Evaluation of the Effects of Combining the Cultivating Awareness and Resilience in Education (CARE) and PATHS Programs on 1st–4th-Grade Teacher and Student Outcomes

A Quasi-Experimental Evaluation during the COVID-19 Pandemic

Peter D. Rehder[1], Yibing Li[1], Tara Hofkens[2], & Karime Cameron[2]

[1]*American Institutes for Research*

[2]*Rand Corporation*

September 2025

AIR®
Advancing Evidence.
Improving Lives.

Contents

Acknowledgements 1

Abstract 2

Introduction 3
- Promoting Alternative THinking Strategies (PATHS) 5
- Cultivating Awareness and Resilience in Education (CARE) 5
- Current Study 7

Evaluation Design and Methods 7
- Quasi-Experimental Design 9
- Participants 10
- Procedure 12
- Measures 14
- Analysis Plan 21

Implementation Evaluation Results 24
- CARE Program Adherence and Quality 24
- Educator Attendance, Perceived Usefulness, Satisfaction, and Engagement 25
- Uptake of the CARE program 26

Impact Evaluation Results 27
- Teacher Descriptive Statistics 27
- Teacher Main Results at Post-Test and Follow-Up 28
- Student Descriptive Statistics 32
- Student Main Results at Post-Test and Follow-Up 33

Discussion 37
- Implementation Quality 38
- Effects of CARE on Teacher Outcomes 38
- Effects of CARE on Student Outcomes 39
- Study Limitations 40
- Conclusion 41

References 42

Appendix A. Supplemental Exhibits 49

Exhibits

Exhibit 1. Logic Model 8

Exhibit 2. Overview of CARE Program Adherence 24

Exhibit 3. Facilitation Quality 24

Exhibit 4. Overview of CARE attendance 25

Exhibit 5. Usefulness of CARE 25

Exhibit 6. Satisfaction with CARE 26

Exhibit 7. Engagement in CARE Trainings 26

Exhibit 8. Frequency of CARE practice 26

Exhibit 9. Benefits of CARE practices 27

Exhibit 10. Teacher Survey Outcome Results 28

Exhibit 11. CLASS Observation Results 31

Exhibit 12. PATHS Classroom Observation Results 31

Exhibit 13. Student Outcome Main Effect Model Results 33

Exhibit 14. SY2021-22 Number of Disciplinary Incidents by Classroom Proportion of FRPL Students 35

Exhibit 15. SY2021-22 Number of Disciplinary Incidents by Classroom Proportion of Students with LEP 36

Exhibit 16. SY2021-22 IAR Mathematics Scores by Classroom Proportion of Students with LEP 37

Exhibit A1. Teacher Baseline Equivalence Comparisons 49

Exhibit A2. Student Baseline Equivalence Comparisons 50

Exhibit A3. Teacher Outcome Group Means, Standard Deviations, and Sample Sizes Across Timepoints 51

Exhibit A4. Student Outcome Group Means, Standard Deviations, and Sample Sizes Across Timepoints 56

Exhibit A5. Teacher Survey Outcome Bivariate Correlations 61

Exhibit A6. CLASS Outcome Bivariate Correlations 67
Exhibit A7. PATHS Observation Outcome Bivariate Correlations 68
Exhibit A8. Student Model Bivariate Correlations 69
Exhibit A9. Student Continuous Outcome Model Fit Statistics 74
Exhibit A10. Student Categorical Outcome Model Fit Statistics 75

Acknowledgements

The authors would like to thank all those whose effort allowed us to conduct this evaluation. We are particularly grateful for their support, given the very challenging conditions of implementing and evaluating programs in schools during the COVID-19 pandemic. We thank Chicago Public Schools Office of Social and Emotional Learning, led by Adam King, and the district staff and school staff who supported program implementation and data collection; the teachers and school leaders who participated in multiple aspects of the evaluation; and the parents who consented to their children's participation in the evaluation. We also thank our partners at the University of Virginia, led by Principal Investigator Patricia Jennings, and CREATE for Education and SEL Worldwide, led by Mark Greenberg and Dorothy Morelli, for their tireless work on program implementation and thought partnership throughout the entirety of the project. We are also grateful to current and former AIR staff who led and contributed to this evaluation, especially Kimberly Kendziora and Tara Zuber. Finally, we thank the U.S. Department of Education's EIR grant program through the Office of Elementary and Secondary Education for its financial support of the project (grant number: U411C190159).

Abstract

Teachers' social-emotional skills are required not only to deliver social-emotional learning (SEL) curricula for students but also to model social competence and engage in classroom processes that support effective pedagogy and student learning. The purpose of this study was to evaluate the potential impacts of supplementing an evidence-based SEL program, Promoting Alternative THinking Strategies (PATHS), with a teacher wellness program, Cultivating Awareness and Resilience in Education (CARE) on grade 1–4 teachers' SEL and well-being, their implementation of the PATHS Curriculum, and students' outcomes. To do this, we randomly assigned 34 elementary schools in Chicago Public Schools (CPS) to receive either (1) the CARE and PATHS programs together (CARE+PATHS group) or (2) the PATHS program without CARE (PATH Only group). We assess CARE implementation and program impact by collecting surveys of teachers, observations of CARE trainings and classrooms, and district administrative data. Although the onset of the COVID-19 pandemic forced a change to virtual program implementation and a change from a randomized controlled trial to a quasi-experimental design, results indicated high implementation fidelity. We found limited effects of CARE on teachers' social-emotional competence and well-being. We found that teachers in the CARE+PATHS group scored more poorly on some aspects of mindfulness and had more depressive symptoms than teachers in the PATHS Only group. However, teachers in the CARE+PATHS group showed greater regard for student perspectives than the PATHS Only group, as measured by independent classroom observations. Further, students in the CARE+PATHS group had better attendance and fewer disciplinary incidents than students in the PATHS Only group. Our findings must be interpreted with caution, but they provide encouraging preliminary evidence of CARE's positive effects on key student outcomes.

Introduction

During the 2023-24 school year, more than 26 million students from over 56,000 public schools benefited from Title I programs nationwide (James, 2025). Title I, the largest elementary and secondary education program within the U.S. Department of Education, finances additional academic support and learning opportunities that are often required to help disadvantaged students and high needs schools. Although high-needs schools have experienced years of intensive school improvement planning, many have not shown sufficient progress on student outcomes. Achievement gaps between disadvantaged students and their peers are large and have existed for decades. One critical factor contributing to these gaps is a lack of support for students' development of cognitive and non-cognitive skills that are essential for academic success. Social-emotional competencies, such as persistence, motivation, and engagement, play a foundational role in learning and long-term academic success (Casillas et al., 2012; Denham et al., 2010; Scales et al., 2006; Weare & Nind, 2011).

No meaningful or sustainable gains in achievement can be achieved until students are fully engaged and connected to school. Students must be taught foundational skills that support productive engagement in their schoolwork and healthy relationships with teachers and peers. Neuroscientific evidences suggests that SEL programs may improve students' executive function by building greater cognitive-affective regulation in the prefrontal cortex (Blum et al., 2004). Students who become more self-aware and confident about their learning abilities try harder in school (Duckworth & Seligman, 2005). Students who set high academic goals, have self-discipline, motivate themselves, manage their stress, and organize their approach to work learn more and get better grades (Elliot & Dweck, 2005). Likewise, students who use problem-solving skills to overcome obstacles and make responsible decisions about studying and completing homework do better academically (Riggs et al., 2006).

Interpersonal, instructional, and environmental factors can also support student achievement by creating positive classrooms and school climates that promote student SEL, engagement, and learning. These supports include high expectations and support for academic success; caring teacher–student relationships; commitment to school and peers; and teaching that creates safe and orderly environments (Jennings et al., 2017; Hamre & Pianta, 2006; Hawkins et al., 2004; Catalano et al., 2002). It is likely that a combination of improvements in student SEL skills, the school environment, teacher practices and expectations, and teacher–student relationships can contribute to students' behavior change and academic performance (Schaps et al., 2004; Kusche et al., 2011). A meta-analytic review of 213 studies (270,034 students) examined the effects of SEL programs across diverse student outcomes and concluded that students exposed to an SEL intervention demonstrated enhanced motivation, positive social

behaviors, fewer conduct problems, and less emotional distress (effect sizes 0.22–0.57; Durlak et al., 2011). Academic performance was significantly improved, with mean effect sizes for test scores of 0.27, which translated to an 11 percentile point improvement on standardized achievement scores. A more recent meta-analysis of follow-up studies of 82 SEL interventions (97,000 students) six months to 18 years after implementation found the benefits of SEL programs to be durable over time and across diverse samples (Taylor et al., 2017). A separate study showed an excellent return on investment for six evidence-based SEL programs (five of which were elementary) with an $11 return for every dollar invested (Belfield et al., 2015).

High-needs schools are often embedded in neighborhoods that are affected by poverty, crime, and trauma, creating a high level of chronic stress (Hackman et al., 2010). Teacher stress and attrition, particularly in high-poverty schools, negatively affect student behavior and achievement (Alliance for Excellent Education, 2014; Greenberg et al., 2016; Hoglund et al., 2015). Teachers experiencing high levels of stress and frustration may transmit these feelings directly to students via "stress-contagion" (Wethington, 2000; Milkie & Warner, 2011; Oberle & Schonert-Reichl, 2016). The prosocial classroom theoretical model suggests that teachers' social-emotional competencies support their ability to cope with the demands of teaching and prevent burnout (Jennings & Greenberg, 2009). When teachers lack the social-emotional competences required to manage the demands of teaching, their well-being erodes and leads to a deterioration of the classroom climate, triggering a "burnout cascade" (Jennings & Greenberg, 2009, p. 492). In contrast, teachers with higher levels of these competences are able to better cope with the demands of the classroom, maintain a positive classroom climate, build and maintain supportive relationships with their students, and establish consistent classroom interactions that promote student learning. Programming that supports teachers' social-emotional skills may improve the implementation quality and effectiveness of SEL programs, especially in high-needs schools.

In the current study, we used a quasi-experimental approach to evaluate the effect of supplementing an evidence-based SEL program, Promoting Alternative THinking Strategies (PATHS; Kusché & Greenberg, 2012), with a teacher wellness program, Cultivating Awareness and Resilience in Education (CARE; Jennings, 2016). PATHS has been empirically demonstrated to promote social-emotional skills, positive peer relationships, engagement, and learning (Durlak et al, 2011) while CARE has been shown to improve teachers' social-emotional skills and classroom interactions (Jennings et al., 2017) and student engagement, motivation, and reading competence (Brown et al., 2025). In this quasi-experimental trial, we compared the outcomes of teachers and students who were randomly assigned to receive both the CARE and PATHS programs to those who were randomly assigned to receive only the PATHS program.

Promoting Alternative THinking Strategies (PATHS)

PATHS is one of the most extensively researched and effectively replicated classroom-based SEL programs available to schools (Greenberg et al., 1998). Numerous federal agencies and registries have identified PATHS as an effective program, and it is recognized as a proven program by the stringent criteria of the Blueprints for Healthy Youth Development. PATHS provides teachers at each grade level with a scope and sequence of lessons that directly teach SEL skills. It also includes daily practices to promote skill generalization and teaching strategies to create a positive classroom climate and help teachers integrate the curriculum with other areas of instruction (e.g., language arts, social studies, and science). Results from efficacy trials indicate that PATHS leads to significant improvements in children's emotional understanding, self-control, frustration tolerance, and social problem-solving skills (Riggs et al., 2006; Greenberg et al., 1995; Bierman et al., 2008a; Bierman et al., 2008b). PATHS also shows effects on cognitive outcomes, with several studies showing significant improvement in executive function, inhibitory control, working memory, efficiency of problem-solving, and planning ability compared to controls. The skills children gain from PATHS facilitate positive coping, classroom behavior, and academic engagement (Bierman et al., 2008a; Bierman et al., 2008b; Conduct Problems Prevention Research Group, 1999; Greenberg et al., 1995).

Cultivating Awareness and Resilience in Education (CARE)

CARE is a comprehensive, manualized professional learning program designed to promote and support teachers' SEL and well-being via training in: (a) Emotion Skills Instruction, (b) Mindfulness/ Stress Management, and (c) Caring & Listening Practices. The original design of CARE involved 30 hours of group instruction, presented as a series of five 6-hour sessions over the course of a school year. In response to reductions in time for teachers' professional development, the CARE model was reduced to two full days delivered in-person and one day delivered virtually or in person. Often there is a one-day booster session in the spring. Following best practices in adult learning, CARE introduces material sequentially, using a blend of didactic, experiential and interactive learning processes (Jennings, 2016).

Two randomized controlled trials (RCTs) evaluating CARE have found that CARE improved teachers' general well-being, self-efficacy, time-related stress, personal accomplishment, stress-related physical symptoms, emotion regulation, mindfulness, and psychological distress compared to comparison teachers who did not receive CARE (effect sizes = .32–.80; Jennings, 2016; Jennings et al., 2017). CARE also improved teachers' emotional support of students, positive classroom climate, productivity, and teacher sensitivity in independent classroom observations (effect sizes = .22–.23; Jennings et al., 2017); and teacher-reported student engagement in learning, motivation, and reading competence (effect sizes = .07–.12; Brown et al., 2025).

Emotion Skills Instruction

CARE combines didactic instruction with experiential activities to help teachers understand, recognize, and regulate emotional responses in the classroom. Teachers learn the brain science associated with the stress response, which prepares them to teach this content to their students. Reflective practices and role-plays support teachers' recognition of emotional states and their exploration of their habitual emotional patterns, tendencies, and reactivity profile.

Mindfulness/Stress Reduction Practices

CARE introduces mindfulness practices beginning with short periods of silent reflection, body awareness, or breath awareness and extend to activities that bring mindfulness into aspects of daily living such as eating, walking, and listening to others. Through these activities, teachers bring greater awareness into the classroom and modulate their physical/emotional state thus improving their relationships with students, classroom management, and instructional practices.

Caring and Listening Practices

To promote empathy and compassion, CARE introduces "caring practice" and "mindful listening." Caring practice involves a silent reflection during which one generates feelings of kindness and compassion by mentally offering well-being, happiness, and peace—first to oneself, then to a loved one, then to a neutral colleague or acquaintance, and finally to a person who one finds challenging.

CARE Delivery

For CATALYZE the original plan was to start with two 6-hour days of training in the summer, followed by one 6-hour day of training in the fall, another 6-hour training in the winter, and ending with a 6-hour booster session in the spring, for a total of 30 hours of training. We also planned for teachers to have 30-minute coaching calls with CARE facilitators between CARE trainings.

Due to the onset of the COVID-19 pandemic when our study was set to begin, we adapted the CARE program to be implemented in a virtual setting by reducing the number of hours and omitting or adapting some content and activities. We reduced the total training time to 12 hours, providing teachers with two days of training split into two 2-hour sessions (4 hours/day) in August, two hours in November, and two hours in January. We did a thorough review and removed activities and practices that could not be done well in a virtual format (e.g., mindful walking and mindful eating) and modified some activities that required interpersonal interactions (e.g., mindful listening and role play). We also incorporated several elements into the virtual format to improve engagement, including breakout rooms and polls, and using the chat to support participation and communication. Finally, in lieu of coaching calls, CARE

facilitators provided weekly email/text reminders to support teachers in applying CARE skills, emotion skills, and mindfulness practices to their teaching.

Current Study

The current study aimed to evaluate the potential impacts of CARE on grade 1–4 teachers' SEL and well-being, their implementation of the PATHS Curriculum, and students' outcomes. To do this, we randomly assigned 34 elementary schools in Chicago Public Schools (CPS) to receive either (1) the CARE and PATHS programs together (CARE+PATHS group) or (2) the PATHS program without CARE (PATH Only group). The CARE and PATHS programs were delivered by researchers from the University of Virginia (UVA) in partnership with SEL Worldwide (PATHS) and CREATE for Education (CARE). They conducted the implementation evaluation while the American Institutes for Research (AIR) conducted an independent impact evaluation. AIR independently conducted random assignments of schools, collected all impact data, and conducted all impact analyses. As shown in the logic model in Exhibit 1, we hypothesized that receiving the CARE program would improve (1) teachers' mindfulness, emotion regulation, psychological distress, and time-related stress; (2) the implementation quality of the PATHS curriculum, (3) and the quality of classroom environments and teacher–relationships compared to PATHS Only teachers. We also hypothesized that students in the CARE+PATHS group would show (4) improvements in SEL, engagement, motivation, behavior, and academic achievement compared to students in the PATHS Only group. Our hypotheses were pre-registered in the Registry of Efficacy and Effectiveness Studies. To our knowledge, this is the first study to examine whether providing a teacher wellness program can catalyze the effects of a student-focused SEL program to improve both teacher and student outcomes.

Evaluation Design and Methods

Our original plan for the evaluation of implementation and impact was to conduct a cluster RCT. We recruited 34 elementary schools from CPS selected by the district based on school need and readiness for intervention. We oversampled schools that were serving high proportions of high-needs students. We used a one-to-one matching procedure in which schools were paired based on school demographics, including their proportion of students eligible for free or reduced-price lunch and their proportion of students identifying as either Black or African American or Hispanic or Latino/a/e/x. After matching schools into pairs, we randomly assigned one school from each pair to the control group, which received training for their teachers to administer the PATHS program for students (i.e., the PATHS Only group), and one school to the intervention group, which received PATHS training and the CARE program for their teachers (i.e., the CARE+PATHS group).

Exhibit 1. Logic Model

Key Components (Inputs)	Direct Components (Outputs)	Outcomes		
		Proximal		Distal
CARE Workshops Emotion skills instruction Mindful awareness and stress reduction practices Caring and listening practices Workbook Phone coaching	Teachers use CARE strategies for their own well-being	Increased adaptive emotion regulation and mindfulness	Improved student social-emotional skills	Improved student academic achievement
	Teachers use CARE strategies when interacting with students	Decreased psychological distress and time-related stress	Improved student behavior	Improved student attendance and reduced absenteeism
PATHS SEL curriculum Summer training day Fall training day Classroom kit	Teachers implement PATHS program in their classrooms	Improved teacher-student interaction quality	Improved student engagement in learning	Reduced student disciplinary incidents
			Improved student motivation to learn	

For the implementation evaluation, the UVA team planned to assess the quality of CREATE's implementation of CARE within the CARE+PATHS group. For the impact evaluation, the AIR team planned to conduct intent-to-treat (ITT) analyses comparing teacher and student outcomes between the CARE+PATHS and PATHS Only groups across a one-year implementation period and a one-year follow up during school years 2021–22 and 2022–23. However, the integrity of our cluster RCT design was significantly hampered by the COVID-19 pandemic.

