Spring 2022 and spring 2023	Not collected	Total Score
Student outcomes						
7	Prosocial behavior	Prosocial Behavior Scale	Student survey (self-report, peer report, and perceived teacher behavior)	Spring 2022 and spring 2023	Fall 2021 and fall 2022	Total Rasch Scale Score
8	Classroom engagement	Classroom Engagement Inventory	Student survey (self-report)	Spring 2022 and spring 2023	Fall 2021 and fall 2022	Subscale Rasch Score
9	Classroom climate	ED School Climate Surveys (item subset TBD)	Student survey (self-report)	Spring 2022 and spring 2023	Fall 2021 and Fall 2022	Total Rasch Scale Score
10	Collaboration	Collaboration Scale	Student survey (self-report)	Spring 2022 and spring 2023	Fall 2021 and fall 2022	Total Rasch Scale Score
11	Teacher–student relationship	Teacher–Student Relationship Scale	Student survey (self-report)	Spring 2022 and spring 2023	Fall 2021 and fall 2022	Total Rasch Scale Score
12 and 12a	Mathematical knowledge and ability	MAP Grade 5 Mathematics Assessment	Student-level extant data	Spring 2022 and spring 2023	Spring 2021 and spring 2022	Total Scale Score
13 and 13a	Science knowledge and ability	MAP Grade 5 Science Assessment	Student-level extant data	Spring 2022 and spring 2023	Spring 2021 and spring 2022	Total Scale Score

[a] The two time points presented represent the timing of measurement for Cohorts 1 and 2, respectively.

Sample Sizes and Attrition

For student-reported teacher use of strategies, we used student survey data aggregated to the teacher level. The analytic sample included 18 out of 21 treatment schools and 17 out of 20 control schools. Hence, the cluster-level attrition for all survey-based student outcome measures meets the criteria for "low" attrition under cautious assumptions (14.6% overall attrition; 0.7% differential attrition). As for teacher level, we determined the sample size at randomization based on the number of teachers in the roster at randomization who were from nonattritting schools at post-assessment. The analytic sample did not include any teacher joiners. The sample included 25 out of 32 treatment teachers and 26 out of 30 control teachers from nonattritting schools (17.7% overall attrition; 8.5% differential attrition). Teacher-level attrition meets the criteria for "low" attrition under optimistic assumptions.

For observation-based instructional quality outcomes, the analytic sample included 17 out of 21 treatment schools and 17 out of 20 schools (17.1% overall attrition; 4.0% differential attrition). The cluster-level attrition meets the criteria for "low" under cautious assumptions. Teacher-level attrition for the observation-based outcomes is high, with 25 out of 30 treatment teachers and 21 out of 30 control teachers included in the analytic sample (23.3% overall attrition; 13.3% differential attrition).

For survey-based student outcome measures (e.g., affective engagement), the analytic sample included 18 out of 21 treatment schools and 17 out of 20 control schools. Hence, the cluster-level attrition for all survey-based student outcome measures meets the criteria for "low" category under cautious assumptions (14.6% overall attrition; 0.7% differential attrition). As for student level, we determined the sample size at randomization based on the number of consented students at enrollment in nonattritting schools during the implementation year (e.g., 2021–22 for Cohort 1). The analytic samples included student joiners who were enrolled into study schools after randomization. However, the risk of bias from student joiners for the current school-level randomization is considered low (What Works Clearinghouse, 2022). Across all outcome measures, student-level attrition was low. The highest overall student attrition for any survey-based student outcome was 24.5%, with a differential attrition rate of 2.2%.

For standardized student outcome measures (i.e., math and science achievement), no attrition occurred at the cluster level for both treatment and control groups. As for student level, we determined the sample sizes at randomization based on the number of students with available premeasure scores. Across all outcome measures, student-level attrition was low. The highest

overall student attrition for any student achievement outcome measure was 0.3%, with a differential attrition rate of 0.3%, as well.

Exhibit 7 presents sample sizes at randomization and in analytic sample by treatment condition for various teacher outcomes.

Exhibit 7. Sample Sizes at Randomization and in Analytic Sample for Teacher Outcomes

Outcome measure	Control group				Treatment group			
	Schools		Teachers		Schools		Teachers	
	# Randomized	# Analytic sample	# Randomized [a]	# Analytic sample	# Randomized	# Analytic sample	# Randomized [a]	# Analytic sample
Survey-based outcomes								
Student-reported Praise	20	17	30	26	21	18	32	25
Student-reported Induction	20	17	30	26	21	18	32	25
Observation-based outcomes								
Overall instructional quality	20	17	30	21	21	17	30	25
Emotional support	20	17	30	21	21	17	30	25
Classroom organization	20	17	30	21	21	17	30	25
Instructional support	20	17	30	21	21	17	30	25
Student engagement	20	17	30	21	21	17	30	25

[a] The randomized sample consists of teachers in nonattritting schools.

Exhibit 8 presents sample sizes at randomization and in analytic sample by treatment condition for various student outcomes.