In CPS, schools experienced multiple lockdowns intended to prevent the spread of COVID in schools, as well as closures resulting from teacher strikes protesting the return to in-person instruction without proper COVID safety protocols in place (Tareen, 2022). Like school districts around the country, CPS struggled with teacher hiring and retention, and many staff had to take on additional responsibilities to cover basic school functions, resulting in less capacity to participate in new programs and research (Lieberman, 2022; Will, 2022). For our study, the effects of COVID on schools resulted in significantly reduced participation in the CARE program and in the evaluation. Only 64.5% of teachers from schools in the CARE+PATHS group attended Days 1 and 2 of the CARE training. Only 70% of teachers completed the baseline teacher well-being survey, and only 32.2% completed the baseline teacher survey of student well-being. Likewise, only 21.5% of parents gave consent for their child's well-being to be rated by teachers. These low participation rates combined to create conditions in which an unacceptable amount of missing data would threaten the internal validity of the school-level RCT as a test of CARE+PATHS. Therefore, we made the decision to change the design to a teacher-level quasi-experimental design (QED) that kept the random assignment of schools intact but matched and compared teachers who received sufficient dosage of the CARE program in the CARE+PATHS group to demographically similar teachers from the PATHS Only group.

Quasi-Experimental Design

Days 1 and 2 of the CARE are where most of the program content is provided for teachers. Therefore, the UVA implementation team determined that teachers who fully attended Days 1 and 2 of CARE had received an adequate dosage of CARE to have potentially observable benefits for them and their students that could be compared to teachers and students in the PATHS Only group as part of the impact evaluation. Seventy-eight 1st–4th-grade teachers in the CARE+PATHS group met the CARE attendance criteria. We conducted propensity score matching to draw a sample of 78 teachers from the PATHS Only group. The independent variables used in the propensity score matching model included teachers' baseline mean mindfulness, psychological distress, time urgency, and mindful practices (i.e., counterfactual) scores; their years of teaching experience, education level, classroom student-teacher ratio, grade(s) taught, gender, and race/ethnicity; and their PATHS training attendance rate.

The propensity score matching procedure successfully resulted in a matched sample of N = 156 teachers (78 CARE+PATHS, 78 PATHS Only) from 32 schools. For analyses of student outcomes that used district administrative records, we had permission to access deidentified records for all students in participating teachers' classrooms, resulting in a sample of N = 3,866 students in the matched sample (2,024 CARE+PATHS, 1,842 PATHS Only). For analyses of student outcomes that used teacher reports, active parental consent was required, resulting in a sample of N = 292 students (134 CARE+PATHS, 158 PATHS Only). For all the teacher and student outcomes, the sample demonstrated baseline equivalence (see Exhibits A1 and A2 in Appendix A).[1]

For the implementation evaluation, we did not limit analyses to the matched sample of 156 teachers. Where applicable, we included data for all teachers who attended CARE training, which included teachers in the original study sample and additional teachers and clinicians in CARE+PATHS schools. Due to the COVID-related stress that school staff were experiencing at the time of implementation, CPS asked the implementation team to offer CARE to Kindergarten through 8th grade teachers and staff at the participating CARE+PATHS schools, including special education teachers, dual language teachers, and school clinicians. Of the 644 eligible educators invited to participate (121 original study sample, 523 additional educators), 102 did not participate in any of the CARE training sessions, resulting in a sample of N = 542 teachers served attending at least one CARE training session (84.2%). Although these additional teachers and staff did not meet inclusion criteria for the impact evaluation, we included them in the implementation evaluation to provide a complete picture of CARE implementation quality. When applicable, we also provide information that is specific to the impact evaluation sample.

Participants

Teacher Sample

The implementation sample included 542 educators who attended the CARE trainings. 85.1% of these educators were women and 14.9% were men. 38.9% were White, 36.2% were Black or African American, 19.5% were Hispanic or Latino/a/e/x, 2.5% were Asian or Filipino, 1.7% were some other race or ethnicity, and 1.2% were more than one race or ethnicity. 2.5% taught prekindergarten, 15.4% taught kindergarten, 14.1% taught 1st grade, 13.6% taught 2nd grade, 15.3% taught 3rd grade, 16.1% taught 4th grade, 16.8% taught 5th grade, 11.9% taught 6th grade, 12.1% taught 7th grade, 12.2% taught 8th grade, and 30% of teachers taught multiple grades. 65.3% of educators were regular teachers, 24.7% were special educators, and 6% were

[1] We conducted baseline equivalence analyses for impact outcome using a model similar to the main impact model without covariates included. We conducted independent samples T-tests for teacher outcomes and means-as-outcomes multilevel models with intervention group as the only independent variable for student outcomes.

supportive staff members (e.g., social workers, clinicians, school counselors, case manager) or "Other" (Resource, Residents, Clerk, Librarian, School administrator, SEL leader, or Specials teacher).

The impact evaluation sample was comprised of the 156 matched teachers. 76.8% of teachers were women and 23.2% were men. 36.2% of teachers were Black or African American, 35.4% were White, 23.1% were Hispanic or Latino/a/e/x, 1.5% were Asian or Filipino, 0.8% were American Indian or Native American, 1.5% were more than one race or ethnicity, and 1.5% were some other race or ethnicity. 19.5% of teachers had a bachelor's degree, 7.1% had some graduate education, and 73.5% of teachers held a Master's/Specialist degree or higher. 31.3% of teachers taught 1st grade, 25.8% taught 2nd grade, 16.4% taught 3rd grade, 18.8% taught 4th grade, and 7.8% taught more than one grade. Teachers had an average of 15.77 years of teaching experience (SD = 9.66), and their classrooms had an average student–teacher ratio of 11.36 (SD = 4.96). There were no significant demographic differences between the CARE+PATHS and PATHS Only groups. However, teachers in the CARE+PATHS group had 2.36 more students per teacher in their classrooms than teachers in the PATHS Only group, $T(122) = 2.72$, $SE = .87$, Hedges' $g = .49$, $p = .008$. Of these 156 teachers, N = 64 (32 CARE+PATHS, 32 PATHS Only) completed the post-test teacher well-being survey in spring 2022, N = 48 (24 CARE+PATHS, 24 PATHS Only) completed the follow-up teacher well-being survey in spring 2023, N = 88 (46 CARE+PATHS, 42 PATHS Only) participated in an observation classroom environment quality, and N = 65 (37 CARE+PATHS, 28 PATHS Only) participated in a classroom observation of their PATHS implementation quality.

Student Sample

The full student sample included N = 3,866 students, all those from classrooms of teachers in the propensity-matched sample. 50.2% of students were girls and 49.8% were boys. 52.9% of students were Black or African American, 39.0% were Hispanic or Latino/a/e/x, 5.9% were White, 1.0% were more than one race or ethnicity, 0.8% were Asian, 0.2% were American Indian or Native American, and 0.1% were Hawaiian or Pacific Islander. 26.5% of students were 1st graders, 29.3% were 2nd graders, 18.7% were 3rd graders, and 25.6% were 4th graders. 81.4% of students qualified for free or reduced-price lunch (FRPL), 12.4% were special education students, and 15.1% of students had limited English proficiency (LEP). Students in the PATHS Only group were more likely to be Black or African American than students in the CARE+PATHS group, $t(3414.20) = -3.51$, $SE = .02$, Hedges' $g = -.12$, $p < .001$. Students in the CARE+PATHS group were more likely to be White than students in the PATHS Only group, $t(3335.84) = 4.21$, $SE = .01$, Hedges' $g = .14$, $p < .001$.

The teacher-report subsample of students included N = 292 students who had parental consent for teachers to complete the surveys. Among this subsample, 51.9% of students were girls and

48.1% were boys. 42.8% of students were Black or African American, 48.6% were Hispanic or Latino/a/e/x, 5.1% were White, 1.0% were more than one race or ethnicity, 1.3% were Asian, 0.3% were American Indian or Native American, and 0.3% Hawaiian or Pacific Islander. 25.0% of students were 1st graders, 34.6% were 2nd graders, 18.2% were 3rd graders, and 22.3% were 4th graders. 85.6% of students qualified for free or reduced-price lunch (FRPL), 10.3% were special education students, and 27.7% of students had limited English proficiency (LEP). Students in the CARE+PATHS group were more likely to be Black or African American, $t(277.80) = 2.78$, $SE = .06$, Hedges' $g = .33$, $p = .006$, and to be eligible for FRPL, $t(279.20) = 2.88$, $SE = .04$, Hedges' $g = .33$, $p = .004$, than students in the PATHS Only group. Students in the PATHS Only group were more likely to be White than students in the CARE+PATHS group, $t(253.34) = -2.17$, $SE = .03$, Hedges' $g = -.24$, $p = .031$. Of these 292 students, N = 237 (116 CARE+PATHS, 121 PATHS Only) were rated by their teachers at post-test in spring 2022 and N = 63 (40 CARE+PATHS, 23 PATHS Only) were rated by their teachers at follow-up in spring 2023.

Procedure

Implementation Evaluation Procedure

The implementation evaluation examined observed program adherence and facilitation quality, and multiple measures of educator-reported responsiveness and uptake. Where applicable, we incorporated measures that were developed in previous trials of CARE for evaluating implementation (see Doyle et al., 2019). We also revised existing measures and introduced several additional or alternative measures based on the modifications made in response to COVID. Since we had to provide CARE trainings virtually, we video-recorded all trainings in order to rate facilitators' program adherence and trained research assistants (RAs) to rate facilitation quality after each training. We also took attendance and asked teachers to complete self-report surveys after each CARE training segment (i.e., at the end of Day 2, Day 3, and the booster) to assess participant responsiveness to and uptake of the CARE program.

CARE Trainings and Recordings. A certified CARE facilitator provided each training on a UVA Zoom account. CARE trainings took place in 2-hour segments. Days 1 and 2 in August consisted of two 2-hour segments: an a.m. and p.m. session separated by a one-hour break for lunch. Day 3 and the Booster consisted of one 2-hour training each. In total, CREATE offered 64 CARE trainings. Each treatment school had their own training for Days 1 and 2, after which CREATE combined some schools for Day 3 and the Booster.

Trained RAs from UVA provided technical assistance to CARE facilitators and participants. The RAs recorded the trainings on the Zoom platform and saved them on a secure server for later coding of facilitation adherence. We trained the RAs who provided technical support to code the recordings for adherence. The RAs completed at least four hours of coding training, during

which they viewed a recording of a CARE training provided by an expert facilitator. The video served as a benchmark for exemplary facilitation of the online version of the CARE program. We also discussed examples of non-exemplary facilitation behaviors for contrast. Coders also received and reviewed a CARE coding manual that we revised from a previous efficacy trial of the CARE program to align with virtual delivery of CARE. Coders practiced coding recordings of CARE trainings from the pilot before coding for the full study. To minimize coder bias, we randomly assigned the trained RAs to rate videos of CARE trainings that they did not facilitate.

Educator Self-Report Post-Session Surveys. At the end of each CARE session, educators completed an online self-report survey about their satisfaction with, engagement with, and perceived usefulness of the CARE program, as well as their use and benefits of CARE practices (after the Booster only). We reserved the last 5-10 minutes of each training for educators to complete surveys online. The project manager emailed the survey link to educators who attended the training but did not complete the survey during the session.

Impact Evaluation Procedure

Our impact evaluation compared teacher and student outcomes of the randomly assigned, propensity-matched groups to each other before and after PATHS training was provided to both groups and CARE training was provided to the CARE+PATHS group. We collected baseline data before program implementation, post-test data at the end of the implementation school year, and follow-up data at the end of the next school year.

Baseline Data Collection. In spring 2021, we asked teachers from participating schools to complete an online baseline teacher well-being survey. The survey assessed several aspects of teacher well-being, including mindfulness, adaptive emotion regulation, psychological distress, and time urgency. It also included questions about teacher demographic information. At the same time point, we also accessed student administrative records for the 2020–21 school year (SY2020–21). Records included attendance, disciplinary incidents, English language arts (ELA) and mathematics standardized test scores, and student demographic information.

In fall 2021, we asked teachers to complete an online baseline student well-being survey. We randomly selected eight students from each teacher's classroom and asked teachers to report on each of them. The survey assessed teachers' perceptions of student engagement, motivation, reading and writing competence, mathematics competence, SEL, and behavior problems. We also planned to conduct baseline observations of classroom environments and teacher–student relationships in fall 2021, but we were unable to do so because of COVID-related school closures and visiting restrictions.

Post-Test and Follow-Up Data Collection. In spring 2022 and spring 2023, we again asked teachers to complete surveys of their own well-being and their students' well-being, and we

accessed SY2021–22 and SY2022–23 student administrative records. For the spring 2023 student well-being survey, we asked their teacher for the new academic year to complete the survey. All measures collected from each of these sources were the same as those collected at baseline. In spring 2022, we also conducted observations of classroom environment quality using the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008). We conducted separate observations of teachers' PATHS implementation quality. Due to COVID-related school closures, we could not conduct baseline classroom observations. We did ask teachers to report on student well-being in spring 2023, but only N = 63 students were rated by their teachers. As a result, we did not have enough statistical power to examine teacher-reported student outcomes at follow-up, and we do not report them here.

Measures

Implementation Evaluation Measures

Program Adherence. We measured CARE facilitator program adherence with observer-rated measures of the extent to which the facilitator (1) completed CARE activities and (2) met the intended learning objectives. Each segment of the CARE training includes multiple activities that introduce concepts and/or provide an opportunity for educators to practice a CARE skill. CARE coders rated whether each activity was completed (1) or not completed (0) by the facilitator. We then computed the percentage of activities each facilitator completed. We considered facilitators to have adequately completed program activities if they met a predetermined fidelity threshold of at least 80% of activities completed.

The activities for the CARE training are designed and facilitated in such a way as to meet specific learning objectives for each segment. Coders rated the extent to which each learning objective was met on a 0–4 scale (0 = activity/skill did not meet any objectives to 4 = all objectives were met at an exemplary level). We computed a mean score with higher scores indicating that learning objectives were met. We considered facilitators to have adequately completed the learning objectives if they met a predetermined fidelity threshold of an average rating of at least 3.0.

We double-coded 42% of the CARE trainings to prevent rater drift and to assess inter-rater reliability. We resolved "significant disagreements" through consensus. A disagreement was considered "significant" if coders disagreed about whether an activity was completed or not, or the ratings of the extent to which learning objectives were met was off by more than 1 point. "Minor disagreements" referred to ratings that were off by 1 point, which we resolved by creating an average rating of the two scores. We calculated that inter-rater reliability via percent agreement between coders. Across all double-coded cases, 90% of the codes had no significant discrepancies and 82% were in perfect agreement.

Facilitation Quality. Coders rated the quality of program delivery by CARE facilitators with a form used in previous CARE studies. The form, based on the Iowa Strengthening Families Facilitator Rating Form (Iowa State University Extension and Outreach, 2010), assesses positive (clarity when explaining concepts and activities, acceptance and friendliness, level of preparation, comfort level with material, pace of delivery, ability to engage participants) and negative facilitation skills (losing track of time, ineffective answers to questions, being critical of participants, acting distant, providing therapeutic advice/problem solving to individual participants during session). In addition, given the changes to CARE due to the pandemic, we added a question about the quality of the zoom platform rated on a 1-5 scale. We rated facilitation skills on a 0-4 scale (e.g., 0 = Facilitator performed very few or none of the behavioral indicators to 4 = Facilitator exemplified all behavioral indicators at all or almost all opportunities; reversed for negative facilitation behaviors).

Educator Attendance. We took attendance at CARE sessions by having RAs take a screenshot of the educators who were in the session thirty minutes after it started and thirty minutes before it ended. Educators who were present for both screenshots were considered to have attended the session.

Perceived Usefulness of CARE. We asked educators to rate their agreement with the following statement to assess the usefulness of the CARE material covered each session day: "Overall, I feel like I will use the concepts and activities taught today in my life and my work as a teacher." Educators responded on a 1 to 5 scale (1 = Strongly Disagree to 5 = Strongly Agree).

Satisfaction with CARE. We measured satisfaction with the CARE training by asking educators to rate how satisfied they were with the program content, the skill of the facilitators, the length of the program, the setting/atmosphere, the program design, and the communication they received in advance. We also asked educators to rate their overall satisfaction with the CARE training. Educators responded on a 5-point scale (1 = highly unsatisfied to 5 = highly satisfied). We computed a mean score for all 8 items, with higher scores indicating greater satisfaction. Internal reliability for the overall means ranged from $\alpha = 0.93–0.98$.

Engagement with CARE. After each training, we asked educators to report on their engagement with the following item: "Overall, how engaged were you in the training today?" Educators responded on a 4-point scale (1 = not very engaged in most session activities/discussions to 4 = highly engaged in all session activities/discussions).

Frequency of CARE practice. We assessed the frequency with which educators used CARE practices at the end of the Booster training in January by asking "On average, how often have you used each one of these practices so far this year?": taking three deep breaths, breath awareness practice, caring practice, body scan, setting an intention, mindful listening,

experiencing joy, recognizing scripts, and meditation. Teachers responded on a 6-point scale (1 = have not practiced; 2 = less than 1 time a week, 3 = 1 times a week, 4 = 2-3 times a week, 5 = 4-6 times a week, 6 = 1 time per day or more). We computed a mean score for all 9 items, with higher scores indicating higher frequency of CARE practice. Internal reliability was $\alpha = 0.86$.

Perceived Benefits of CARE practices. We asked educators to rate the extent to which they agreed that as a result of participating in the CARE training, they are: more self-aware, more committed to staying in their profession, better able to take control of their time, have an improved sense of overall well-being, feel better able to establish and maintain supportive relationships with their students, feel better able to manage classroom behaviors effectively and compassionately, feel better able to promote awareness and concentration among their students, and feel less stress associated with their job. We also asked if they agreed that this type of training should be integrated into preparation and in-service training for all teachers. Educators responded on a 5-point scale (1 = strongly disagree to 5 = strongly agree). We computed a mean score for all 9 items, with higher scores indicating higher perceived benefits of CARE practices. Internal reliability was $\alpha = 0.96$.

Teacher Impact Evaluation Measures

Mindfulness. We measured mindfulness using the Five Factor Mindfulness Questionnaire (FFMQ; Baer et al., 2004), a multi-dimensional measure of mindfulness asking respondents to rate their agreement with various statements using 5-point Likert scale items. The FFMQ has five subscales that include Observing, Describing, Acting with Awareness, Non-Judgmental, and Non-Reactive mindfulness. Example items include "I pay attention to sensations, such as the wind in my hair or sun on my face", "I criticize myself for having irrational or inappropriate emotions", "I'm good at finding words to describe my feelings", "When I do things, my mind wanders off and I'm easily distracted", and "I perceive my feelings and emotions without having to react to them." We computed mean scores for all items and for each subscale, with higher scores indicating greater mindfulness. Internal reliability for the overall and subscale means ranged from $\alpha = .63–.91$ at baseline, $\alpha = .67–.94$ at post-test, and $\alpha = .74–.94$ at follow-up.

Emotion Regulation. We measured emotion regulation using the Cognitive Reappraisal and Expressive Suppression subscales of the Emotion Regulation Questionnaire (ERQ; Gross & John, 2003), a brief questionnaire asking respondents to rate their agreement with various statements about their emotions using 7-point Likert scale items. The Cognitive Reappraisal subscale consists of five items, and example items include "When I want to feel more positive emotion (such as joy or amusement), I change what I'm thinking about" and "I control my emotions by changing the way I think about the situation I'm in". The Expressive Suppression subscale consists of five items, and example items include "I keep my emotions to myself" and "I control my emotions by not expressing them". We computed mean scores for each subscale,

with higher scores indicating better emotion regulation. Internal reliabilities for Cognitive Reappraisal and Expressive Suppression, respectively, were α = .87 and α = .75 at baseline, α = .92 and α = .65 at post-test, and α = .82 and α = .83 at follow-up.

Depressive Symptoms. We measured depressive symptoms using the eight-item Patient Health Questionnaire depression scale (PHQ-8; Kroenke et al., 2008), a brief screener asking respondents about their experience of various depressive symptoms over the past two weeks using 4-point Likert scale items. Example items include "feeling down, depressed, or hopeless" and "little interest or pleasure in doing things". We computed a mean score of all the items, with higher scores indicating more depressive symptoms. Internal reliability was α = .89 at baseline, α = .90 at post-test, and α = .88 at follow-up.

Anxiety Symptoms. We measured anxiety symptoms using the seven-item Generalized Anxiety Disorder Scale (GAD-7; Spitzer et al., 2006), a brief screener asking respondents about their experience of various anxiety symptoms over the past two weeks using 4-point Likert scale items. Example items include "feeling nervous, anxious or on edge" and "not being able to stop or control worrying". We computed a mean score of all the items, with higher scores indicating more anxiety symptoms. Internal reliability was α = .94 at baseline, α = .92 at post-test, and α = .91 at follow-up.

Emotional Exhaustion. We measured emotional exhaustion using the Emotional Exhaustion subscale of the Maslach Burnout Inventory: Educators Survey (MBI; Maslach et al., 1997), a questionnaire that assesses burnout among teachers by asking respondents to rate the frequency that they experience various job-related feelings using 7-point Likert scale items. The Emotional Exhaustion subscale consists of nine items, and example items include "I feel emotionally drained from my work" and "I feel used up at the end of the workday". We computed a mean score of all the subscale items, with higher scores indicating more emotional exhaustion. Internal reliability was α = .94 at baseline, α = .94 at post-test, and α = .96 at follow-up.

Sleep Quality. We measured sleep quality using the four-item Patient-Reported Outcomes Measurement Information System Sleep Disturbance Questionnaire (PROMIS; Buysse et al., 2010), a brief screener asking respondents to rate their sleep quality over the past seven days using 5-point Likert scale items. Example items include "in the past 7 days, my sleep was refreshing" and "in the past 7 days, I had a problem with my sleep". We computed a mean score of all the items, with higher scores indicating better sleep quality. Internal reliability was α = .83 at baseline, α = .79 at post-test, and α = .76 at follow-up.

Time-Related Stress. We measured time-related stress using the Time Urgency Scale (TUS; Conte et al., 1991), a questionnaire asking respondents to rate their agreement with

statements about their time urgency using 5-point Likert scale items. The Task-Related Hurry subscale consists of three items, and examples include “I often feel very pressed for time” and “I am often in a hurry”. The Speech Patterns subscale consists of five items, and examples include “I talk more rapidly than most people” and “I often try to persuade others to my point of view”. The General Hurry subscale consists of five items, and examples include “I am slow doing things” and “I usually work fast”. We computed mean scores for all items and for each subscale, with higher scores indicating greater time-related stress. Internal reliability for the overall and subscale means ranged from $\alpha = .71–.89$ at baseline, $\alpha = .53–.90$ at post-test, and $\alpha = .74–.94$ at follow-up.

Classroom Environment Quality. We measured classroom environment quality using the Classroom Assessment Scoring System (CLASS, Pianta et al., 2008), a naturalistic classroom observation in which independent observers rate teacher–student interactions on a 1-7 scale across three domains, each of which includes multiple dimensions. The Emotional Support domain includes four dimensions: positive climate, negative climate, teacher sensitivity, and regard for student perspectives. The Classroom Organization domain includes three dimensions: behavior management, productivity, and instructional learning formats. The Instructional Support domain includes three dimensions: concept development, quality of feedback, and language modeling. Observers completed three 15-minute observation cycles for each teacher. We computed mean scores averaging the three cycles for each dimension and for each dimension, with higher scores indicating higher quality classroom environments. All observers passed CLASS interrater reliability certification for observing Kindergarten to 3rd grade and upper elementary classrooms. However, due to difficulty accessing schools for observations during the COVID pandemic, we were unable to double code observations to assess interrater reliability within our sample.

PATHS Implementation Quality. We measured teachers’ PATHS implementation quality using both classroom observations and teacher self-reports. These measures were collected at post-test and follow-up only because teachers had no knowledge of how to implement PATHS at baseline. We conducted a naturalistic classroom observation using the PATHS Lesson Observation Form (Conduct Problems Prevention Research Group, 1999), in which independent observers rate teachers’ overall implementation quality with a 1-7 global rating, as well as specific aspects of PATHS implementation rated on a 1-4 scale across three domains, each of which includes multiple dimensions. The PATHS Lesson Observation domain included four dimensions: Teaching of PATHS Concepts, Level of Disruption during PATHS Lessons, Pacing of PATHS Lessons, and Teacher Affect and Energy during PATHS Lessons. The PATHS Modeling and Generalization domain included four dimensions: PATHS Language, Kid of the Day, PATHS Concepts, and Student Social-Emotional Competence. The General Teaching Style and Behavior domain included five dimensions: Teacher Openness to Consultation, Teacher Ability to

Problem Solve, Teacher Level of Punitive Discipline, Teacher Interpersonal Style, and Teacher Classroom Management and Discipline Skills. Certified reliable coders observed for approximately 45 minutes which included a PATHS lesson. We examined the global rating and each dimension score individually, and we computed mean scores for each domain, with higher scores indicating higher quality PATHS implementation. Due to difficulty accessing schools for observations during the COVID pandemic, we were unable to double code observations to assess interrater reliability within our sample.

For teacher reports, we measured PATHS implementation quality with two single-item questions and a five-item scale (Domitrovich et al, 2019). The first question asked teachers how many PATHS lessons they taught in the past month, and they could respond with any number. The second question asked teachers how regularly they implemented PATHS lessons on a 1-5 scale ranging from "not at all" to "frequently, two lessons a week". The five-item scale asked about their implementation of PATHS principles, and example items included "I project positive feeling, energy, and enthusiasm during lessons" and "I actively model PATHS strategies by using my own feeling faces, the CSP, Does Turtle, etc." We examined the two single item questions individually and computed a mean score of the five-item scale, with higher scores indicating higher quality PATHS implementation. Internal reliability for the five-item scale was $\alpha = .87$ at post-test, and $\alpha = .77$ at follow-up.

Student Impact Evaluation Measures

Attendance. We measured student attendance using attendance rates reported in district administrative records. Attendance rate was calculated by the district as the number of days a student attended school divided by the number of days they were enrolled during the school year. In addition, we computed a dichotomous variable reflecting chronic absenteeism. Students with attendance rates less than or equal to 90% were scored as chronically absent, and students with attendance rates greater than 90% were scored as not chronically absent.

Conduct Problems. We measured conduct problems using both district administrative records and teacher reports. Administrative records included every behavioral disciplinary incident that was reported to the district for each student, as well as whether students were given in- or out-of-school suspensions for each incident. Across the school years studied, very few students were given suspensions of either type, so we focused our analysis on the occurrence of disciplinary incidents. We computed a sum of the number of disciplinary incidents for each student during the academic year, as well as a dichotomous variable reflecting whether they had one or more disciplinary incidents during the school year.

For teacher reports, we measured conduct problems using the Teacher Observation of Classroom Adaptation–Revised (TOCA-R; Werthamer-Larsson et al., 1991), a questionnaire

asking respondents to rate their agreement with statements about students' behavior in class using 6-point Liker scale items. The Aggression subscale includes seven items, and example items include "breaks rules" and "physically fights". The Inattention/Hyperactivity subscale has six items, and example items include "pays attention" and "stays on task". We computed mean scores for all items and for each subscale, with higher scores indicating greater aggression and inattention/hyperactivity. Internal reliability for the overall and subscale means ranged from α = .86–.96 at baseline, α = .88–.96 at post-test, and α = .91–.97 at follow-up.

English Language Arts Achievement. We measured English language arts (ELA) achievement using formative assessment (grades 1-2) and standardized test (grades 3-4) scores retrieved from district administrative records and teacher reports. Grade 1 and 2 students were administered the mCLASS Text Reading and Comprehension (TRC; Amplify, 2013), an assessment of students' reading proficiency and instructional reading level. Students' scores were reported in the administrative records using a 4-point proficiency rating ranging from "far below proficient" to "above proficient". We examined TRC proficiency scores as a multinomial categorial outcome. Grade 3 and 4 students were assessed by the ELA portion of the Illinois Assessment of Readiness (IAR), which assesses grade-appropriate reading literary text, informational text, vocabulary, written expression, and knowledge of language and conventions. Students' raw IAR scores were reported in the administrative records, and we standardized these scores for analyses.

For teacher report, we measured ELA achievement using three items from the Reading Skills subscale of the Academic Competence Evaluation Scale – Teacher Form (ACES; DiPerna & Elliot, 1999), which asks respondents to rate students' reading comprehension, reading fluency, and written communication using a 5-point Likert scale ranging from "far below grade level" to "far above grade level". We computed a mean score of all items, with higher scores indicating greater ELA achievement. Internal reliability was α = .95 at baseline, α = .97 at post-test, and α = .97 at follow-up.

Mathematics Achievement. We measured mathematics achievement using standardized test scores (grades 3 and 4) retrieved from district administrative records and teacher reports. The district did not report any math assessments for 1st and 2nd grade students, but teacher reports of math achievement included students from grades 1-4. Grade 3 and 4 students were assessed by the math portion of the IAR, which assesses grade-appropriate operations and algebraic thinking, numbers and operations in base ten, numbers and operations with fractions, measurement and data, and geometry. Students' raw IAR scores were reported in the administrative records, and we standardized these scores for analyses.

For teacher report, we measured math achievement using three items from the Mathematics Skills subscale of the ACES, which asks respondents to rate students' computation, measurement, problem solving using a 5-point Likert scale ranging from "far below grade level" to "far above grade level". We computed a mean score of all items, with higher scores indicating greater math achievement. Internal reliability was α = .96 at baseline, α = .98 at post-test, and α = .97 at follow-up.

Student Motivation. We measured student motivation using three items from the Motivation subscale of the ACES, which asks teachers to report on how frequently students exhibit behaviors that contribute to their academic performance using 5-point Likert scale items. The items included "prefers challenging tasks", "persists when task is difficult", and "assumes responsibility for own learning". We computed a mean score of all items, with higher scores indicating greater student motivation. Internal reliability was α = .91 at baseline, α = .95 at post-test, and α = .95 at follow-up.

Student Engagement. We measured student engagement using three items from the Engagement subscale of the ACES. The items included "asks questions about tests and projects", "participates in class discussions", and "volunteers to answer questions. We computed a mean score of all items, with higher scores indicating greater student engagement. Internal reliability was α = .89 at baseline, α = .92 at post-test, and α = .91 at follow-up.

Social-Emotional Learning. We measured SEL using the Social Skills Improvement System (SSIS; Gresham & Elliot, 2008) and the Teacher Rating of Student Social and Emotional Competencies (TRSSEC; Collaborative for Academic, Social, and Emotional Learning [CASEL] & AIR, 2013), two questionnaires that ask respondents to rate how frequently students display various socially appropriate classroom behaviors using 4-point Likert scale items. The SSIS consists of 10 items, and example items include "interacts well with other children", and "shows concern for others". We used five items from the TRSSEC, and example items include "accepts when things don't go his or her way" and "respects other people's viewpoints". We computed mean scores for each questionnaire, with higher scores indicating greater SEL. Internal reliability of the SSIS was α = .93 at baseline, α = .95 at post-test, and α = .92 at follow-up. Internal reliability of the TRSSEC was α = .84 at baseline, α = .91 at post-test, and α = .90 at follow-up.

Analysis Plan

Implementation Analyses

We conducted all implementation analyses in Stata 18.5 (StataCorp, 2024). To assess and describe implementation of the CARE program, we examined descriptive statistics of all

observer-rated implementation measures and educator-reported responsiveness and uptake measures for the full implementation sample as well as for the impact (sub)sample.

Impact Analyses

We conducted all impact analyses using path analyses in Mplus 8 (Muthén & Muthén, 1998–2024). To assess teacher survey outcomes, we used single-level univariate path analyses that regressed each outcome on treatment status, baseline scores, teacher gender, and teacher years of teaching experience as covariates. For teacher observation outcomes, we used single-level multivariate path analyses that regressed domain and subscale outcomes on treatment status, teacher gender, and years of teaching experience (e.g., one model for the three CLASS domains and one model for ten subscales). We used the Satorra-Bentler scaled, mean-adjusted chi-square difference test to compare these conditional models (saturated) to their respective null models, in which the effects of the predictors on the outcomes were constrained to zero (Satorra & Bentler, 2001).

For student outcomes, we conducted multilevel path analyses. We used two-level random intercept models that regressed each outcome on baseline scores and student demographic covariates at level one; and treatment status, teacher gender and years of teaching experience, and classroom proportions of student demographic covariates (compositional effects) at level two. The level one (student) model was:

$$Y_{sj} = \pi_{0j} + \pi_{1j}B_{sj} + \pi_{2j}C_{sj} + \varepsilon_{sj}$$

Where Y_{sj} is the outcome score for student *s* taught by teacher *j*, B_{sj} is the baseline score, C_{sj} is a vector of student demographic covariates, π_{0j} is the average score on the outcome for students taught by teacher *j*, π_{1j} is the association between baseline and outcome scores for students taught by teacher *j*, π_{2j} is the association between student demographic covariates and outcome scores for students taught by teacher *j*, and ε_{sj} is the random error associated with student *s* taught by teacher *j*. Student covariates included grade level, gender, FRPL status (dichotomous), LEP status (dichotomous), special education status (dichotomous), and race/ethnicity (dummy codes for Black, Hispanic, and White)[2].

[2] The number of American Indian or Native American, Asian, Hawaiian or Pacific Islander, and multiracial students was too small to include dummy variables for those groups. Students in those racial/ethnic groups were included in analyses but were scored as 0s for all three dummy variables.

The level two (teacher) model was:

$$\pi_{0j} = \beta_{00} + \beta_{01}T_j + \beta_{02}G_j + \beta_{03}E_j + \beta_{04}S_j + r_{0j}$$
$$\pi_{1j} = \beta_{10}$$
$$\pi_{2j} = \beta_{20}$$

Where T_j is the treatment status of teacher *j*, G_j is the gender of teacher *j*, E_j is the years of teaching experience of teacher *j*, S_j is a vector of classroom proportions of student demographic covariates, β_{00} is the average student outcome score across all study teachers, β_{01} is the pooled difference between CARE+PATHS and PATHS Only teachers on the student outcome, β_{02} is the association between teacher gender and the classroom average outcome score, β_{03} is the association between teacher years of experience and the classroom average outcome score, β_{04} is the association between classroom proportions of student demographic covariates and the classroom average outcome score, β_{10} is the average association between baseline and outcome scores across all students taught by study teachers, β_{20} is the average association between student demographic covariates and outcome scores across all students taught by study teachers, and r_{0j} is the random error associated with teacher *j*. After testing main effect models for each outcome, we tested exploratory moderation models that added level two interaction effects between intervention status and classroom proportions of FRPL status, LEP status, and race dummy codes. Each moderator was tested in a separate model run. Significant interaction effects were probed using simple slopes at the mean and plus/minus one standard deviation of the moderator.

We used the Satorra-Bentler scaled, mean-adjusted chi-square difference test to compare nested models (Satorra & Bentler, 2001). For continuous student outcomes, we assessed model fit using the root mean square error of approximation (RMSEA), standardized root-mean square residual within (SRMRw) and between (SRMRb), comparative fit index (CFI), and Tucker-Lewis Index (TLI). Good fit was defined as CFI and TLI values ≥ 0.95, RMSEA values ≤ 0.06, and SRMR values ≤ 0.08 (Hu & Bentler, 1999). For categorical student outcomes (i.e., having one or more disciplinary incident, chronic absenteeism, TRC), we conducted logistic multilevel path analyses, and exact fit statistics cannot be computed because outcomes are not normally distributed. Therefore, we used relative fit indices to compare nested models, including chi-square difference tests, Akaike information criterion (AIC), Bayesian information criterion (BIC), and sample-size adjusted BIC (aBIC). For all teacher and student outcome models, we handled missing data using full information maximum likelihood methods (FIML; Enders, 2025; Enders & Bandalos, 2001). Exhibits A3 and A4 in Appendix A show the means, standard deviations, and analytic sample sizes for both intervention groups across timepoints with full data versus listwise deletion.

Implementation Evaluation Results

CARE Program Adherence and Quality

CARE facilitators demonstrated high adherence to the CARE program. On average, they completed 85-95% of planned activities across all trainings. Facilitators also met learning objectives at an average proficiency of 3.3 to 3.6 on a 0 to 4 scale across trainings, indicating that they met or exceeded the fidelity thresholds for both activities completed and learning objectives met (see Exhibit 2).

Exhibit 2. Overview of CARE Program Adherence

Component	CARE Day 1 (N=18)		CARE Day 2 (N=18)		CARE Day 3 (N=17)	CARE Booster (N=11)
	A.M.	P.M.	A.M.	P.M.		
Average Percent of Activities Completed (SD)	95 (5)	92 (8)	93 (4)	87 (11)	85 (15)	89 (10)
Average Learning Objectives (SD)	3.6 (0.2)	3.5 (0.4)	3.6 (0.5)	3.4 (0.5)	3.4 (0.4)	3.3 (0.3)

Note. A.M. = morning training session; P.M. = afternoon training session; N = sample size; SD = standard deviation. Average learning objectives met was rated on a 0–4 scale.

Facilitation quality was consistently high across trainings, with the average observer-reported ratings falling between 3.8 and 4.0 on a 4-point scale, indicating that CARE facilitators exemplified all behavioral indicators at all or almost all opportunities. The quality of the Zoom platform was also excellent, with observer ratings falling between 4.7 and 4.9 on a 5-point scale (see Exhibit 3).