Exhibit 8. Sample Sizes at Randomization and in Analytic Sample for Student Outcomes

Outcome measure	Control group				Treatment group			
	Schools		Students		Schools		Students	
	# Randomized	# Analytic sample	# Randomized [a]	# Analytic sample	# Randomized	# Analytic sample	# Randomized [a]	# Analytic sample
Survey-based outcomes								
Prosocial behavior– Peer	20	17	440	344	21	18	477	362
Prosocial behavior–Teacher	20	17	440	343	21	18	477	362
Prosocial behavior–Self	20	17	440	343	21	18	477	362
Affective engagement	20	17	440	344	21	18	477	363
Cognitive engagement	20	17	440	344	21	18	477	363
Behavioral engagement	20	17	440	344	21	18	477	362
Classroom climate	20	17	440	342	21	18	477	361
Collaboration	20	17	440	344	21	18	477	362
Teacher–student relationship	20	17	440	337	21	18	477	355
Standardized achievement outcomes								
Math achievement	20	20	737 [b]	734	21	21	662 [a]	661
Science achievement	20	20	736 [b]	733	21	21	661 [a]	661

[a] The randomized sample consists of students enrolled in nonattritting schools.

[b] The randomized sample for achievement outcomes consists of any student with a baseline score (i.e., from the prior year) for a given outcome.

Data Analysis and Findings

Baseline Equivalence

To examine baseline equivalence, we calculated the standardized mean differences of baseline covariates, using the most inclusive sample for each outcome domain (e.g., student math achievement).

For student-reported teachers' use of strategies, treatment and control groups did not differ on the premeasure of each outcome measure. That is, the standardized mean difference effect sizes between the two groups on premeasures were lower than 0.25 standard deviation (*SD*).

The two groups, however, differed on five measures pertinent to years of teaching experience and age, with standardized mean differences on these measures greater than 0.25.

For teacher instructional quality outcomes, treatment and control teachers differed on five measures pertinent to race/ethnicity and years of teaching experience, with standardized mean differences on these measures greater than 0.25. Because we did not collect premeasures of instructional quality, we used student-reported teachers' use of strategies at baseline as proxy premeasures within the same WWC outcome domain. Results indicate that treatment and control teachers did not differ on the premeasure of student-reported teachers' use of praise. However, the two groups differed on the premeasure of student-reported teachers' use of induction, with a standardized mean difference greater than 0.25.

For the survey-based student outcomes treatment and control groups were generally equivalent on baseline characteristics with one exception: treatment and control groups were not equivalent on the school-level aggregated English language arts (ELA) measure, with a standardized mean difference of 0.34. However, the two groups did not differ on the premeasure of each survey-based outcome measure and gender. For student achievement outcomes, treatment and control groups were generally equivalent on baseline characteristics. Importantly, treatment and control groups did not differ on the premeasure of each outcome measure.

Baseline equivalence assessments are presented in Exhibits 9–12.

Exhibit 9. Baseline Equivalence for Teacher Use of Strategies Outcomes

Measure	Control group			Treatment group			Treatment – Control difference	Effect size (*d*)
	Sample size	Mean	Standard Deviation (*SD*)	Sample size	Mean	(*SD*)		
Baseline praise	26	0.07	0.29	25	0.02	0.33	−0.04	-0.14
Baseline induction	26	−0.05	0.32	25	0.05	0.45	0.09	0.24
Baseline prosocial teacher practices	26	0.01	0.37	25	0.08	0.43	0.07	0.17
Asian	26	0.04	0.20	25	0.00	0.00	−0.04	-0.27
White	26	0.81	0.40	25	0.96	0.20	0.15	0.47
Total teaching experience: 0 years	21	0.00	0.00	24	0.13	0.34	0.13	0.50
Total teaching experience: 1–3 years	21	0.29	0.46	24	0.21	0.41	−0.08	-0.17
Total teaching experience: 4–6 years	21	0.24	0.44	24	0.25	0.44	0.01	0.03
Total teaching experience: 7+ years	21	0.48	0.51	24	0.42	0.50	−0.06	-0.12
Grade 5 teaching experience: 0 years	21	0.05	0.22	24	0.21	0.41	0.16	0.47
Grade 5 teaching experience: 1–3 years	21	0.43	0.51	24	0.25	0.44	−0.18	-0.37
Grade 5 teaching experience: 4–6 years	21	0.38	0.50	24	0.33	0.48	−0.05	-0.10
Grade 5 teaching experience: 7+ years	21	0.14	0.36	24	0.21	0.41	0.07	0.17
Age: 21–30	21	0.29	0.46	24	0.29	0.46	0.01	0.01
Age: 31–40	21	0.38	0.50	24	0.33	0.48	−0.05	-0.10
Age: 41–50	21	0.24	0.44	24	0.25	0.44	0.01	0.03
Age: 51–60	21	0.05	0.22	24	0.04	0.20	−0.01	-0.03
Age: 61–70	21	0.05	0.22	24	0.08	0.28	0.04	0.14

Note. The baseline equivalence presented in the exhibit is based on the analytic samples for the student-reported praise and induction outcomes. The impacts for these outcomes were estimated on the student-level survey data using the same models as other student survey outcomes. The two outcomes shared the same teacher sample. Teaching experience and age were collected separately from teacher surveys and were not included in the student-level estimation models but they are presented here to characterize the samples. The sample size is lower for these characteristics because of higher missingness on the teacher survey; some teachers for whom we had student survey data for the outcomes did not complete the teacher surveys and thus we do not have their age and experience information. Baseline praise, induction, and prosocial teacher practices are measured using the

aggregated student-reported data on teacher practices. Teacher race/ethnicity, teaching experience, and age measure the proportion of the sample represented by a given characteristic (e.g., teachers with one to three years of total teaching experience).