Exhibit 3. Facilitation Quality

Component	CARE Day 1 (N=18)	CARE Day 2 (N=18)	CARE Day 3 (N=15)	CARE Booster (N=11)
Zoom Quality (SD)	4.8 (0.4)	4.7 (0.4)	4.7 (0.7)	4.9 (0.2)
Facilitator delivery	3.9 (0.2)	3.9 (0.2)	3.8 (0.3)	4.0 (0.1)

Note. N = sample size; SD = standard deviation. Zoom quality was rated on a 1–5 scale, and facilitator delivery was rated on a 1–4 scale.

Educator Attendance, Perceived Usefulness, Satisfaction, and Engagement

On average, teachers in the impact sample attended more than double the hours of CARE training than the full implementation sample (see Exhibit 4). This reflects the fact that we selected the impact sample based on their level of participation in Days 1 and 2 of training. We also had less communication with the additional participants throughout the project due to their lack of involvement in other aspects of the research, which may have led them to attend fewer CARE sessions.

Exhibit 4. Overview of CARE attendance

Sample	Attendance at CARE Trainings				Hours of Participation in CARE Training	
	Day 1 N (%)	Day 2 N (%)	Day 3 N (%)	Booster N (%)	Mean	SD
Impact sample	78 (100)	78 (100)	65 (85.5)	54 (70.1)	11.1	1.3
Implementation sample	208 (32.3)	211 (32.8)	288 (44.7)	308 (47.8)	4.7	4.3

Note. N = sample size; SD = standard deviation.

Educators indicated that they agreed or strongly agreed that the concepts and activities taught in the CARE trainings were useful in their lives and in their work as a teacher (see Exhibit 5). Participants reported being satisfied with the CARE trainings, with average satisfaction scores ranging from 4.1–4.4 on a 5-point scale across training sessions (see Exhibit 6). Participants also reported being engaged or highly engaged in the trainings each day, with average self-report ratings from 3.5–3.6 on a 4-point scale (see Exhibit 7).

Exhibit 5. Usefulness of CARE

CARE Training	Implementation Sample			Impact Sample		
	N	Mean	SD	N	Mean	SD
Day 1	147	4.4	0.6	58	4.5	0.7
Day 2	134	4.6	0.6	50	4.7	0.6
Day 3	174	4.3	0.8	44	4.3	0.7
Booster	172	4.3	0.9	30	4.6	0.7

Note. N = sample size; SD = standard deviation. Usefulness of CARE was rated on a 1–5 scale.

Exhibit 6. Satisfaction with CARE

CARE Training	Implementation Sample			Impact Sample		
	N	Mean	SD	N	Mean	SD
Day 1 & 2	131	4.4	0.7	49	4.4	0.5
Day 3	159	4.1	1.0	37	4.3	0.6
Booster	162	4.1	1.1	27	4.4	0.9

Note. N = sample size; SD = standard deviation. Satisfaction with CARE was rated on a 1–5 scale.

Exhibit 7. Engagement in CARE Trainings

CARE Training	Implementation Sample			Impact Sample		
	N	Mean	SD	N	Mean	SD
Day 1	144	3.6	0.4	57	3.6	0.4
Day 2	132	3.6	0.4	50	3.6	0.4
Day 3	172	3.5	0.6	43	3.5	1.2

Note. N = sample size; SD = standard deviation. Engagement in CARE trainings was rated on a 1–4 scale.

Uptake of the CARE program

After the booster sessions in January and February, educators reported that they used most CARE skills on a weekly basis on average (from one time to multiple times per week; see Exhibit 8). They reported using three deep breaths, mindful listening, and experiencing joy at least 2-3 times per week on average, whereas they used the body scan and recognizing scripts less than once per week on average. Educators also reported that they benefitted from the CARE trainings.

Exhibit 8. Frequency of CARE practice

CARE Practice	Implementation Sample			Impact Sample		
	N	Mean	SD	N	Mean	SD
3 Deep Breaths	164	4.3	1.5	27	4.5	1.3
Breath Awareness	163	3.9	1.6	27	3.6	1.7
Caring Practice	161	3.9	1.5	27	3.8	1.4
Body Scan	163	2.9	1.7	26	3.0	1.6
Setting Intention	161	3.4	1.6	26	3.6	1.4
Mindful Listening	159	4.4	1.3	27	4.1	1.3
Experiencing Joy	160	4.4	1.5	27	4.0	1.3
Recognizing Scripts	161	2.8	1.6	27	2.9	1.6

Note. N = sample size; SD = standard deviation. Frequency of CARE practice was rated on a 1–6 scale.

Mean CARE benefit scores ranged from 3.9–4.2 on a 5-point scale, indicating that they agreed that benefited personally and professionally from receiving CARE training (see Exhibit 9). Importantly, they agreed that this type of training should be integrated into preparation and in-service training for all teachers (implementation sample M = 4.1, SD = 1.0; impact sample M = 4.25, SD = 0.70).

Exhibit 9. Benefits of CARE practices

CARE Training	Implementation Sample			Impact Sample		
	N	Mean	SD	N	Mean	SD
Day 1 & 2	131	4.2	0.6	49	4.2	0.5
Day 3	156	4.0	0.6	38	3.9	0.6
Booster	161	4.0	0.8	28	4.1	0.6

Note. N = sample size; SD = standard deviation. Benefits of CARE practices were rated on a 1–5 scale.

Impact Evaluation Results

Teacher Descriptive Statistics

Exhibit A5 in Appendix A shows bivariate correlations for self-reported teacher survey variables. In general, baseline (spring 2021), post-test (spring 2022), and follow-up (spring 2023) scores for each variable were positively correlated with each other. Teachers' FFMQ Acting with Awareness post-test score and FFMQ Overall and FFMQ Observing follow-up scores were moderately negatively correlated with intervention status, indicating the PATHS Only had higher scores than the CARE+PATHS group on those mindfulness indicators. Depressive symptoms were moderately positively correlated with intervention status at post-test marginally positively correlated at follow-up, indicating greater depressive symptoms in the CARE+PATHS group than the PATHS Only group. TUS General Hurry scores at post-test were moderately negatively correlated with intervention status, indicating lower general time-related stress in the CARE+PATHS group than the PATHS Only group. No other teacher self-reported outcomes were significantly correlated with intervention status.

Exhibits A6 and A7 in Appendix A show bivariate correlations of the post-test CLASS and PATHS observations conducted in spring 2022. With a few exceptions, the CLASS domain and subscale scores were generally positively correlated with each other. Regard for Student Perspectives subscale scores were positively correlated with intervention status, indicating higher scores in the CARE+PATHS group than the PATHS Only group. Productivity subscale scores were also marginally positively correlated with intervention status. For the PATHS observation, four subscale scores (Level of Disruption during a PATHS Lesson, Pacing of PATHS Lesson, Teaching

Affect and Energy during PATHS Lesson, and Teacher Level of Punitive Discipline) and one domain score (PATHS Lesson Observation) were highly skewed, with almost all observed teachers receiving the highest or second highest (or lowest) score. In addition, the Ability to Problem Solve subscale only had n = 7 observations. Therefore, we did not analyze those outcomes. The remaining PATHS observation scores were positively correlated with each other, but none were significantly correlated with intervention status.

Teacher Main Results at Post-Test and Follow-Up

Exhibit 10 displays the results from models comparing intervention group differences in teacher's self-reported survey outcomes at the post-test (spring 2022) and follow-up (spring 2023). At post-test, teachers in the CARE+PATHS group had significantly lower General Hurry scores than the PATHS Only group, but there were no significant group differences in Task- or Speech-Related Hurry. Counter to our hypotheses, the CARE+PATHS group had lower scores on the Acting with Awareness subscale of the FFMQ at post-test and the Observing subscale at follow-up than the PATHS Only group. However, there were no significant group differences in the overall mindfulness score or the other mindfulness subscales. Likewise, the CARE+PATHS group had significantly more depressive symptoms than the PATHS Only group at both the post-test and follow-up. On average, however, their depressive symptoms were not clinically significant, as the mean scores in both groups were below clinical cutoffs for positive depression screens (Wu et al., 2020). We found no other significant group differences in self-reported teacher emotion regulation, burnout, or symptoms of anxiety. There were also no differences in PATHS implementation outcomes.

Exhibit 10. Teacher Survey Outcome Results

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
FFMQ – Overall						
Post-test	3.43(0.57)	3.65(0.50)	-0.23	0.10	-.13	.160
Follow-up	3.35(0.57)	3.71(0.35)	-0.35	0.12	-.20	.089
FFMQ – Observing						
Post-test	3.42(0.68)	3.48(0.77)	-0.06	0.15	-.01	0.932
Follow-up	3.19(0.69)	3.78(0.57)	-0.58	0.19	-.34	**.018**
FFMQ – Describing						
Post-test	3.74(0.82)	3.92(0.76)	-0.18	0.14	-.11	.206
Follow-up	3.54(0.89)	3.71(0.61)	-0.17	0.17	-.08	.428
FFMQ – Acting with Awareness						
Post-test	3.35(0.76)	3.77(0.72)	-0.41	0.14	-.24	**.010**

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
Follow-up	3.21(1.03)	3.73(0.77)	-0.53	0.20	-.19	.058
FFMQ – Non-reactive						
Post-test	3.22(0.65)	3.47(0.63)	-0.26	0.12	-.14	.137
Follow-up	3.55(0.91)	3.89(0.63)	-0.34	0.21	-.19	.174
FFMQ – Non-judgmental						
Post-test	3.41(0.71)	3.63(0.69)	-0.22	0.12	-.07	.423
Follow-up	3.28(0.68)	3.41(0.53)	-0.13	0.18	-.02	.884
ERQ – Reappraisal						
Post-test	5.50(0.91)	5.18(1.27)	0.33	0.24	.20	.075
Follow-up	5.38(0.89)	5.67(0.79)	-0.28	0.24	-.17	.216
ERQ – Suppression						
Post-test	3.18(1.10)	3.39(1.01)	-0.22	0.24	-.02	.834
Follow-up	3.43(1.30)	3.64(1.41)	-0.21	0.32	.03	.767
PHQ-8						
Post-test	1.76(0.69)	1.40(0.43)	0.36	0.12	.21	**.021**
Follow-up	1.67(0.64)	1.38(0.44)	0.29	0.13	.26	**.020**
GAD-7						
Post-test	1.92(0.70)	1.67(0.73)	0.25	0.13	.13	.167
Follow-up	1.86(0.79)	1.61(0.66)	0.25	0.16	.17	.080
MBI						
Post-test	4.23(1.62)	3.65(1.46)	0.58	0.29	-.02	.852
Follow-up	4.22(1.94)	3.40(1.57)	0.82	0.37	.12	.242
PROMIS						
Post-test	3.57(0.72)	3.74(0.85)	-0.17	0.18	.01	.958
Follow-up[1]	3.62(0.73)	3.75(0.73)	-0.13	0.24	-.12	.451
TUS – Overall						
Post-test	3.17(0.55)	3.15(0.56)	0.02	0.09	-.01	.929
Follow-up	3.19(0.64)	3.17(0.58)	0.02	0.11	-.06	.557
TUS – Task-Related Hurry						
Post-test	3.59(0.96)	3.46(1.02)	0.13	0.18	.06	.480
Follow-up	3.57(1.14)	3.59(0.97)	-.03	0.22	-.03	.783
TUS – Speech Patterns						

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
Post-test	2.90(0.81)	2.62(0.73)	0.28	0.12	.09	.271
Follow-up	2.80(0.97)	2.67(0.77)	0.13	0.17	-.07	.474
TUS – General Hurry						
Post-test	3.01(0.72)	3.36(0.56)	-0.36	0.12	-.24	**.013**
Follow-up	3.20(0.64)	3.25(0.64)	-0.05	0.13	-.03	.758
Number of PATHS Lessons Taught in the Past Month						
Post-test[1]	8.63(10.67)	9.63(10.61)	-1.00	2.88	-.03	.837
Follow-up[1]	8.89(7.97)	11.88(12.01)	-2.99	3.62	-.19	.313
Frequency of PATHS Lessons Taught						
Post-test	3.67(1.11)	3.86(1.04)	-0.19	0.29	-.08	.569
Follow-up	3.76(1.09)	3.53(1.42)	0.24	0.44	.10	.578
Self-Reported PATHS Implementation Quality						
Post-test[1]	4.01(0.88)	4.10(0.95)	-0.09	0.24	-.03	.822
Follow-up[1]	4.15(0.80)	4.44(0.59)	-0.29	0.25	-.16	.377

Note. M = Mean; SD = standard deviation; SE = standard error. Effect sizes were calculated as standardized betas.
[1]Model fit was not significantly better than the null model

Exhibit 11 displays the results comparing intervention group differences in classroom environment quality and teacher–student relationships as measured by the CLASS observation at post-test (spring 2022). There were no significant group differences in the three CLASS domain scores. However, when we examined the CLASS subscales, we found that the CARE+PATHS group had significantly higher Regard for Student Perspectives and Productivity subscale scores than the PATHS Only group. Behaviors that indicate Regard for Student Perspectives include flexibility and student focus, support for student autonomy and leadership, student expression, connections to current life, meaningful peer interactions, and minimal restriction of student movement. Behaviors that indicate Productivity include maximizing learning time, routines, transitions, and preparation. We found no other significant group differences in CLASS subscale scores but in all cases the means favored the CARE+PATHS group.

Exhibit 11. CLASS Observation Results

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
Emotional Support	5.37(0.72)	5.12(0.83)	0.25	0.17	.17	.113
Positive Climate	5.33(1.15)	5.20(1.16)	0.13	0.25	.05	.660
Negative Climate	1.23(0.42)	1.33(0.66)	-0.10	0.12	-.13	.225
Teacher Sensitivity	5.20(1.01)	4.89(1.24)	0.31	0.24	.13	.235
Regard for Student Perspectives	4.20(1.06)	3.74(0.95)	0.46	0.21	.24	**.018**
Classroom Organization	5.49(0.74)	5.20(1.07)	0.29	0.19	.18	.098
Behavior Management	5.72(1.08)	5.62(1.28)	0.11	0.24	.09	.366
Productivity	6.02(0.95)	5.56(1.30)	0.46	0.24	.23	**.027**
Instructional Learning Formats	4.72(0.98)	4.42(1.14)	0.30	0.23	.13	.226
Instructional Support	3.02(1.00)	2.80(0.86)	0.22	0.20	.11	.286
Concept Development	2.70(0.90)	2.34(0.84)	0.35	0.19	.19	.071
Quality of Feedback	3.31(1.21)	3.29(1.20)	0.02	0.26	.00	.993
Language Modeling	3.05(1.17)	2.78(1.00)	0.27	0.23	.13	.247

Note. M = Mean; SD = standard deviation; SE = standard error. Effect sizes were calculated as standardized betas.

Exhibit 12 displays the results comparing intervention group differences in observed PATHS implementation quality. Counter to our hypotheses, there were no significant group differences in teachers' PATHS implementation quality. Taken together, we found no evidence across teacher self-reports and independent observations that CARE improved teachers' implementation of PATHS.

Exhibit 12. PATHS Classroom Observation Results

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
Global Rating[1]	4.46(1.30)	4.86(1.46)	-0.40	0.34	-.11	.357
Teaching of PATHS Concepts	3.27(1.07)	3.55(0.62)	-0.05	0.24	-.09	.479
PATHS Modeling and Generalization	2.64(0.92)	2.87(0.95)	-0.23	0.23	-.10	.422
PATHS Language	2.46(1.12)	2.86(1.08)	-0.40	0.28	-.18	.158
Kid of the Day	2.70(1.33)	2.82(1.19)	-0.12	0.32	-.06	.630
PATHS Concepts	2.27(1.15)	2.61(1.23)	-0.34	0.29	-.12	.310
Student Social-Emotional Competence	3.11(0.94)	3.18(1.06)	-0.07	0.24	.01	.921

General Teaching Style and Behavior	3.51(0.75)	3.55(0.62)	-0.05	0.17	-.01	.910
Teacher Openness to Consultation	3.41(0.87)	3.50(0.75)	-0.10	0.21	-.05	.715
Teacher Interpersonal Style	3.41(0.93)	3.39(0.88)	0.01	0.23	.01	.916
Teacher Classroom Management and Discipline Skills	3.49(0.95)	3.44(0.89)	0.04	0.23	.05	.687

Note. M = Mean; SD = standard deviation; SE = standard error. Effect sizes were calculated as standardized betas.
[1]Model fit was not significantly better than the null model

Student Descriptive Statistics

Exhibit A8 in Appendix A shows bivariate correlations for student outcome variables and covariates. In general, baseline (SY2020–21 for district records, fall 2021 for teacher reports), post-test (SY2021–22 for district records, spring 2022 for teacher reports), and follow-up (SY2022–23 for district records, spring 2023 for teacher reports) scores for each variable were positively correlated with each other. Attendance rate had a small positive correlation with intervention status at all three timepoints, indicating better attendance in the CARE+PATHS group than the PATHS Only group. Chronic absenteeism had a small negative correlation with intervention status at post-test and follow-up but not at baseline, indicating lower likelihood of chronic absenteeism in the CARE+PATHS than the PATHS Only group. Disciplinary incidents (both the total number and having one or more) had a small negative correlation with intervention status at post-test but not at follow-up, indicating fewer disciplinary incidents in the CARE+PATHS group than the PATHS Only Group. TOCA-R Overall scores were not correlated with intervention status, but the baseline Aggression subscale score had a small positive correlation with intervention status, indicating greater aggression in the CARE+PATHS group than the PATHS Only group at baseline. TRC scores at post-test, ACES Reading Skills scores at post-test and follow-up, and ACES Mathematics at baseline and post-test had small negative correlations with intervention status, indicating lower teacher-reported academic competence in the CARE+PATHS group than the PATHS Only group. However, IAR ELA and Mathematics scores were not significantly correlated with intervention status. ACES Motivation and Engagement, SSIS, and TRSSEC scores were not significantly correlated with intervention status.

In general, FRPL status was moderately negatively correlated with positive outcomes (e.g., attendance rate, standardized test scores) across timepoints and was moderately positively correlated with negative outcomes (e.g., chronic absenteeism, disciplinary incidents) across timepoints. LEP status had the opposite pattern of being positively correlated with some positive outcomes and negatively correlated with some negative outcomes. Being Black or African American was also moderately negatively correlated with several positive outcomes and

positively correlated with several negative outcomes across timepoints. The opposite was true for being White and Hispanic or Latino/a/e/x, although not as consistently across outcomes.

Student Main Results at Post-Test and Follow-Up

Main Effects

Model fit was good for all final student outcome models (see Exhibits A9 and A10 in Appendix A). Exhibit 13 displays the results from main effect models comparing intervention group differences in student outcomes at post-test and follow-up. Partially consistent with our hypotheses, there was no significant group difference in attendance rate during the implementation year (SY2021–22), but students of teachers in the CARE+PATHS group had 1.2% better attendance in the follow-up year (SY2022–23) than students of teachers in the PATHS Only group. We found the same pattern in chronic absenteeism, with no group differences in the implementation year (SY2021–22) but 4.7% less chronic absenteeism in the CARE+PATHS group compared to the PATHS Only group in the follow-up year (SY2022–23).

Also partially consistent with our hypotheses, students of teachers in the CARE+PATHS group had significantly fewer disciplinary incidents reported to the district in SY2021–22 than students of teachers in the PATHS Only group. They also had a lower likelihood of having at least one disciplinary incident, with 1.6% of students (n = 32) in the CARE+PATHS group having at least one incident compared to 5.7% of students (n = 115) in the PATHS Only group. We did not find any other significant main effects of CARE on other student behavioral or academic achievement outcomes.

Exhibit 13. Student Outcome Main Effect Model Results

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
Attendance Rate						
Post-test	0.880(0.109)	0.872(0.116)	0.009	0.005	.09	.422
Follow-up	0.915(0.079)	0.902(0.092)	0.012	0.004	.42	**.001**
Chronic Absenteeism						
Post-test	0.447(0.497)	0.488(0.500)	-0.041	0.103	-.16	.148
Follow-up	0.309(0.462)	0.356(0.479)	-0.047	0.099	-.21	**.049**
Number of Disciplinary Incidents						
Post-test	0.019(0.170)	0.109(0.560)	-0.090	0.028	-.37	**<.001**
Follow-up	0.095(0.531)	0.131(0.684)	-0.035	0.036	-.15	.184
One or More Disciplinary Incidents						
Post-test	0.016(0.125)	0.062(0.242)	-0.047	0.489	-.37	**<.001**

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
Follow-up	0.057(0.232)	0.060(0.238)	-0.003	0.394	-.16	.234
TOCA-R Overall						
Post-test	1.96(0.96)	1.87(0.81)	0.09	0.10	.11	.615
TOCA-R Aggression						
Post-test	1.58(0.87)	1.51(0.69)	0.07	0.11	-.05	.818
TOCA-R Inattention/Hyperactivity						
Post-test	2.41(1.39)	2.31(1.24)	0.10	0.15	.20	.264
TRC						
Post-test	0.99(1.25)	1.45(1.31)	0.46	0.22	-.01	.893
IAR – ELA						
Post-test	706.57(34.63)	709.42(33.88)	-2.85	0.08	-.07	.448
Follow-up	715.11(34.39)	716.49(34.02)	-1.49	0.07	-.08	.437
ACES – Reading Skills						
Post-test	2.56(1.06)	2.91(1.13)	-0.36	0.15	-.25	.187
IAR – Mathematics						
Post-test	706.06(30.96)	707.74(27.46)	-1.68	0.09	-.12	.295
Follow-up	712.86(30.95)	711.37(30.06)	1.49	0.07	.03	.774
ACES Mathematics Skills						
Post-test	2.66(0.90)	3.11(0.99)	-0.45	0.12	-.24	.194
ACES – Motivation						
Post-test	3.51(1.23)	3.72(1.09)	-0.21	0.18	-.14	.514
ACES – Engagement						
Post-test	3.79(1.14)	3.94(0.98)	-0.15	0.16	-.23	.227
SSIS						
Post-test	3.27(0.64)	3.37(0.60)	-0.10	0.08	-.16	.461
TRSSEC						
Post-test	3.20(0.89)	3.26(0.92)	-0.06	0.15	.17	.491

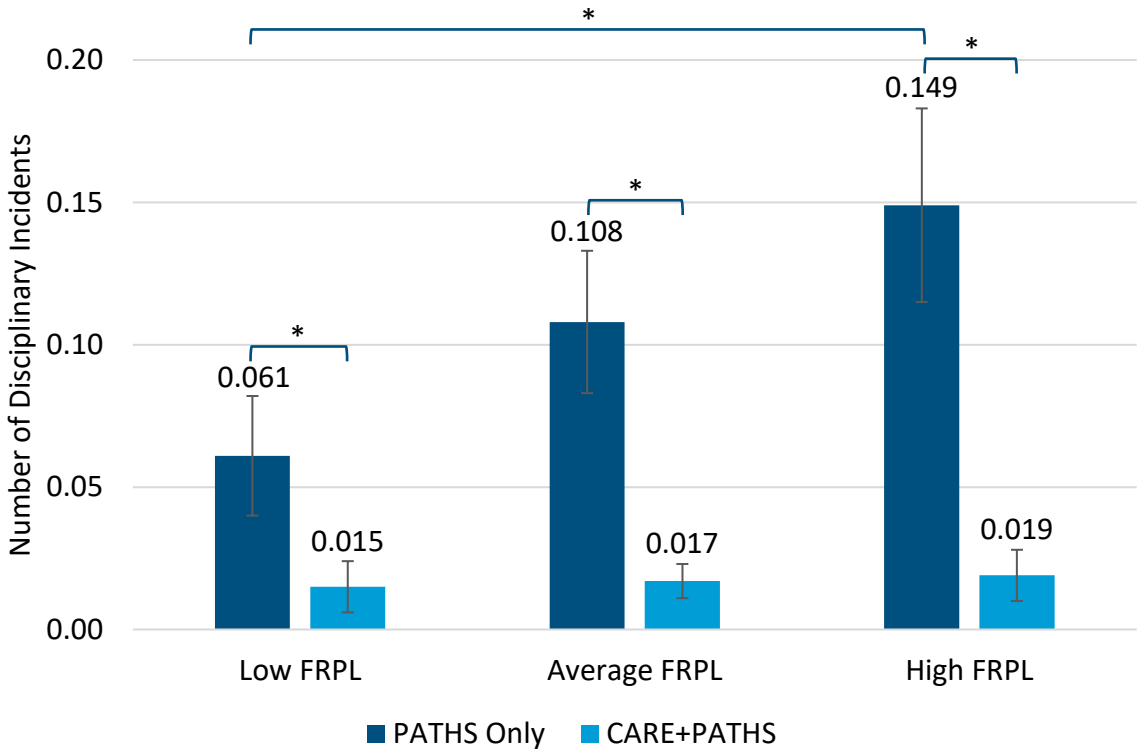
Note. M = Mean; SD = standard deviation; SE = standard error. Effect sizes were calculated as standardized betas.

Moderation Effects

Moderation analyses revealed additional patterns in the effects of CARE on student outcomes. We found that the interaction between intervention status and the classroom proportion of students qualifying for FRPL significantly predicted students' number of disciplinary incidents in

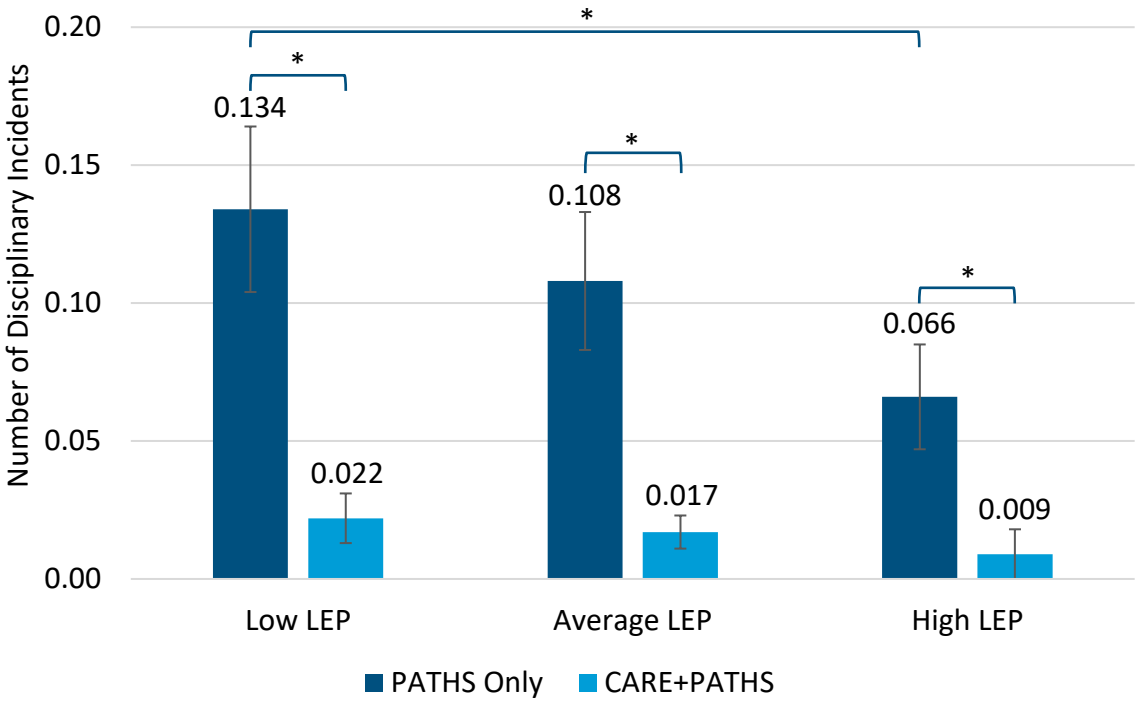
SY2021–22, *b* = -.23, SE = .08, β = -.71, *p* = .004. Exhibit 14 displays the results of the simple slopes analysis probing the interaction effect. Students in the CARE+PATHS group had significantly fewer disciplinary incidents than those in the PATHS Only group in classrooms with all levels of FRPL-eligible students, but the group difference was larger in classrooms with a high proportion of FRPL-eligible students compared to those with a low proportion of FRPL-eligible students.

Exhibit 14. SY2021-22 Number of Disciplinary Incidents by Classroom Proportion of FRPL Students



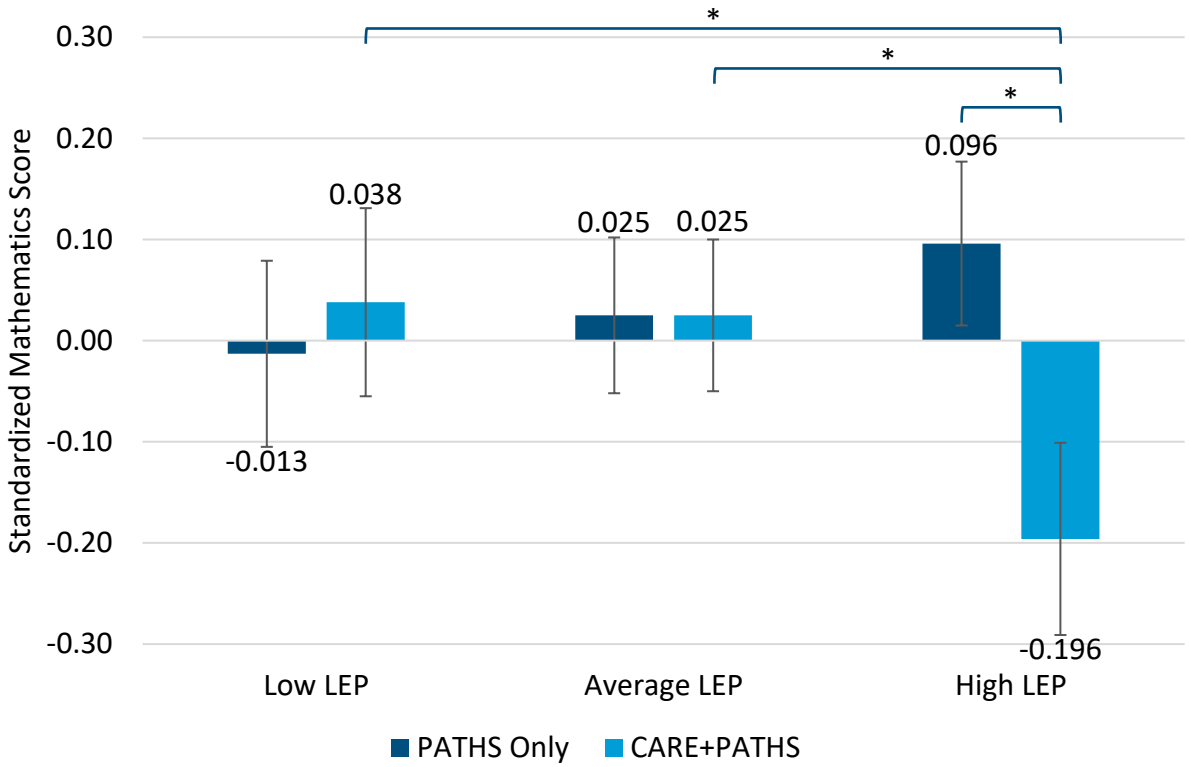
We also found that the interaction between intervention status and the classroom proportion of students classified as LEP significantly predicted students' number of disciplinary incidents in SY2021–22 (see Exhibit 15), *b* = .14, SE = .05, β = .18, *p* = .003. Students in the CARE+PATHS group had significantly fewer disciplinary incidents than those in the PATHS Only group in classrooms with all levels of students with LEP, but the group difference was larger in classrooms with a low proportion of students with LEP compared to those with a high proportion of students with LEP.

Exhibit 15. SY2021-22 Number of Disciplinary Incidents by Classroom Proportion of Students with LEP



Lastly, we found that the interaction between intervention status and the classroom proportion of students classified as LEP significantly predicted students' mathematics standardized test scores on the IAR in SY2021–22 (see Exhibit 16), b = -1.02, SE = .30, β = -.33, p = .001. Counter to our hypotheses, CARE+PATHS group students in classrooms with a high proportion of students with LEP had lower SY2021–22 math scores than PATHS-only students in similar classrooms, but not in classrooms with low or average proportions of students with LEP. Further, CARE+PATHS group students in classrooms with a high proportion of students with LEP had worse math scores than CARE+PATHS group students in classrooms with a low proportion of students with LEP.

Exhibit 16. SY2021-22 IAR Mathematics Scores by Classroom Proportion of Students with LEP



Discussion

The goals of the current study were to evaluate the quality of implementation of the CARE program and the effects of combining CARE for teachers with the PATHS SEL program for students on (1) teachers' social-emotional competence and well-being; (2) classroom environment quality and teacher–student relationships; (3) teachers' implementation of PATHS; and (4) students SEL, behavior, and academic achievement. Although our plan was to implement CARE in-person and to evaluate impact with an RCT, the COVID-19 pandemic significantly compromised our study. The CREATE and UVA teams had to deliver CARE via Zoom, and the AIR team had to change the impact evaluation from an RCT to a QED. Despite the major pandemic-related challenges, such as low program participation by teachers and virtual delivery where many teachers had their cameras turned off, we implemented CARE training with fidelity. Likewise, our QED approach allowed us to make some limited causal inferences about the effects of CARE, despite the original RCT design being compromised.

Implementation Quality

This was the first project to implement the CARE program in a virtual setting. Although CARE is designed for in-person delivery, CARE facilitators delivered the program with high program adherence (i.e., program activities and learning objectives completed) and high facilitation quality. On average and across all trainings, educators who completed the implementation surveys found the CARE content to be useful and engaging, and felt they were benefiting from and were satisfied with the program. They also used the CARE practices frequently but often were at home instead of in-person with their students in a normal classroom environment. On average, educators reported using 6 of the 8 CARE practices at least once a week, and 3 of the 8 practices at least 2-3 times per week. Although educator attendance was low overall, implementation was successful for those who did attend, as CARE facilitators delivered the program with high fidelity and educators responded positively to it.

Effects of CARE on Teacher Outcomes

With respect to teacher SEC and well-being, findings were largely inconsistent with our hypotheses. For most outcomes, there were no significant group differences between the CARE+PATHS and PATHS Only groups. We found five significant group differences, but four of those were in the unexpected direction with the CARE+PATHS group having worse mindfulness scores and greater depressive symptoms than the PATHS Only group. Teacher participation in the post-test and follow-up well-being surveys was very low, which limited our power to detect effects and is likely the primary explanation for null and counterintuitive results.

However, we can cautiously interpret these findings. CARE is designed to help teachers be more aware of their feelings and sensations and use strategies to regulate emotions and manage stress, but it was not designed to address the major stresses brought on by a global pandemic. Therefore, it is possible that receiving CARE training helped teachers become attuned enough to their own mental states to realize when they were not doing well without having the ability to overcome the major stresses of working in schools through multiple lockdowns, teacher strikes, and potential exposure to COVID. If so, teachers from the CARE+PATHS group may have been less likely to endorse survey items reflecting mindful observing (e.g., “I pay attention to how my emotions affect my thoughts and behavior”) and more likely to endorse items reflecting less mindful awareness (e.g., “I am easily distracted”) and depressive symptoms. Teachers in the PATH Only group, on the other hand, may have been equally stressed but less likely to endorse survey items reflecting that stress (Baer, 2019).

Our findings on classroom environment quality and teacher–student relationships were partially consistent with hypotheses and previous evaluations of CARE. We found no group differences in the CLASS domains, but analyses of the subscales indicated higher scores for the subscales of Regard for Student Perspectives (part of the Emotional Support domain) and

Productivity in the CARE+PATHS group than the PATHS Only group. This is in line with a previous RCT evaluation found that CARE significantly improved Emotional Support domain scores and Productivity (Jennings et al., 2017). Further, all CLASS scores were higher in the CARE+PATHS group than the PATH Only group, suggesting the power to detect effects may have again been limited by our small observation sample size (N = 88). Nevertheless, better Regard for Student Perspectives and Productivity are consistent with the CARE logic model. CARE may help teachers manage the moment-to-moment stresses of the classroom, enabling them to respond more sensitively and supportively to student needs and to create a more consistent and productive learning environment.

We also found that teachers' PATHS implementation (both self-reported and independently observed) was not significantly improved by CARE. It is possible that the potential benefits of CARE were not large enough (given its unusual pandemic-related model of delivery) to outweigh the quality of PATHS training teachers that both intervention groups received.

Effects of CARE on Student Outcomes

Although our analyses of teacher-reported student outcomes were also significantly limited by low sample size, our access to district administrative records provided enough statistical power to properly examine several student outcomes. Our finding that CARE+PATHS students had better attendance and less chronic absenteeism in the follow-up year (SY2022–23) compared to PATHS Only students is notable, given the large increase in chronic absenteeism in the United States during and after the pandemic (Dee, 2024). Although there was only a 1.2% group difference in attendance rate, the average attendance rate in both groups was close to the cutoff for chronic absenteeism (90% attendance), resulting in a 4.7% difference in chronic absenteeism between the groups. We did not have the ability to examine mediators of CARE effects on student outcomes, but it is possible that improved attendance could be partly explained by our CLASS observation results. The more supportive and productive classroom environments created by CARE+PATHS teachers may have made school a more positive experience for students, leading to an increased desire to attend school. However, future research with more statistical power to test mediation is needed to test this hypothesis.