Exhibit 10. Baseline Equivalence for Teacher Observation Outcomes

Measure	Control group			Treatment group			Treatment – Control difference	Effect size (*d*)
	Sample size	Mean	Standard deviation	Sample size	Mean	Standard deviation		
Baseline praise	21	0.09	0.29	25	0.02	0.33	−0.07	-0.21
Baseline induction	21	−0.06	0.32	25	0.06	0.44	0.12	0.31
Baseline prosocial teacher practices	21	−0.01	0.37	25	0.07	0.43	0.08	0.19
Asian	21	0.05	0.22	25	0.00	0.00	−0.05	-0.32
White	21	0.95	0.22	25	1.00	0.00	0.05	0.32
Total teaching experience: 0 years	21	0.00	0.00	25	0.12	0.33	0.12	0.48
Total teaching experience: 1–3 years	21	0.29	0.46	25	0.20	0.41	−0.09	-0.19
Total teaching experience: 4–6 years	21	0.24	0.44	25	0.28	0.46	0.04	0.09
Total teaching experience: 7+ years	21	0.48	0.51	25	0.40	0.50	−0.08	-0.15
Grade 5 teaching experience: 0 years	21	0.05	0.22	25	0.20	0.41	0.15	0.45
Grade 5 teaching experience: 1–3 years	21	0.43	0.51	25	0.24	0.44	−0.19	-0.39
Grade 5 teaching experience: 4–6 years	21	0.38	0.50	25	0.36	0.49	−0.02	-0.04
Grade 5 teaching experience - 7+ years	21	0.14	0.36	25	0.20	0.41	0.06	0.15
Age: 21–30	21	0.29	0.46	25	0.32	0.48	0.03	0.07
Age: 31–40	21	0.38	0.50	25	0.32	0.48	−0.06	-0.12
Age: 41–50	21	0.24	0.44	25	0.24	0.44	0.00	0.00
Age: 51–60	21	0.05	0.22	25	0.04	0.20	−0.01	-0.04

Measure	Control group			Treatment group			Treatment – Control difference	Effect size (*d*)
	Sample size	Mean	Standard deviation	Sample size	Mean	Standard deviation		
Age: 61–70	21	0.05	0.22	25	0.08	0.28	0.03	0.13

Note. Teaching experience and age were collected from teacher surveys.

Exhibit 11. Baseline Equivalence for Student Survey Outcomes

Measure	Control group			Treatment group			Treatment – Control difference	Effect size (*d*)
	Sample size	Mean	Standard deviation	Sample size	Mean	Standard deviation		
Prosocial behavior of peers baseline	344	-0.01	1.01	363	0.00	0.91	0.01	0.01
Prosocial behavior of teacher baseline	344	0.04	1.01	363	-0.03	0.89	-0.07	-0.07
Prosocial behavior of self baseline	344	0.03	1.02	363	-0.02	0.90	-0.05	-0.05
Affective engagement baseline	344	-0.08	0.83	363	0.03	1.01	0.11	0.12
Cognitive engagement baseline	344	-0.03	0.94	363	0.00	0.97	0.03	0.03
Behavioral engagement baseline	344	-0.02	0.92	363	0.00	0.98	0.01	0.01
Collaboration baseline	344	0.01	0.95	363	-0.02	0.94	-0.04	-0.04
Classroom climate baseline	344	0.01	0.96	363	-0.03	0.94	-0.03	-0.03
Teacher–student relationship baseline	344	0.00	1.00	363	-0.03	0.86	-0.03	-0.03
ELA baseline–school level	344	-0.10	0.30	363	0.00	0.29	0.10	0.34
Math baseline – school level	344	0.01	0.20	363	0.02	0.28	0.02	0.07
Female	344	0.49	0.50	363	0.46	0.50	-0.03	-0.06

Note. The baseline equivalence presented in the exhibit is based on the analytic sample for the affective engagement outcome measure, which was most inclusive across all survey-based outcomes.

Exhibit 12. Baseline Equivalence for Student Achievement Outcomes

Measure	Control group			Treatment group			Treatment – Control difference	Effect size (*d*)
	Sample size	Mean	Standard deviation	Sample size	Mean	Standard deviation		
ELA baseline	733	−0.06	1.03	660	0.05	0.88	0.10	0.11
Math baseline	734	0.01	0.94	661	0.10	0.85	0.09	0.10
Female	734	0.48	0.50	661	0.46	0.50	−0.02	-0.04
American Indian	734	0.00	0.05	661	0.00	0.05	0.00	0.01
Pacific Islander	734	0.00	0.00	661	0.00	0.05	0.00	0.08
Hispanic	734	0.02	0.15	661	0.02	0.15	0.00	0.01
White	734	0.89	0.31	661	0.95	0.21	0.06	0.23
Asian	734	0.01	0.08	661	0.00	0.05	0.00	-0.05
Black	734	0.03	0.18	661	0.00	0.04	−0.03	-0.24
Multirace	734	0.04	0.20	661	0.01	0.11	−0.03	-0.19
Economic disadvantage	734	0.60	0.49	661	0.71	0.45	0.11	0.24
Limited English proficient	734	0.18	0.38	661	0.17	0.38	0.00	-0.01
Special education	734	0.03	0.17	661	0.01	0.09	−0.02	-0.15

Note. ELA = English language arts. The baseline equivalence presented in the exhibit is based on the analytic sample for the math outcome measure.