Our finding that CARE+PATHS students had significantly fewer disciplinary incidents reported to the school district (both in total number and likelihood of having one or more incident) is also practically significant. The PATHS Only group had more than 3.5 times more students who had at least one disciplinary incident in the post-test year (SY2021–22), a difference indicating that 83 more students had disciplinary incidents in the PATHS Only group. In addition, moderation analyses showed that CARE was more effective at preventing disciplinary incidents in higher need classrooms with more FRPL-eligible (low income) students. Again, the more supportive and productive classroom environments created by CARE+PATHS teachers may have played a

role in reducing disciplinary issues among their students. They may have been better at preventing behavior problems or may have been better equipped to respond to behavior problems on their own in class instead of reporting students to the office. However, it was less effective (although still effective) in classrooms with more English learners (i.e., students with LEP). It is possible that teachers were somewhat less able to support students and address behavior problems when they faced potential language barriers. This also seemed to be true for mathematics achievement in classrooms with a high proportion of English learners. Perhaps CARE+PATHS teachers had some difficulty supporting both SEL and academic competence of their students when also addressing language barriers.

The practical significance of CARE's effect on reducing disciplinary incidents is further evidenced by an absence of an effect in the follow-up school year. When students moved up a grade to a new teacher, who may or may not have participated in CARE, the CARE+PATHS average number of disciplinary incidents and likelihood of having one or more incident rose to be statistically equal that of the PATHS Only group. The effect of CARE subsiding when students moved to a new teacher suggests that something about the classroom environment was helping prevent disciplinary issues, likely the more supportive and productive environments we observed. However, it also suggests that the positive effect of CARE was not powerful to result in long-lasting improvements in students' behavior across school years.

Study Limitations

As we described earlier in the report, our study has several limitations stemming from COVID-related disruptions. First, CARE had to be implemented virtually instead of in-person, the way it is designed to be delivered. The CARE practices are designed be taught in-person, allowing teachers to try them together with the CARE instructor. Indeed, the activities that require face-to-face interaction had to be removed. The same kind of practices and connections cannot be replicated over video call. Further, most teachers attended CARE sessions with their cameras turned off, which made it harder for CARE instructors to build a sense of community and provide real-time feedback to teachers during the sessions. Second, many fewer teachers who were eligible for the study participated in CARE than expected. Although the UVA team was able to provide CARE for many more teachers than originally planned at the request of participating schools, most of the additional teachers were not eligible for the study. Among grade 1–4 teachers in schools randomly assigned to the CARE+PATHS group, only 63.9% attended Days 1 and 2 of CARE. As a result, our original plan to conduct ITT analyses was not feasible.

In addition to the limitations on CARE implementation, teacher participation in surveys was significantly lower than expected. Likewise, we could not conduct as many classroom observations as originally planned, and we were completely unable to do CLASS observations at

baseline because of school closures. Consequently, our teacher outcome analyses had to be simplified to single-level models with few covariates that could not account for shared variance among teachers associated with working at the same school. Our teacher outcome results should be interpreted with caution and be considered preliminary. Our student outcome models that used district administrative records provided much more statistical power to detect CARE effects and confidence in our findings. However, they were still situated in a QED treatment-on-the-treated approach, which reduced our ability to make causal inferences.

Conclusion

Our study, like many others, was significantly and negatively impacted by the COVID pandemic. Despite the challenges, the UVA team was able to implement a virtual version of CARE with high fidelity. We also found significant positive effects of CARE classroom environments and student attendance and discipline. Although the findings must be interpreted with caution, they are partially consistent with a previous RCT evaluation of CARE (Jennings et al., 2017) and provide encouraging preliminary evidence of CARE's positive effects on key student outcomes. Because the current study was compromised, future research is needed to conduct a more robust evaluation of CARE. Such a study is already underway—Project ENGAGE is testing the impacts of CARE when implementation is scaled up in a larger number of urban, suburban, and rural school districts post-pandemic.

References

Alliance for Excellent Education. (2014). On the path to equity: Improving the effectiveness of beginning teachers [Press release]. Retrieved from http://all4ed.org/press/teacher-attrition-costs-united-states-up-to-2-2-billion-annually-says-new-alliance-report/

Amplify (2013). mCLASS Reading 3D: Text Reading and Comprehension. Retrieved from https://www.ode.state.or.us/wma/teachlearn/testing/resources/mclass_reading_3d_blueprints.pdf

Baer, R. (2019). Assessment of mindfulness by self-report. *Current Opinion in Psychology*, *28*, 42–48. https://doi.org/10.1016/j.copsyc.2018.10.015

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of Mindfulness by Self-Report: The Kentucky Inventory of Mindfulness Skills. *Assessment*, *11*(3), 191–206. https://doi.org/10.1177/1073191104268029

Belfield, C., Bowden, A. B., Klapp, A., Levin, H., Shand, R., & Zander, S. (2015). The economic value of social and emotional learning. *Journal of Benefit-Cost Analysis*, *6*(3), 508-544. https://doi.org/10.1017/bca.2015.55

Blum, R. W., Libbey, H. P., Bishop, J. H., & Bishop, M. (2004). School connectedness—Strengthening health and education outcomes for teenagers. *Journal of School Health*, *74*(7), 231–235.

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., Blair, C., Nelson, K. E., & Gill, S. (2008a). Promoting Academic and Social-Emotional School Readiness: The Head Start REDI Program. *Child Development*, *79*(6), 1802–1817. https://doi.org/10.1111/j.1467-8624.2008.01227.x

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008b). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology*, *20*(3), 821–843. https://doi.org/10.1017/S0954579408000394

Brown, J. L., Jennings, P. A., Rasheed, D. S., Cham, H., Doyle, S. L., Frank, J. L., Davis, R., & Greenberg, M. T. (2025). Direct and moderating impacts of the CARE mindfulness-based professional learning program for teachers on children's academic and social-emotional outcomes. *Applied Developmental Science*, *29*(1), 14–33. https://doi.org/10.1080/10888691.2023.2268327

Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., Johnston, K. L., Shablesky-Cade, M. A., & Pilkonis, P. A. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep*, *33*(6), 781–792. https://doi.org/10.1093/sleep/33.6.781

Casillas, A., Robbins, S., & Allen, J. (2012). Predicting Early Academic Failure in High School from Prior Academic Achievement, Psychosocial Characteristics, and Behavior. *Journal of Educational Psychology*, *104*(2), 407–420. https://doi.org/10.1037/a0027180

Catalano, R. F., Hawkins, J. D., Berglund, L., Pollard, J. A., & Arthur, M. W. (2002). Prevention science and positive youth development: Competitive or cooperative frameworks? *Journal of Adolescent Health*, *31*(Suppl6), 230–239. https://doi.org/10.1016/S1054-139X(02)00496-2

Collaborative for Academic, Social, and Emotional Learning & American Institutes for Research. (2013b). *Teacher rating of student social and emotional competencies*. Washington, DC, and Chicago, IL: Author.

Conduct Problems Prevention Research Group (1999). Initial impact of the Fast Track prevention trial for conduct problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology*, *67*(5), 648–657. https://doi-org.air.idm.oclc.org/10.1037/0022-006X.67.5.648

Conte, J. M., Ringenbach, K. L., Moran, S. K., & Landy, F. J. (2001). Criterion-validity evidence for time urgency: Associations with burnout, organizational commitment, and job involvement in travel agents. *Applied H.R.M. Research*, *6*(1–2), 129–134.

Dee, T. S. (2024). Higher chronic absenteeism threatens academic recovery from the COVID-19 pandemic. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *121*(3), 1–3. https://doi.org/10.1073/pnas.2312249121

Denham, S. A., & Brown, C. (2010). "Plays Nice with Others": Social-Emotional Learning and Academic Success. *Early Education and Development*, *21*(5), 652–680. https://doi.org/10.1080/10409289.2010.497450

DiPerna, J. C., & Elliott, S. N. (1999). Development and validation of the academic competence evaluation scales. *Journal of Psychoeducational Assessment*, *17*(3), 207–225. https://doi.org/10.1177/073428299901700302

Domitrovich, C. E., Li, Y., Mathis, E. T., & Greenberg, M. T. (2019). Individual and organizational factors associated with teacher self-reported implementation of the PATHS curriculum. *Journal of School Psychology*, *76*, 168–185. https://doi.org/10.1016/j.jsp.2019.07.015

Doyle, S. L., Jennings, P. A., Brown, J. L., Rasheed, D., DeWeese, A., Frank, J. L., ... & Greenberg, M. T. (2019). Exploring relationships between CARE program fidelity, quality, participant responsiveness, and uptake of mindful practices. *Mindfulness*, *10*(5), 841-853. https://doi.org/10.1007/s12671-018-1034-9

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents. *Psychological Science (0956-7976)*, *16*(12), 939–944. https://doi.org/10.1111/j.1467-9280.2005.01641.x

Durlak, J. A., Weissberg, R. P., & Dymnicki, A. B. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development*, *82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Elliot, A. J., & Dweck. C. S. (Eds). (2005). *Handbook of competence and motivation.* New York. Guilford Press.

Enders, C. K. (2025). Missing data: An update on the state of the art. *Psychological Methods, 30*(2), 322–339. https://doi.org/10.1037/met0000563

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling*, *8*(3), 430–457.

Greenberg, M. T., Brown J. L., Abenavoli, R. M. (2016). Teacher stress and health effects on teachers, students, and schools. (Issue Brief). Retrieved from Robert Wood Johnson Foundation website: http://www.rwjf.org/content/dam/farm/reports/issue_briefs/2016/rwjf430428

Greenberg, M. T., Kusché, C. A., Cook, E. T., & Quamma, J. P. (1995). Promoting emotional competence in school-aged children: The effects of the PATHS curriculum. *Development and psychopathology*, *7*(1), 117-136. https://doi.org/10.1017/S0954579400006374

Greenberg, M.T., Kusché, C. & Mihalic, S.F. (1998). *Blueprints for Violence Prevention, Book Ten: Promoting Alternative Thinking Strategies (PATHS)*. Boulder, CO: Center for the Study and Prevention of Violence.

Gresham, F. M., & Elliott, S. N. (2008). *Social Skills Improvement System: Rating Scales.* Bloomington, MN: Pearson Assessments.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*(2), 348–362. https://doi.org/10.1037/0022-3514.85.2.348

Hackman, D. A., Farah, M. J., & Meaney, M. J. (2010). Socioeconomic status and the brain: Mechanistic insights from human and animal research. *Nature Reviews Neuroscience*, *11*(9), 651–659. https://doi.org/10.1038/nrn2897

Hamre, B. K., Pianta, R. C. (2006). Student-teacher relationships. In G. G. Bear & K. M. Minke (eds.) *Children's needs III: Development, prevention, and intervention* (pp. 59–71). Bethesda, MD: National Association of School Psychologists.

Hawkins, J. D., Smith, B. H., & Catalano, R. F. (2004). Social development and social and emotional learning. In J. E. Zins, R. P. Weissberg, M. C. Wang, & H. J. Walberg (Eds.) *Building academic success on social and emotional learning. What does the research say?* (pp. 135–150). New York: Teachers College Press.

Hoglund, W. L. G., Klingle, K. E., & Hosan, N. E. (2015). Classroom risks and resources: Teacher burnout, classroom quality and children's adjustment in high needs elementary schools. *Journal of School Psychology*, *53*(5), 337–357. https://doi.org/10.1016/j.jsp.2015.06.002

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Iowa State University Extension and Outreach. (2010). *Strengthening families program: For parents and youth ages 10–14 facilitator delivery ratings*. http://www.episcenter.psu.edu/ebp/strengthening/sfpfidelityforms

James, S. (2025) The impact of Title I funding on local school programs. *Education Walkthrough*. https://educationwalkthrough.com/title-i-funding/

Jennings, P. A. (2016). CARE for Teachers: A mindfulness-based approach to promoting teachers' well-being and improving performance. In K. Schonert-Reichl & R. Roeser (Eds). *The handbook of mindfulness in education: Emerging theory, research, and programs* (p. 133-148). New York: Springer-Verlag.

Jennings, P. A., Frank, J. L., Doyle, S., Yoonkyung Oh, Rasheed, D., DeWeese, A., Cham, H., Brown, J. L., Davis, R., DeMauro, A. A., & Greenberg, M. T. (2017). Impacts of the CARE for Teachers Program on Teachers' Social and Emotional Competence and Classroom Interactions. *Journal of Educational Psychology*, *109*(7), 1010–1028. https://doi-org.air.idm.oclc.org/10.1037/edu0000187

Jennings, P. A**.** & Greenberg, M. (2009). The prosocial classroom: Teacher social and emotional competence in relation to child and classroom outcomes. *Review of Educational Research, 79*(1), 491–525. https://doi.org/10.3102/0034654308325693

Kroenke, K., String, T., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Kusché, C. A., & Greenberg, M. T. (2011). PATHS: An introduction to the curriculum. *PATHS Program Holding*. https://content.pathsprogram.com/paths-current-users/502005-D-4c2c-DP-OT-EG.pdf

Kusché, C. A., & Greenberg, M. T. (2012). The PATHS Curriculum: Promoting emotional literacy, prosocial behavior, and caring classrooms. In *Handbook of school violence and school safety: International research and practice, 2nd ed.* (pp. 435–446).

Lieberman, M. (2022). All teaching shortages are not equal: 4 takeaways from new research. *Education Week*. https://www.edweek.org/leadership/all-teaching-shortages-are-not-equal-4-takeaways-from-new-research/2022/12

Maslach, C., Jackson, S. E., Leiter, M. P., Zalaquett, C. P., & Wood, R. J. (1997). Maslach Burnout Inventory (3rd ed.). In C. P. Zalaquett & R. J. Wood (Eds.), *Evaluating stress: A book of resources* (pp. 191–218). Scarecrow Education.

Milkie, M. A., & Warner, C. H. (2011). Classroom Learning Environments and the Mental Health of First Grade Children. *Journal of Health and Social Behavior*, *52*(1), 4–22. https://doi.org/10.1177/0022146510394952

Muthén, L.K. and Muthén, B.O. (1998–2024). *Mplus User's Guide* (8th Ed.). Los Angeles, CA: Muthén & Muthén.

Oberle, E., & Schonert-Reichl, K. A. (2016). Stress contagion in the classroom? The link between classroom teacher burnout and morning cortisol in elementary school students. *Social Science & Medicine*, *159*, 30–37. https://doi.org/10.1016/j.socscimed.2016.04.031

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K–3*. Paul H. Brookes Publishing.

Riggs, N. R., Greenberg, M. T., Kusché, C. A., & Pentz, M. A. (2006). The Mediational Role of Neurocognition in the Behavioral Outcomes of a Social-Emotional Prevention Program in Elementary School Students: Effects of the PATHS Curriculum. *Prevention Science*, *7*(1), 91–102. https://doi.org/10.1007/s11121-005-0022-1

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Scales, P. C., Benson, P. L., & Roehlkepartain, E. C. (2006). The Role of Developmental Assets in Predicting Academic Achievement: A Longitudinal Study. *Journal of Adolescence*, *29*(5), 691–708. https://doi.org/10.1016/j.adolescence.2005.09.001

Schaps, E., Battistich, V., & Solomon, D. (2004). Community in school as key to student growth: Findings from the Child Development Project. In J. E. Zins, R. P. Weissberg, M. C. Wang, & H. J. Walberg (Eds.) *Building academic success on social and emotional learning. What does the research say?* (pp. 189–205). New York: Teachers College Press.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

StataCorp. (2024). *Stata Statistical Software: Release 18.5*. College Station, TX: StataCorp LLC.

Tareen, S. (2022, January 11). After five days of canceled classes, Chicago union leaders approve plan to return to school. *Associated Press*. https://www.pbs.org/newshour/nation/after-five-days-of-canceled-classes-chicago-union-leaders-approve-plan-to-return-to-school

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting Positive Youth Development through School-Based Social and Emotional Learning Interventions: A Meta-Analysis of Follow-Up Effects. *Child Development*, *88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*(4), 585–602. https://doi.org/10.1007/BF00937993

Wethington, E. (2000). Contagion of stress. In *Advances in group processes* (pp. 229-253). Emerald Group Publishing Limited.

Will, M. (2022, December 16). The teaching profession in 2022 (in charts). *Education Week*. https://www.edweek.org/teaching-learning/the-teaching-profession-in-2022-in-charts/2022/12

Weare, K., & Nind, M. (2011). Mental health promotion and problem prevention in schools: what does the evidence say? *Health Promotion International*, *26*(suppl_1), i29–i69. https://doi.org/10.1093/heapro/dar075

Wu, Y., Levis, B., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., Rice, D. B., Boruff, J., Cuijpers, P., Gilbody, S., Ioannidis, J. P. A., Kloda, L. A., McMillan, D., Patten, S. B., Shrier, I., Ziegelstein, R. C., Akena, D. H., Arroll, B., Ayalon, L., Baradaran, H. R., … Thombs, B. D. (2020). Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychological medicine*, *50*(8), 1368–1380. https://doi.org/10.1017/S0033291719001314

Appendix A. Supplemental Exhibits

Exhibit A1. Teacher Baseline Equivalence Comparisons

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
FFMQ – Overall	3.37(0.41)	3.48(0.47)	-0.12	0.08	-.26	.153
FFMQ – Observing	3.28(0.57)	3.39(0.66)	-0.11	0.11	-.18	.341
FFMQ – Describing	3.58(0.67)	3.67(0.72)	-0.09	0.13	-.13	.471
FFMQ – Acting with Awareness	3.38(0.68)	3.62(0.87)	-0.24	0.15	-.30	.104
FFMQ – Non-reactive	3.17(0.51)	3.24(0.52)	-0.08	0.10	-.15	.432
FFMQ – Non-judgmental	3.42(0.49)	3.50(0.57)	-0.08	0.10	-.14	.445
ERQ – Reappraisal	5.19(0.95)	5.31(0.92)	-0.12	0.18	-.13	.496
ERQ – Suppression	3.46(1.17)	3.77(1.00)	-0.31	0.20	-.28	.133
PHQ-8	1.68(0.71)	1.59(0.55)	0.09	0.12	.15	.431
GAD-7	1.85(0.75)	1.86(0.86)	-0.01	0.15	-.01	.973
MBI	3.80(1.49)	3.54(1.49)	0.26	0.28	.18	.350
PROMIS	3.27(1.01)	3.49(0.86)	-0.22	0.18	-.23	.220
TUS – Overall	3.16(0.49)	3.15(0.60)	0.00	0.10	.01	.971
TUS – Task-Related Hurry	3.45(0.94)	3.40(1.02)	0.05	0.19	.05	.806
TUS – Speech Patterns	2.85(0.70)	2.86(0.80)	-0.01	0.14	-.01	.958
TUS – General Hurry	3.17(0.67)	3.20(0.67)	-0.03	0.13	-.04	.833

Note. M = Mean; SD = standard deviation; SE = standard error. Effect sizes were calculated using Hedges' g.

Exhibit A2. Student Baseline Equivalence Comparisons

Measure	CARE+PATHS M(SD)	PATHS Only M(SD)	Mean Difference	SE	Effect Size	*p*-value
Attendance rate	0.911(0.118)	0.901(0.138)	0.01	0.01	.08	.400
Chronic Absenteeism	0.281(0.450)	0.300(0.458)	-0.02	0.18	-.04	.643
TOCA-R – Overall[1]	1.98(0.88)	1.91(0.69)	0.07	0.15	.20	.477
TOCA-R – Aggression	1.49(0.86)	1.28(0.41)	0.21	0.12	.56	.059
TOCA-R – Inattention/Hyperactivity	2.56(1.29)	2.66(1.25)	-0.10	0.27	-.09	.710
TRC	1.17(1.27)	1.26(1.23)	-0.09	0.54	-.32	.153
IAR – ELA	705.77(37.11)	699.22(33.28)	6.54	0.18	.08	.719
ACES – Reading Skills	2.26(1.11)	2.45(0.96)	-0.19	0.23	-.12	.638
IAR – Mathematics	705.58(33.56)	702.02(32.12)	3.57	.04	.21	.849
ACES – Mathematics Skills	2.34(1.00)	2.62(0.85)	-0.28	0.21	-.22	.333
ACES – Motivation	3.29(1.13)	3.41(1.08)	-0.12	0.19	-.20	.538
ACES – Engagement	3.70(1.10)	3.57(1.11)	.13	0.24	.06	.774
SSIS	3.17(0.65)	3.24(0.53)	-0.06	0.12	-.16	.525
TRSSEC	3.22(0.87)	3.18(0.79)	.04	0.16	.01	.975

Note. M = Mean; SD = standard deviation; SE = standard error. Effect sizes were calculated as standardized betas.
[1]Only four students had a reported disciplinary incident in the 2020-21 school year. Therefore, we used teacher-reported TOCA-R Overall scores as the baseline measure for disciplinary incident outcomes and for TOCA-R outcomes.

Exhibit A3. Teacher Outcome Group Means, Standard Deviations, and Sample Sizes Across Timepoints

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
Five Facet Mindfulness Questionnaire – Overall						
Full Sample	3.37(0.41) 54	3.48(0.47) 64	3.43(0.57) 32	3.65(0.50) 32	3.35(0.57) 24	3.71(0.35) 24
Baseline–Post-Test Sample	3.35(0.46) 27	3.48(0.50) 25	3.36(0.53) 27	3.65(0.54) 25	–	–
Baseline–Follow-Up Sample	3.37(0.37) 19	3.59(0.46) 20	–	–	3.32(0.61) 19	3.68(0.37) 20
Five Facet Mindfulness Questionnaire – Observing						
Full Sample	3.28(0.57) 54	3.39(0.66) 64	3.42(0.68) 32	3.48(0.77) 32	3.19(0.69) 24	3.78(0.57) 24
Baseline–Post-Test Sample	3.32(0.57) 27	3.48(0.55) 25	3.36(0.69) 27	3.52(0.64) 25	–	–
Baseline–Follow-Up Sample	3.23(0.71) 19	3.58(0.55) 20	–	–	3.29(0.70) 19	3.65(0.51) 20
Five Facet Mindfulness Questionnaire – Describing						
Full Sample	3.58(0.67) 54	3.67(0.72) 64	3.74(0.82) 32	3.92(0.76) 32	3.54(0.89) 24	3.71(0.61) 23
Baseline–Post-Test Sample	3.57(0.81) 27	3.61(0.69) 25	3.62(0.81) 27	3.81(0.76) 25	–	–
Baseline–Follow-Up Sample	3.51(0.69) 19	3.73(0.71) 20	–	–	3.51(0.97) 19	3.64(0.58) 20
Five Facet Mindfulness Questionnaire – Acting with Awareness						
Full Sample	3.38(0.68) 54	3.62(0.87) 64	3.35(0.76) 32	3.77(0.72) 32	3.21(1.03) 24	3.73(0.77) 24
Baseline–Post-Test Sample	3.41(0.67)	3.54(0.88)	3.25(0.61)	3.68(0.75)	–	–

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
	27	25	27	25		
Baseline–Follow-Up Sample	3.47(0.52) 19	3.65(0.80) 20	–	–	3.11(0.98) 19	3.67(0.73) 20
Five Facet Mindfulness Questionnaire – Non-Reactive						
Full Sample	3.17(0.51) 54	3.24(0.52) 64	3.22(0.65) 32	3.47(0.63) 32	3.55(0.91) 24	3.89(0.63) 24
Baseline–Post-Test Sample	3.08(0.50) 27	3.24(0.55) 25	3.20(0.63) 27	3.57(0.58) 25	–	–
Baseline–Follow-Up Sample	3.17(0.51) 19	3.39(0.52) 20	–	–	3.34(0.85) 19	3.98(0.58) 20
Five Facet Mindfulness Questionnaire – Non-Judgmental						
Full Sample	3.42(0.49) 54	3.50(0.57) 64	3.41(0.71) 32	3.63(0.69) 32	3.28(0.68) 24	3.41(0.53) 23
Baseline–Post-Test Sample	3.37(0.53) 27	3.53(0.60) 25	3.40(0.61) 27	3.65(0.68) 25	–	–
Baseline–Follow-Up Sample	3.46(0.54) 19	3.58(0.64) 20	–	–	3.37(0.67) 19	3.45(0.55) 20
Emotion Regulation Questionnaire – Reappraisal						
Full Sample	5.19(0.95) 52	5.31(0.92) 60	5.50(0.91) 30	5.18(1.27) 32	5.38(0.89) 23	5.67(0.79) 23
Baseline–Post-Test Sample	5.22(0.98) 25	5.35(1.08) 23	5.38(0.84) 25	5.03(1.32) 23	–	–
Baseline–Follow-Up Sample	5.38(0.75) 18	5.39(0.94) 19	–	–	5.36(0.92) 18	5.54(0.78) 19
Emotion Regulation Questionnaire – Suppression						
Full Sample	3.46(1.17) 52	3.77(1.00) 61	3.18(1.10) 30	3.39(1.01) 32	3.43(1.30) 23	3.64(1.41) 23

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
Baseline–Post-Test Sample	3.26(1.08) 25	3.67(0.85) 23	3.22(1.10) 25	3.39(1.13) 23	–	–
Baseline–Follow-Up Sample	3.72(1.01) 18	4.01(0.98) 19	–	–	3.43(1.28) 18	3.89(1.39) 19
Patient Health Questionnaire						
Full Sample	1.68(0.71) 52	1.59(0.55) 60	1.76(0.69) 30	1.40(0.43) 32	1.67(0.64) 22	1.38(0.44) 23
Baseline–Post-Test Sample	1.76(0.66) 25	1.54(0.56) 24	1.75(0.66) 25	1.36(0.41) 24	–	–
Baseline–Follow-Up Sample	1.65(0.83) 17	1.54(0.43) 18	–	–	1.77(0.67) 17	1.36(0.46) 18
Generalized Anxiety Disorder Scale						
Full Sample	1.85(0.75) 52	1.86(0.86) 60	1.92(0.70) 30	1.67(0.73) 32	1.86(0.79) 22	1.61(0.66) 23
Baseline–Post-Test Sample	2.03(0.80) 25	1.86(0.89) 24	1.97(0.66) 25	1.66(0.76) 24	–	–
Baseline–Follow-Up Sample	1.81(0.85) 17	1.72(0.69) 18	–	–	1.99(0.82) 17	1.56(0.66) 18
Maslach Burnout Inventory						
Full Sample	3.80(1.49) 52	3.54(1.49) 62	4.23(1.62) 30	3.65(1.46) 32	4.22(1.94) 23	3.40(1.57) 23
Baseline–Post-Test Sample	4.16(1.37) 25	3.40(1.59) 24	4.26(1.56) 25	3.66(1.61) 23	–	–
Baseline–Follow-Up Sample	4.05(1.53) 18	3.33(1.60) 19	–	–	4.48(2.01) 18	3.25(1.62) 19
Patient-Reported Outcomes Measurement Information System Sleep Disturbance Questionnaire						
Full Sample	3.27(1.01)	3.49(0.86)	3.57(0.72)	3.74(0.85)	3.62(0.73)	3.75(0.73)

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
	52	61	30	33	21	23
Baseline–Post-Test Sample	3.11(1.01) 25	3.63(1.01) 24	3.63(0.56) 25	3.76(0.94) 24	–	–
Baseline–Follow-Up Sample	3.48(1.17) 16	4.58(0.85) 18	–	–	3.48(0.71) 16	3.67(0.91) 18
Time Urgency Scale – Overall						
Full Sample	3.16(0.49) 52	3.15(0.60) 62	3.17(0.55) 30	3.15(0.56) 32	3.19(0.64) 23	3.17(0.58) 23
Baseline–Post-Test Sample	3.18(0.56) 25	3.11(0.71) 24	3.21(0.56) 25	3.14(0.64) 24	–	–
Baseline–Follow-Up Sample	3.27(0.53) 18	3.14(0.69) 19	–	–	3.34(0.57) 18	3.20(0.63) 19
Time Urgency Scale – Task-Related Hurry						
Full Sample	3.45(0.94) 52	3.40(1.02) 62	3.59(0.96) 30	3.46(1.02) 32	3.57(1.14) 23	3.59(0.97) 23
Baseline–Post-Test Sample	3.60(0.87) 25	3.56(1.09) 24	3.64(0.88) 25	3.44(1.13) 24	–	–
Baseline–Follow-Up Sample	3.59(0.81) 18	3.51(1.02) 19	–	–	3.80(1.00) 18	3.60(1.06) 19
Time Urgency Scale – Speech Patterns						
Full Sample	2.85(0.70) 52	2.86(0.80) 62	2.90(0.81) 30	2.62(0.73) 32	2.80(0.97) 23	2.67(0.77) 23
Baseline–Post-Test Sample	2.88(0.86) 25	2.66(0.83) 24	2.96(0.84) 25	2.65(0.73) 24	–	–
Baseline–Follow-Up Sample	2.92(0.73) 18	2.71(0.73) 19	–	–	2.99(0.91) 18	2.79(0.70) 19
Time Urgency Scale – General Hurry						

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
Full Sample	3.17(0.67) 52	3.20(0.67) 62	3.01(0.72) 30	3.36(0.56) 32	3.20(0.64) 23	3.25(0.64) 23
Baseline–Post-Test Sample	3.07(0.78) 25	3.11(0.72) 24	3.03(0.73) 25	3.34(0.56) 24	–	–
Baseline–Follow-Up Sample	3.30(0.73) 18	3.21(0.84) 19	–	–	3.22(0.72) 18	3.22(0.65) 19

Note. M = mean; SD = standard deviation; N = sample size. Rows marked "Full Sample" reflect all data available from participants at each time point. Rows labeled "Baseline–Post-Test Sample" reflect data only from participants with data available from both the baseline and post-test timepoints. Rows labeled "Baseline–Follow-Up Sample" reflect data only from participants with data available from both the baseline and follow-up timepoints.

Exhibit A4. Student Outcome Group Means, Standard Deviations, and Sample Sizes Across Timepoints

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
Attendance Rate						
Full Sample	0.911(0.118) 1631	0.901(0.138) 1482	0.880(0.109) 1889	0.872(0.116) 1727	0.915(0.079) 1555	0.902(0.092) 1440
Baseline–Post-Test Sample	0.916(0.113) 1512	0.905(0.136) 1373	0.888(0.102) 1512	0.881(0.107) 1373	–	–
Baseline–Follow-Up Sample	0.924(0.106) 1280	0.911(0.129) 1208	–	–	0.919(0.071) 1280	0.906(0.091) 1208
Chronic Absenteeism						
Full Sample	0.281(0.450) 1631	0.300(0.458) 1482	0.447(0.497) 1889	0.488(0.500) 1727	0.309(0.462) 1555	0.356(0.479) 1440
Baseline–Post-Test Sample	0.270(0.444) 1512	0.288(0.453) 1373	0.413(0.493) 1512	0.457(0.498) 1373	–	–
Baseline–Follow-Up Sample	0.240(0.427) 1280	0.272(0.445) 1208	–	–	0.285(0.452) 1280	0.333(0.471) 1208
Number of Disciplinary Incidents[1]						
Full Sample	1.98(0.88) 60	1.91(0.69) 115	0.019(0.170) 2024	0.109(0.560) 1842	0.095(0.531) 2024	0.131(0.684) 1842
Baseline–Post-Test Sample	1.98(0.88) 60	1.91(0.69) 115	0.000(0.000) 60	0.104(0.466) 115	–	–
Baseline–Follow-Up Sample	1.98(0.88) 60	1.91(0.69) 115	–	–	0.233(0.945) 60	0.070(0.413) 115
One or More Disciplinary Incidents[1]						
Full Sample	1.98(0.88) 60	1.91(0.69) 115	0.016(0.125) 2024	0.062(0.242) 1842	0.057(0.232) 2024	0.060(0.238) 1842
Baseline–Post-Test Sample	1.98(0.88)	1.91(0.69)	0.000(0.000)	0.070(0.256)	–	–

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
	60	115	60	115		
Baseline–Follow-Up Sample	1.98(0.88) 60	1.91(0.69) 115	–	–	0.100(0.303) 60	0.035(0.184) 115
Teacher Observation of Classroom Adaptation–Revised – Overall						
Full Sample	1.98(0.88) 60	1.91(0.69) 115	1.96(0.96) 116	1.87(0.81) 121	2.15(1.03) 40	1.91(1.05) 18
Baseline–Post-Test Sample	2.00(0.93) 51	1.87(0.69) 78	2.01(1.12) 51	1.87(0.78) 78	–	–
Baseline–Follow-Up Sample	2.04(1.01) 16	1.85(0.59) 7	–	–	2.58(1.17) 16	1.53(0.84) 7
Teacher Observation of Classroom Adaptation–Revised – Aggression						
Full Sample	1.49(0.86) 60	1.28(0.41) 115	1.58(0.87) 115	1.51(0.69) 121	1.71(0.94) 40	1.64(0.88) 18
Baseline–Post-Test Sample	1.52(0.91) 51	1.27(0.38) 78	1.73(1.06) 51	1.42(0.56) 78	–	–
Baseline–Follow-Up Sample	1.46(1.24) 16	1.14(0.27) 7	–	–	2.20(1.27) 16	1.41(0.60) 7
Teacher Observation of Classroom Adaptation–Revised – Inattention/Hyperactivity						
Full Sample	2.56(1.29) 60	2.66(1.25) 116	2.41(1.39) 116	2.31(1.24) 121	2.67(1.41) 40	2.36(1.44) 18
Baseline–Post-Test Sample	2.58(1.33) 51	2.58(1.23) 78	2.32(1.43) 51	2.40(1.27) 78	–	–
Baseline–Follow-Up Sample	2.73(1.30) 16	2.67(1.03) 7	–	–	3.02(1.34) 16	1.67(1.16) 7
mCLASS Text Reading and Comprehension[2]						
Full Sample	1.17(1.27) 316	1.26(1.23) 470	0.99(1.25) 333	1.45(1.31) 342	–	–

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
Baseline–Post-Test Sample	0.94 (1.17) 157	1.20(1.19) 229	1.30(1.33) 157	1.62(1.31) 229	–	–
Illinois Assessment of Readiness – English Language Arts						
Full Sample	705.77(37.11) 202	699.22(33.28) 178	706.57(34.63) 787	709.42(33.88) 670	715.11(34.39) 1091	716.49(34.02) 963
Baseline–Post-Test Sample	707.95(38.19) 175	700.01(33.85) 145	722.24(33.15) 175	714.46(31.06) 145	–	–
Baseline–Follow-Up Sample	710.90(38.31) 154	700.47(33.98) 132	–	–	732.49(31.93) 154	721.76(30.40) 132
Academic Competence Evaluation Scale – Reading Skills						
Full Sample	2.26(1.11) 69	2.45(0.96) 115	2.56(1.06) 115	2.91(1.13) 121	2.63(1.06) 38	3.22(1.09) 23
Baseline–Post-Test Sample	2.37(1.18) 51	2.61(1.00) 72	2.53(1.09) 51	2.78(1.12) 72	–	–
Baseline–Follow-Up Sample	2.28(1.35) 12	2.69(1.14) 13	–	–	2.69(1.18) 12	3.21(1.03) 13
Illinois Assessment of Readiness – Mathematics						
Full Sample	705.58(33.56) 213	702.02(32.12) 175	706.06(30.96) 782	707.74(27.46) 661	712.86(30.95) 1088	711.37(30.06) 965
Baseline–Post-Test Sample	709.59(33.45) 175	703.20(30.43) 145	711.61(29.34) 175	709.72(25.35) 145	–	–
Baseline–Follow-Up Sample	712.16(33.26) 154	702.68(30.93) 132	–	–	721.58(28.75) 154	710.07(26.52) 132
Academic Competence Evaluation Scale – Mathematics Skills						
Full Sample	2.34(1.00) 69	2.62(0.85) 108	2.66(0.90) 115	3.11(0.99) 121	2.74(1.12) 35	3.29(1.01) 22
Baseline–Post-Test Sample	2.43(1.04)	2.79(0.83)	2.61(0.92)	3.09(0.95)	–	–

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
	51	72	51	72		
Baseline–Follow-Up Sample	2.31(1.39) 12	2.92(0.85) 13	–	–	2.53(1.14) 12	3.36(1.00) 13
Academic Competence Evaluation Scale – Motivation						
Full Sample	3.29(1.13) 69	3.41(1.08) 114	3.51(1.23) 114	3.72(1.09) 118	3.35(1.26) 39	3.70(1.25) 23
Baseline–Post-Test Sample	3.42(1.12) 51	3.46(1.00) 72	3.56(1.31) 51	3.57(1.13) 72	–	–
Baseline–Follow-Up Sample	3.13(1.24) 16	3.28(1.16) 13	–	–	3.13(1.39) 16	3.59(1.07) 13
Academic Competence Evaluation Scale – Engagement						
Full Sample	3.70(1.10) 69	3.57(1.11) 114	3.79(1.14) 115	3.94(0.98) 119	3.58(1.22) 39	3.64(1.08) 23
Baseline–Post-Test Sample	3.80(1.13) 51	3.71(1.10) 72	3.92(1.10) 51	3.82(1.04) 72	–	–
Baseline–Follow-Up Sample	3.54(0.97) 16	3.36(1.13) 13	–	–	3.54(1.16) 16	3.49(0.93) 13
Social Skills Improvement System						
Full Sample	3.17(0.65) 69	3.24(0.53) 114	3.27(0.64) 115	3.37(0.60) 119	3.09(0.60) 39	3.26(0.62) 23
Baseline–Post-Test Sample	3.23(0.66) 51	3.31(0.52) 79	3.28(0.71) 51	3.36(0.59) 79	–	–
Baseline–Follow-Up Sample	3.19(0.77) 16	3.10(0.56) 13	–	–	3.03(0.73) 16	3.23(0.53) 13
Teacher Rating of Student Social and Emotional Competencies						
Full Sample	3.22(0.87) 69	3.18(0.79) 112	3.20(0.89) 115	3.26(0.92) 120	3.13(0.87) 38	3.21(0.94) 23

Sample	Baseline		Post-Test		Follow-up	
	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N	CARE+PATHS M(SD) N	PATHS Only M(SD) N
Baseline–Post-Test Sample	3.22(0.93) 51	3.22(0.83) 76	3.16(0.96) 51	3.19(0.97) 76	–	–
Baseline–Follow-Up Sample	2.96(1.05) 16	3.03(0.85) 13	–	–	3.04(0.95) 16	3.46(0.80) 13

Note. M = mean; SD = standard deviation; N = sample size. Rows marked "Full Sample" reflect all data available from participants at each time point. Rows labeled "Baseline–Post-Test Sample" reflect data only from participants with data available from both the baseline and post-test timepoints. Rows labeled "Baseline–Follow-Up Sample" reflect data only from participants with data available from both the baseline and follow-up timepoints.