Representativeness of Individuals in Clusters

Because the study did not establish baseline equivalence of teachers in the analytic sample for the observation-based outcomes because of lack of baseline data, we examined representativeness of teachers in schools at post-assessment (Exhibit 13). Results indicate that teachers in the analytic sample were representative of schools at post-assessment under optimistic assumptions only, with 21 out of 26 teachers in control schools and 25 out of 30 teachers in treatment schools contributing to posttest means.

Exhibit 13. Postintervention Cluster Sample Sizes and Enrollment for Teacher Observation Outcomes

Outcome measure	# Schools in analysis	Control group		Treatment group	
		# Individuals contributing to posttest mean	# Individuals enrolled in schools	# Individuals contributing to posttest mean	# Individuals enrolled in schools
Overall Instructional Quality	34	21	26	25	30
Emotional Support	34	21	26	25	30
Classroom Organization	34	21	26	25	30
Instructional Support	34	21	26	25	30
Student Engagement	34	21	26	25	30

Results

In this section, we summarize impact analysis results for student and teacher outcomes. These results are organized by research questions (RQs). A detailed description of our analytic approach to estimating program impacts is provided in the appendix.

PAL Classroom had positive but not statistically significant impacts on teachers' use of prosocial strategies (RQ 1). Teachers in schools assigned to PAL Classrooms more frequently used prosocial strategies, according to survey-based responses from their students, including praise (d = 0.06) and induction (d = 0.17), compared with teachers in control schools. These results, however, were not statistically significant (Exhibit 14).

Exhibit 14. Impact Analysis Results for Student-Reported Teacher Use of Prosocial Strategies

Outcome measure	Control group				Treatment group				Treatment–Control difference	Effect size (d)	p-value
	Sample size		Model-adj. mean	Standard deviation	Sample size		Model-adj. mean	Standard deviation			
	# Clusters	# Students			# Clusters	# Students					
Praise	17	345	−0.05	1.13	18	364	0.00	0.63	0.05	0.06	0.30
Induction	17	344	−0.12	0.87	18	363	0.04	1.02	0.16	0.17	0.15

Note. Student survey outcome models included covariates for school-level standardized prior English language arts and math scores (prior-year science scores were unavailable in nearly all schools), cohort, randomization block, baseline survey impact measure, student gender, and an indicator for whether a student completed their outcome survey with a different teacher than their baseline survey.

PAL Classrooms had mostly positive but not statistically significant impacts on teachers' instructional quality (RQs 2–6). Overall, PAL Classrooms had a positive impact on teachers' instruction, including overall instructional quality (*d* = 0.29), student engagement (*d* = 0.36), emotional support (*d* = 0.58), and instructional support (*d* = 0.30). With the exception of emotional support, which was marginally statistically significant (*p* = .05), no other impact estimates were statistically significant. Additionally, teachers in schools assigned to PAL Classrooms demonstrated lower levels of overall classroom organization (*d* = -0.23) compared with teachers in control schools, which was also not statistically significant (Exhibit 15).

Exhibit 15. Impact Analysis Results for Observation-Based Teacher Instructional Quality Outcomes

Outcome measure	Control Group				Treatment Group				Treatment–Control difference	Effect size (*d*)	*p*-value
	Sample size		Model-adj. mean	Standard deviation	Sample Size		Model-adj. mean	Standard deviation			
	# Clusters	# Teachers			# Clusters	# Teachers					
Overall Instructional Quality	17	21	0.58	1.75	17	25	1.13	1.97	0.55	0.29	0.15
Student Engagement	17	21	0.73	2.04	17	25	1.52	2.33	0.79	0.36	0.11
Emotional Support	17	21	0.82	1.91	17	25	1.96	2.01	1.14	0.58*	0.05
Classroom Organization	17	21	0.33	1.59	17	25	−0.12	2.30	−0.45	−0.23	0.52
Instructional Support	17	21	0.46	2.40	17	25	1.17	2.38	0.71	0.30	0.52

* $p < 0.10$.

Note. Observation-based teacher outcome models included covariates for total years teaching overall, total years teaching Grade 5, age, race, teacher-level baseline student-reported praise and induction practices, cohort, and randomization block.

PAL Classrooms had positive impacts on students' prosocial behavior (RQ 7). Results indicate that students in schools assigned to PAL Classrooms perceived higher levels of prosocial behavior in themselves (*d* = 0.19), in their peers (*d* = 0.23), and in their teachers (*d* = 0.20) compared with students in control schools. The impact estimates for students' perceptions of prosocial behaviors in their teachers and their peers were statistically significant (Exhibit 16).

PAL Classroom had positive but not statistically significant impacts on student engagement (RQ 8). Results indicate that students in schools assigned to PAL Classrooms demonstrated higher levels of affective (*d* = 0.17), cognitive (*d* = 0.06), and behavioral (*d* = 0.13) engagement compared with students in control schools. However, none of these impact estimates were statistically significant.

PAL Classroom had a positive but not statistically significant impact on classroom climate (RQ 9). Students in schools assigned to PAL Classrooms indicated more positive classroom climates (*d* = 0.18) compared with students in control schools; however, the impact estimate was not statistically significant.

PAL Classroom had a slightly negative but not statistically significant impact on student collaboration (RQ 10). Results indicate that students in PAL Classrooms perceived slightly lower levels of collaboration with other students in their classrooms (*d* = -0.07) compared with students in control schools; however, this impact estimate was not statistically significant.

PAL Classroom had a positive but not statistically significant impact on student–teacher relationships (RQ 11). Students in schools assigned to PAL Classrooms indicated more positive relationships with their teachers (*d* = 0.22) compared with students in control schools; however, the impact estimate was not statistically significant.

Exhibit 16. Impact Analysis Results for Student Engagement, Student Prosocial Behavior, Student–Teacher Relationships, and Classroom Climate

Outcome measure	Control group				Treatment group				Treatment–Control difference	E Effect size (*d*)	*p*-value
	Sample size		Model-adj. mean	Standard deviation	Sample size		Model-adj. mean	Standard deviation			
	# Clusters	# Students			# Clusters	# Students					
Prosocial behavior–Peer	17	344	−0.16	0.75	18	362	0.05	0.99	0.20	0.23**	0.03
Prosocial behavior–Teacher	17	343	−0.15	0.78	18	362	0.02	0.93	0.18	0.21**	0.04
Prosocial behavior–Self	17	343	−0.14	0.74	18	362	0.02	0.99	0.16	0.19	0.15
Affective engagement	17	344	−0.11	0.85	18	363	0.05	1.00	0.16	0.17	0.17
Cognitive engagement	17	344	−0.05	0.81	18	363	0.00	1.03	0.05	0.06	0.81

Outcome measure	Control group				Treatment group				Treatment–Control difference	E Effect size (*d*)	*p*-value
	Sample size		Model-adj. mean	Standard deviation	Sample size		Model-adj. mean	Standard deviation			
	# Clusters	# Students			# Clusters	# Students					
Behavioral engagement	17	344	−0.09	0.78	18	362	0.03	1.03	0.12	0.13	0.27
Classroom climate	17	342	−0.15	0.77	18	361	0.01	0.97	0.16	0.18	0.09
Collaboration	17	344	−0.03	0.84	18	362	−0.09	0.96	−0.07	-0.07	0.59
Teacher–student relationship	17	337	−0.18	0.80	18	355	0.01	0.96	0.19	0.22	0.09

** $p < 0.05$.

Note. Student survey outcome models included covariates for school-level standardized prior English language arts and math scores (prior-year science scores were unavailable in nearly all schools), cohort, randomization block, baseline survey impact measure, student gender, and an indicator for whether a student completed their outcome survey with a different teacher than their baseline survey.

PAL Classrooms did not have a clear impact on student achievement in math or science (RQs 12–13). Estimated impacts on math and science achievement were relatively small and not statistically significant. Students in schools assigned to PAL Classrooms had slightly higher levels of math achievement (*d* = 0.08) and slightly lower levels of science achievement (*d* = -0.06) after participating in the program compared with students in control schools (Exhibit 17).

Exhibit 17. Impact Analysis Results for Student Achievement Outcomes

	Sample size		Model-adj. mean	Standard deviation	Sample size		Model-adj. mean	Standard deviation	Treatment–control difference	Effect size (*d*)	*p*-value
	# Clusters	# Students			# Clusters	# Students					
Science	20	733	−0.02	0.92	21	661	0.06	0.88	0.07	0.08	0.81
Math	20	734	−0.01	0.84	21	661	−0.06	0.85	−0.05	−0.06	0.48

Note. Student achievement outcome models included covariates for student gender, race, free or reduced-price lunch eligibility, Individualized Education Program status, limited English proficiency/English language learner status, prior-year standardized math score at the student and school levels, prior-year standardized English language arts score at the student and school levels (included in science achievement model only), cohort, randomization block, and school locale.

PAL Classrooms' impact on math and science achievement did not vary by student characteristics (RQ 12a–RQ 13a). In addition to estimating overall impacts on math and science achievement, we explored differential impacts for students eligible for free or reduced-price lunch (FRPL) and students receiving services through an Individualized Education Program (IEP). Our analyses did not find any statistically significant differential treatment effects across subgroups, indicating that impacts on math and science achievement were largely consistent across students eligible for FRPL or receiving services under an IEP (Exhibit 18).

Exhibit 18. Estimates of Moderating Effects of Student Free or Reduced-Price Lunch Status on Student Achievement Outcomes

Outcome	Estimate	*N*
Students eligible for free or reduced-price lunch		
Math achievement	0.06	1,395
	(0.07)	
Science achievement	0.03	1,394
	(0.08)	
Students receiving services under an Individualized Education Plan		
Mathematics achievement	−0.03	1,395
	(0.07)	
Science achievement	−0.13	1,394
	(0.08)	

Note. Standard errors are presented in parentheses. Student achievement outcome models included covariates for student gender, race, free or reduced-price lunch eligibility, Individualized Education Program status, limited English proficiency/English language learner status, prior-year standardized math score at the student and school levels, prior-year standardized English language arts score at the student and school levels (included in science achievement model only), cohort, randomization block, and school locale.

Impact Evaluation Limitations

The impact evaluation had some limitations. One limitation was tied to the administrative records that we requested and used for the evaluation. Our team was unable to use direct student and teacher links across primary and administrative data sources, which limited our ability to adjust for broader student- and teacher-level characteristics in our analyses, and it limited our ability to investigate other sources of impact heterogeneity across other student group characteristics and outcomes. We were, however, able to link primary and administrative data sources at the school level, allowing us to adjust for school compositional characteristics in our analyses.

Another limitation was the lack of a direct baseline measure for the student science achievement outcome and teacher observation-based outcomes. The participating states only administer grade-level science assessments in Grades 5 and 8, so baseline science scores were unavailable for the majority of the sample. However, we were able to include two other baseline standardized scores—math and English language arts—in the science outcome models. Likewise, we did not have baseline data for the teacher observation outcomes (see limitations of the fidelity of implementation study, below). Instead, we aggregated baseline student-reported teacher praise and induction practices from the student survey to the teacher level and included these as covariates in the observation-based teacher instructional quality models.

Finally, student survey data collection and cleaning posed some challenges for analysis. Students were identified in their baseline and outcome surveys by entering their names in an open response question. In cleaning the data, we first attempted to link baseline with outcome survey data using student name and then link survey data to the fall student rosters to identify the intention-to-treat (ITT) sample. Matching on student names using an open response question proved challenging, as students may have changed the spelling of their names, entered nicknames or pseudonyms, or not fully completed the name entry. It is possible that some ITT students were dropped from the analysis sample, either because we were unable to match their baseline and outcome survey responses (making it appear as though they had missing baseline or outcome data) or because we were unable to match their survey data to the school roster (making it appear as though they were a joiner and not an ITT student). Some students additionally may have taken the student survey with multiple teachers. In cleaning the survey data, we kept only student survey responses from the same teacher in baseline and outcome. However, occasionally the only baseline survey data we had for a student was from a different teacher than in outcome. Rather than excluding these students from analysis, we included an indicator for students with a different teacher in baseline and outcome in the impact analyses.

Conclusion

PAL Classrooms was designed as a professional development program for teachers that promotes prosocial education strategies to create positive classroom learning environments. Teachers learn to use techniques such as praise and induction, in place of traditional disciplinary practices, in the classroom.

The results of the implementation evaluation indicate that the key program components were implemented with fidelity, including implementation of professional development activities and teachers' use of prosocial strategies in the classroom. These results provide early evidence of

implementation feasibility and sustainability of the PAL Classrooms program in rural, low-income 5th grade math and science classrooms.

The results of the impact evaluation provide promising early evidence of PAL Classrooms' potential to positively effect proximal teacher and student outcomes. For teachers, the program had the strongest impact on aspects of instructional quality that are most closely related to the PAL Classrooms program, including emotional support, student engagement, and instructional support, as measured by the CLASS-UE. Impact on student-report of teachers' use of prosocial education strategies (i.e., praise and induction) were smaller but still positive.

For students, the program had the strongest impact on prosocial behavior in the classroom, the key outcome of interest. Students in school assigned to PAL Classrooms tended to indicate that they, their peers, and their teachers more frequently used prosocial behaviors in the classroom environment compared to students in control schools. Similarly, students in schools assigned to PAL Classrooms tended to indicate stronger relationships with their teachers, slightly more engagement (cognitive, affective, and behavioral), and more positive classroom climate compared to students in control schools, although these effects did not reach conventional levels of statistical significance

The impact evaluation was not able to identify clear effect on more distal student achievement outcomes. Students in schools assigned to PAL Classrooms had similar levels of overall academic achievement in math and science compared to students in control schools. Additionally, these impact results did not vary by student background characteristics (i.e., FRPL eligibility, or receipt of academic services under an IEP).

Implications

The PAL Classrooms program has early evidence of promise for increasing students' prosocial behavior, and a positive classroom environment. The results of this evaluation provide evidence of sustainability and effective transfer of key concepts and strategies.

For this evaluation, PAL Classrooms was implemented in largely remote and rural elementary schools. Future directions for program development and evaluation should consider how and where to expand implementation and scale the program sustainably. For example, the program may have benefits to students in other grade levels, locales, or states. In considering how and where to expand the PAL Classroom's program to other grade levels or settings, the development team should consider how best to scaffold teachers' implementation of the program to ensure consistent use of tools and strategies.

The PAL Classrooms development team should also further consider program modifications that would more directly demonstrate impacts on student learning outcomes. Although clear,

positive, impacts on math and science achievement were not identified in this evaluation, future avenues of inquiry may consider more proximal learning outcomes including course grades or tests of specific skills. Additionally, outcomes such as academic, commitment, persistence, and interest in schooling would provide a more robust understanding of the plausible connections across proximal and distal outcomes.

References

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43,* 561–573. https://doi.org/10.1007/BF02293814

Bergin, C., Wang, Z., & Bergin, D. A. (2011, April). *Prosocial behavior in fourth to twelfth grade classrooms* [Paper presentation.] Biennial Meeting for the Society for Research in Child Development, Montreal, Quebec, Canada.

Bierman, K. L., Torres, M. M., Domitrovich, C. E., Welsh, J. A., & Gest, S. D. (2009). Behavioral and cognitive readiness for school: Cross-domain associations for children attending Head Start. *Social Development*, *18*(2), 305–323.

Caprara, G. V., Barbaranelli, C., Pastorelli, C., Bandura, A., & Zimbardo, P. G. (2000). Prosocial foundations of children's academic achievement. *Psychological Science*, *11*(4), 302–306.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction*, *13*(5), 533–568.

Dweck, C. (2014). How can you develop a growth mindset about teaching?. *Educational Horizons*, 93(2), 15-15.

Galindo, C., & Fuller, B. (2010). The social competence of Latino kindergartners and growth in mathematical understanding. *Developmental Psychology*, *46*(3), 579–592.

Griffith, J. (2002). A multilevel analysis of the relation of school learning and social environments to minority achievement in public elementary schools. *Elementary School Journal*, *102*(5), 349–366.

Hoglund, W. L., & Leadbeater, B. J. (2004). The effects of family, school, and classroom ecologies on changes in children's social competence and emotional and behavioral problems in first grade. *Developmental Psychology*, *40*(4), 533–544.

Hung, W., Jonassen, D. H., & Liu, R. (2008). Problem-based learning. In J.M. Spector, J. G. van Merriënboer, M.D., Merrill, & M. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 485–506). Routledge.

Huang, F. L., Lewis, C., Cohen, D. R., Prewett, S., & Herman, K. (2018). Bullying involvement, teacher-student relationships, and psychosocial outcomes. *School psychology quarterly: the official journal of the Division of School Psychology, American Psychological Association*, *33*(2), 223–234.

Kate Whiting. (2020) These are the top 10 job skills of tomorrow – and how long it takes to learn them. World Economic Forum. https://www.weforum.org/agenda/2020/10/top-10-work-skills-of-tomorrow-how-long-it-takes-to-learn-them/

Ladd, G. W., Kochenderfer-Ladd, B., Visconti, K. J., Ettekal, I., Sechler, C. M., & Cortes, K. I. (2014). Grade-school children's social collaborative skills: Links with partner preference and achievement. *American Educational Research Journal*, *51*(1), 152–183.

Linacre, J. M. (2015). Winsteps® (Version 3.91.0) [Computer software]. Winsteps.com.

Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development*, *77*(1), 103–117.

Miró-Pérez, A. P. (2020). World Economic Forum: present and future. *Dimensión empresarial*, *18*(2), 1–7.

National Center for Education Statistics. (2022). More than 80 percent of U.S. public schools report pandemic has negatively impacted student behavior and socio-emotional development. Retrieved from https://nces.ed.gov/whatsnew/press_releases/07_06_2022.asp

Pianta, R.C., Hamre, B. K., Mintz, S. (2012) Classroom Assessment Scoring System–Upper Elementary and Secondary Manual. v 1.2.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Exp. ed.). University of Chicago Press.

Robinson, C. C., Mandleco, B., Olsen, S. F., & Hart, C. H. (1995). Authoritative, authoritarian, and permissive parenting practices: Development of a new measure. *Psychological Reports, 77,* 819–830.

Roeser, R. W., Midgley, C., & Urdan, T. C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology*, 88, 408 – 422.

Roseth, C. J., Johnson, D. W., & Johnson, R. T. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin*, *134*(2), 223–246.

Salmela-Aro, K., Upadyaya, K., Vinni-Laakso, J., & Hietajärvi, L. (2021). Adolescents' longitudinal school engagement and burnout before and during COVID-19: The role of socio-emotional skills. *Journal of Research on Adolescence*, *31*(3), 796–807.

Wang, Z., Bergin, C., & Bergin, D. A. (2014). Measuring engagement in fourth to twelfth grade classrooms: The Classroom Engagement Inventory. *School Psychology Quarterly*, *29*(4), 517–535.

Wang, Y., Murphy, K., & Kantaparn, C. (2016). *Technical and administration user guide for the ED School Climate Surveys (EDSCLS).* Center on Safe Supportive Learning Environments. https://safesupportivelearning.ed.gov/edscls

Wentzel, K. R. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology*, *85*(2), 357–364.

What Works Clearinghouse. (2022). *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Appendix. Description of Analytic Approaches

Approach to Estimating Program Effects

To address the confirmatory research questions, we estimated the intention-to-treat (ITT) impact on student and teacher outcomes. For each outcome, we conducted a complete case analysis and used an analytic sample including participants with complete data on outcomes and all covariates, including the baseline covariate. In the first sections, we first describe the analytic model then present analytic results. We organize the results by outcome domain (i.e., student achievement, student survey-based outcomes, student-reported teachers' use of strategies, and observation-based teacher instructional quality).

To estimate the impact of Prosocial and Active Learning (PAL) Classrooms on student achievement, student survey-based outcomes, and student-reported teachers' use of strategies, we fit the following two-level model that accounted for the nesting of students within schools:

$$Y_{ik} = \beta_{0k} + \beta_1 PAL_k + \boldsymbol{\beta}_2 \boldsymbol{X}_{ik} + \boldsymbol{\beta}_3 \boldsymbol{Z}_k + \boldsymbol{\beta}_4 \boldsymbol{W}_k + \varepsilon_{ik} + \upsilon_k,$$

where

Y_{ik} is an outcome measure for student i in school k;

β_{0k} is the intercept (for school k), or average outcome for the control school students in the reference level of student and school covariates;

β_1 is the average difference between PAL school outcomes and control school outcomes;

PAL_k is a school-level indicator for treatment status (*1* for treatment schools, *0* otherwise);

$\boldsymbol{\beta}_2$ is a vector of student covariate effects, representing their relationship with outcomes;

$\boldsymbol{X}_{ik}$ is a vector of student characteristics, including premeasures (i.e., prior-year scores for achievement outcomes and baseline survey measure for survey-based outcomes) and demographic characteristics (i.e., gender, race/ethnicity, eligibility for free or reduced-price lunch, English language learner status, individualized education program status);[15]

$\boldsymbol{\beta}_3$ is a vector of school covariate effects, representing their relationship with outcomes;

[15] For survey-based outcomes, we included only gender due to data unavailability for other demographic covariates.

$\boldsymbol{Z}_k$ is a vector of school characteristics, including school-level prior academic achievement in English language arts and math;

$\boldsymbol{\beta}_4$ is a vector of state and cohort block covariate effects, representing their relationship with outcomes;

$\boldsymbol{W}_k$ is a vector of state and cohort block indicators;

and ε_{ik} and υ_k are the student- and school-level random error terms, respectively.

To estimate the impact of PAL Classrooms on observation-based teacher instructional quality, we fit the following two-level model that accounted for the nesting of teachers within schools:

$$Y_{ik} = \beta_{0k} + \beta_1 PAL_k + \boldsymbol{\beta}_2 \boldsymbol{X}_{ik} + \boldsymbol{\beta}_3 \boldsymbol{Z}_k + \varepsilon_{ik} + \upsilon_k,$$

where

Y_{ik} is an outcome measure for teacher i in school k;

β_{0k} is the intercept (for school k), or average outcome for the control school teachers in the reference level of teacher and school covariates;

β_1 is the average difference between PAL school outcomes and control school outcomes;

PAL_k is a school-level indicator for treatment status (1 for treatment schools, 0 otherwise);

$\boldsymbol{\beta}_2$ is a vector of teacher covariate effects, representing their relationship with outcomes;

$\boldsymbol{X}_{ik}$ is a vector of teacher characteristics, including years of experience teaching overall, years of experience teaching Grade 5, race, age, and prior year prosocial teaching practices;[16]

$\boldsymbol{\beta}_3$ is a vector of state and cohort block covariate effects, representing their relationship with outcomes;

$\boldsymbol{Z}_k$ is a vector of state and cohort block indicators;

and ε_{ik} and υ_k are the student- and school-level random error terms, respectively.

[16] Prior-year teaching practices included teachers' practices related to induction and praise. Practices were reported by students in the student survey and aggregated to the teacher level by creating mean scores for each teacher.

Approach to Assessing Fidelity of Implementation at Various Levels

The implementation study focused on two key components of the PAL Classrooms program: teacher professional development (PD), and teacher prosocial strategies. We developed four fidelity indicators for the PD component and three indicators for the prosocial strategies component. These two components and their indicators address the core elements described in the logic model. The PD indicators focused primarily on adherence (e.g., teachers completing a certain number of hours of an activity), while the teacher prosocial education strategies indicators focused on the quality of implementation.

For each indicator, we used a binary measure to indicate whether the corresponding aspect of the program was implemented with fidelity (1 = yes; 0 = no). The unit of measurement was teachers for all indicators. Then, for each teacher, we determined whether the teacher met the *Adequate* threshold for fidelity at the component level by calculating the number of indicators with fidelity within the component. Particularly, a teacher would receive an *Adequate* rating for the PD component if the teacher had at least three indicators reaching fidelity. Similarly, a teacher would receive an *Adequate* rating for the prosocial strategies component if the teacher met at least two indicators with fidelity.

To calculate program-wide fidelity, we determined whether the study sample met the threshold for *Adequate* fidelity for each component by calculating the proportion of teachers who reached fidelity at the component level. If 80% or more of the teacher sample met the fidelity threshold for a component, we determined that the program component was implemented with fidelity at the sample level.

About the American Institutes for Research®

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit institution that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. AIR's work is driven by its mission to generate and use rigorous evidence that contributes to a better, more equitable world. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit AIR.ORG.

AIR® Headquarters
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | **AIR.ORG**

25214_07/24