[1]Only four students had a reported disciplinary incident in the 2020-21 school year. Therefore, we used teacher-reported Teacher Observation of Classroom Adaptation–Revised scores as the baseline measure for disciplinary incident outcomes.

[2]Chicago Public Schools stopped using the mCLASS Text Reading and Comprehension (TRC) in SY2022-23 and did not report a new formative assessment in the administrative records for that year, so we do not have TRC follow-up data.

Exhibit A5. Teacher Survey Outcome Bivariate Correlations

	Baseline Score	Post-Test Score	Follow-Up Score	Intervention Status	Gender	Years of Experience
Five Facet Mindfulness Questionnaire – Overall						
Baseline Score	–	.71**	.63**	-.13	-.08	.17†
Post-Test Score	–	–	.75**	-.21	-.14	.30*
Follow-Up Score	–	–	–	-.36*	.13	.29†
Number	118	64	48	156	130	111
Mean	3.43	3.54	3.53	0.50	1.08	15.77
Standard Deviation	0.45	0.55	0.50	0.50	0.31	9.66
Five Facet Mindfulness Questionnaire – Observing						
Baseline Score	–	.48**	.42**	-.09	-.06	-.04
Post-Test Score	–	–	.59**	-.04	-.04	.15
Follow-Up Score	–	–	–	-.43**	.10	.04
Number	118	64	48	156	130	111
Mean	3.34	3.45	3.49	0.50	1.08	15.77
Standard Deviation	0.62	0.72	0.69	0.50	0.31	9.66
Five Facet Mindfulness Questionnaire – Describing						
Baseline Score	–	.72**	.70**	-.07	-.06	.13
Post-Test Score	–	–	.75**	-.11	-.26†	.30*
Follow-Up Score	–	–	–	-.11	.09	.26
Number	118	64	47	156	130	111
Mean	3.63	3.83	3.63	0.50	1.08	15.77
Standard Deviation	0.70	0.79	0.76	0.50	0.31	9.66
Five Facet Mindfulness Questionnaire – Acting With Awareness						

	Baseline Score	Post-Test Score	Follow-Up Score	Intervention Status	Gender	Years of Experience
Baseline Score	–	.54**	.66**	-.15	.00	.16
Post-Test Score	–	–	.63**	-.27*	-.07	.23
Follow-Up Score	–	–	–	-.28	-.08	.37*
Number	118	64	48	156	130	111
Mean	3.51	3.56	3.47	0.50	1.08	15.77
Standard Deviation	0.80	0.76	0.94	0.50	0.31	9.66
Five Facet Mindfulness Questionnaire – Non-Reactive						
Baseline Score	–	.65**	.47**	-.07	.02	.10
Post-Test Score	–	–	.62**	-.20	.00	.14
Follow-Up Score	–	–	–	-.22	.11	.23
Number	118	64	48	156	130	111
Mean	3.21	3.34	3.72	0.50	1.08	15.77
Standard Deviation	0.52	0.65	0.80	0.50	0.31	9.66
Five Facet Mindfulness Questionnaire – Non-Judgmental						
Baseline Score	–	.68**	.26	-.07	-.22*	.25**
Post-Test Score	–	–	.13	-.16	-.18	.38**
Follow-Up Score	–	–	–	-.11	.28†	.02
Number	118	64	47	156	130	111
Mean	3.46	3.52	3.34	0.50	1.08	15.77
Standard Deviation	0.53	0.70	0.61	0.50	0.31	9.66
Emotion Regulation Questionnaire – Reappraisal						
Baseline Score	–	.53**	.41*	-.07	.00	.14
Post-Test Score	–	–	.29	.15	-.01	.17

	Baseline Score	Post-Test Score	Follow-Up Score	Intervention Status	Gender	Years of Experience
Follow-Up Score	–	–	–	-.17	-.21	.10
Number	112	62	46	156	130	111
Mean	5.25	5.33	5.52	0.50	1.08	15.77
Standard Deviation	0.93	1.11	0.84	0.50	0.31	9.66
Emotion Regulation Questionnaire – Suppression						
Baseline Score	–	.58**	.60**	-.14	.04	.18†
Post-Test Score	–	–	.73**	-.10	-.01	.10
Follow-Up Score	–	–	–	-.08	.44**	.13
Number	113	62	46	156	130	111
Mean	3.63	3.29	3.54	0.50	1.08	15.77
Standard Deviation	1.09	1.05	1.34	0.50	0.31	9.66
Patient Health Questionnaire						
Baseline Score	–	.67**	.64**	.08	.00	-.11
Post-Test Score	–	–	.57**	.30*	.24	-.35*
Follow-Up Score	–	–	–	.26†	.16	-.32†
Number	112	62	45	156	130	111
Mean	1.63	1.58	1.52	0.50	1.08	15.77
Standard Deviation	0.63	0.59	0.50	0.50	0.31	9.66
Generalized Anxiety Disorder Scale						
Baseline Score	–	.74**	.68**	.00	.13	-.16
Post-Test Score	–	–	.74**	.17	.13	-.32*
Follow-Up Score	–	–	–	.17	.20	-.29†
Number	112	62	45	156	130	111

	Baseline Score	Post-Test Score	Follow-Up Score	Intervention Status	Gender	Years of Experience
Mean	1.86	1.79	1.73	0.50	1.08	15.77
Standard Deviation	0.81	0.72	0.73	0.50	0.31	9.66
Maslach Burnout Inventory						
Baseline Score	–	.70**	.76**	.09	-.05	.00
Post-Test Score	–	–	.74**	.19	.08	-.18
Follow-Up Score	–	–	–	.23	.03	-.11
Number	114	62	46	156	130	111
Mean	3.66	3.93	3.81	0.50	1.08	15.77
Standard Deviation	1.49	1.55	1.79	0.50	0.31	9.66
Patient-Reported Outcomes Measurement Information System Sleep Disturbance Questionnaire						
Baseline Score	–	.49**	.33†	-.12	.07	.07
Post-Test Score	–	–	.58**	-.11	.07	-.05
Follow-Up Score	–	–	–	.02	-.21	-.12
Number	113	63	44	156	130	111
Mean	3.39	3.66	3.69	0.50	1.08	15.77
Standard Deviation	0.94	0.79	0.80	0.50	0.31	9.66
Time Urgency Scale – Overall						
Baseline Score	–	.83**	.82**	.00	-.13	-.04
Post-Test Score	–	–	.80**	.02	-.08	-.16
Follow-Up Score	–	–	–	-.08	-.11	.03
Number	114	62	46	156	130	111
Mean	3.16	3.16	3.18	0.50	1.08	15.77
Standard Deviation	0.55	0.55	0.60	0.50	0.31	9.66

	Baseline Score	Post-Test Score	Follow-Up Score	Intervention Status	Gender	Years of Experience
Time Urgency Scale – Task-Related Hurry						
Baseline Score	–	.74**	.75**	.02	-.04	.02
Post-Test Score	–	–	.80**	.07	-.05	-.07
Follow-Up Score	–	–	–	-.01	-.11	-.05
Number	114	62	46	156	130	111
Mean	3.42	3.52	3.58	0.50	1.08	15.77
Standard Deviation	0.98	0.98	1.05	0.50	0.31	9.66
Time Urgency Scale – Speech Patterns						
Baseline Score	–	.83**	.75**	.00	-.09	-.12
Post-Test Score	–	–	.83**	.18	-.03	-.31*
Follow-Up Score	–	–	–	.08	-.12	-.32†
Number	114	62	46	156	130	111
Mean	2.86	2.75	2.73	0.50	1.08	15.77
Standard Deviation	0.75	0.78	0.87	0.50	0.31	9.66
Time Urgency Scale – General Hurry						
Baseline Score	–	.71**	.81**	-.02	-.16†	.02
Post-Test Score	–	–	.65**	-.27*	-.10	.04
Follow-Up Score	–	–	–	-.04	-.26	.15
Number	114	62	46	156	130	111
Mean	3.18	3.19	3.22	0.50	1.08	15.77
Standard Deviation	0.67	0.66	0.63	0.50	0.31	9.66
Number of PATHS Lessons Taught In The Past Month[1]						
Post-Test Score	–	–	.57*	-.05	-.10	.16

	Baseline Score	Post-Test Score	Follow-Up Score	Intervention Status	Gender	Years of Experience
Follow-Up Score	–	–	–	-.15	-.01	-.06
Number	–	54	34	156	130	111
Mean	–	9.13	10.29	0.50	1.08	15.77
Standard Deviation	–	10.55	10.03	0.50	0.31	9.66
Frequency of PATHS Lessons Taught[1]						
Post-Test Score	–	–	.58**	-.09	.04	.07
Follow-Up Score	–	–	–	.10	.05	-.11
Number	–	55	34	156	130	111
Mean	–	3.76	3.65	0.50	1.08	15.77
Standard Deviation	–	1.07	1.25	0.50	0.31	9.66
Self-Reported Paths Implementation Quality[1]						
Post-Test Score	–	–	.68**	-.05	-.11	.19
Follow-Up Score	–	–	–	-.21	.13	.19
Number	–	55	34	156	130	111
Mean	–	4.05	4.29	0.50	1.08	15.77
Standard Deviation	–	0.91	0.71	0.50	0.31	9.66

Note. *$p < .05$ **$p < .01$ †$p < .10$; Intervention status (0 = PATHS Only, 1 = CARE+PATHS); Gender (0 = male, 1 = female)

[1]Questions about teachers' PATHS implementation were only asked at the post-test and follow-up timepoints because they had not yet received PATHS training at the baseline timepoint.

Exhibit A6. CLASS Outcome Bivariate Correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Emotional Support	–															
2. Class Org	.60**	–														
3. Instruct Support	.39**	.30**	–													
4. Positive Climate	.87**	.54**	.27*	–												
5. Negative Climate	-.64**	-.58**	-.02	-.52**	–											
6. Teacher Sensitivity	.85**	.58**	.35**	.61**	-.48**	–										
7. Reg Stu Pers	.79**	.29**	.50**	.57**	-.31**	.56**	–									
8. Behavior Manage	.43**	.85**	.08	.46**	-.58**	.37**	.08	–								
9. Productivity	.46**	.88**	.07	.39**	-.52**	.50**	.14	.71**	–							
10. Instruct Lrn Frmts	.59**	.70**	.61**	.48**	-.29**	.56**	.50**	.33**	.41**	–						
11. Concept Dev	.34**	.16	.85**	.18†	.00	.29**	.51**	-.01	.01	.42**	–					
12. Qual of Feedback	.36**	.37**	.90**	.28**	-.06	.36**	.37**	.17	.15	.61**	.63**	–				
13. Lang Modeling	.35**	.23*	.91**	.25*	.02	.28**	.47**	.01	.02	.57**	.69**	.71**	–			
14. Inter	.16	.16	.12	.06	-.09	.14	.22*	.05	.20†	.14	.20†	.01	.13	–		
15. Gender	.10	-.03	.07	-.05	-.03	.03	.30*	-.15	.00	.10	.02	.05	.09	-.14	–	
16. Yrs Exp	-.01	.20	-.17	.05	-.25†	-.06	-.15	.41**	.22†	-.20	-.18	-.11	-.19	-.10	-.14	–
Number	88	88	88	88	88	88	88	88	88	88	88	88	88	156	130	111
Mean	5.25	5.35	2.92	5.27	1.28	5.05	3.98	5.67	5.80	4.58	2.53	3.30	2.92	0.50	1.08	15.77
SD	0.78	0.92	0.94	1.15	0.55	1.13	1.03	1.17	1.15	1.06	0.88	1.20	1.09	0.50	0.31	9.66

Note. *$p < .05$ **$p < .01$ †$p < .10$; Class Org = Classroom organization; Instruct Support = Instructional support; Reg Stu Pers = Regard for student perspectives; Behavior Manage = Behavior management; Instruct Lrn Frmts = Instructional learning formats; Concept Dev = Concept development; Qual of Feedback = Quality of feedback; Lang Modeling = Language modeling; Inter = Intervention status (0 = PATHS Only, 1 = CARE+PATHS); Gender (0 = male, 1 = female); Yrs Exp = Teacher years of experience.

Exhibit A7. PATHS Observation Outcome Bivariate Correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Global Rating	–													
2. Model & Gen	.84**	–												
3. Tch Style & Behav	.71**	.66**	–											
4. Teach Concepts	.67**	.53**	.65**	–										
5. PATHS Language	.82**	.88**	.60**	.55**	–									
6. Kid of the Day	.44**	.75**	.29*	.18	.52**	–								
7. PATHS Concepts	.79**	.90**	.56**	.43**	.79**	.51**	–							
8. Student SEC	.76**	.76**	.80**	.64**	.58**	.34**	.66**	–						
9. Open to Consult	.60**	.55**	.83**	.55**	.57**	.28*	.41**	.61**	–					
10. Interpers Style	.60**	.60**	.89**	.44**	.54**	.27*	.55**	.68**	.65**	–				
11. Mng & Disc	.66**	.54**	.86**	.70**	.44**	.19	.44**	.78**	.56**	.69**	–			
12. Inter	-.14	-.12	-.03	-.12	-.18	-.05	-.14	-.04	-.06	.01	.02	–		
13. Gender	.06	-.01	-.01	.14	-.03	.01	-.08	.07	.02	-.10	.00	-.14	–	
14. Yrs Exp	.25	.24	.19	.08	.10	.09	.30†	.35*	.06	.08	.12	-.10	-.14	–
Number	65	65	63	65	65	65	65	65	65	65	62	156	130	111
Mean	4.63	2.73	3.53	3.37	2.63	2.75	2.42	3.14	3.45	3.40	3.47	0.50	1.08	15.77
SD	1.38	0.93	0.69	0.96	1.11	1.26	1.18	0.98	0.81	0.90	0.92	0.50	0.31	9.66

Note. *$p < .05$ **$p < .01$ †$p < .10$; Model & Gen = PATHS modeling and generalization; Tch Style & Behav = PATHS teaching style and behavior; Teach Concepts = Teaching of PATHS concepts; Student SEC = Student social-emotional competence; Open to Consult = Teacher openness to consultation; Interpers Style = Teacher interpersonal style; Mng & Disc = teacher classroom management and discipline skills; Inter = Intervention status (0 = PATHS Only, 1 = CARE+PATHS); Gender (0 = male, 1 = female); Yrs Exp = Teacher years of experience.

Exhibit A8. Student Model Bivariate Correlations

	Base-line	Post-Test	Follow-Up	Inter	Grade	Stu Gen	FRPL	Spec Ed	LEP	Black	Hisp	White	Teach Gen	Yrs Exp
Attendance Rate														
Baseline	–	.46**	.37**	.04*	.09**	.03	-.19**	-.05*	.12**	-.18**	.13**	.07**	.03	.10**
Post-Test	–	–	.58**	.04*	.05**	.02	-.16**	-.03†	.07**	-.12**	.06**	.11**	.00	.08**
Follow-Up	–	–	–	.07**	.02	.02	-.14**	-.03	.05**	-.06**	.02	.06**	-.04†	.07**
Number	3113	3616	2995	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	0.91	0.88	0.91	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.13	0.11	0.09	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Chronic Absenteeism														
Baseline	–	.41**	.34**	-0.02	-.07**	-0.03	.20**	.07**	-.12**	.20**	-.14**	-.09**	-0.03	-.12**
Post-Test	–	–	.51**	-.04*	-.05**	-0.03†	.18**	.06**	-.05**	.13**	-.05**	-.13**	-0.01	-.08**
Follow-Up	–	–	–	-.05**	-0.03	0.00	.13**	0.03†	-0.03†	.06**	-0.01	-.07**	.04*	-.09**
Number	3113	3616	2995	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	0.29	0.47	0.33	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.45	0.50	0.47	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Number of Disciplinary Incidents[1]														
Baseline	–	0.07	.17*	0.04	0.12	-0.13†	.17*	-0.01	-.16*	.22**	-0.12	-0.04	-0.07	-.33**
Post-Test	–	–	.29**	-.11**	.08**	-.05**	.04*	.05**	-.06**	.09**	-.07**	-0.03†	-0.02	0.00
Follow-Up	–	–	–	-0.03†	.08**	-.04*	-0.01	0.03†	-.06**	.07**	-.06**	-0.01	0.03	-0.01
Number	175	3866	3866	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	1.94	0.06	0.11	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.76	0.41	0.61	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
One or More Disciplinary Incidents[1]														
Baseline	–	0.02	.21**	0.04	0.12	-0.13†	.17*	-0.01	-.16*	.22**	-0.12	-0.04	-0.07	-.33**
Post-Test	–	–	.23**	-.12**	.10**	-.05**	.04**	0.03	-.08**	.11**	-.08**	-.03*	-0.02	0.01

	Base-line	Post-Test	Follow-Up	Inter	Grade	Stu Gen	FRPL	Spec Ed	LEP	Black	Hisp	White	Teach Gen	Yrs Exp
Follow-Up	–	–	–	-0.01	.10**	-.05**	0.02	.04*	-.08**	.10**	-.09**	0.01	.06**	-0.02
Number	175	3866	3866	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	1.94	0.04	0.06	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.76	0.19	0.23	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Teacher Observation of Classroom Adaptation–Revised – Overall														
Baseline	–	.82**	.51*	0.04	0.12	-0.13†	.17*	-0.01	-.16*	.22**	-0.12	-0.04	-0.07	-.33**
Post-Test	–	–	.63**	0.05	-0.01	-.21**	.14*	.15*	-0.11†	.22**	-.14*	-0.09	-0.11	-.32**
Follow-Up	–	–	–	0.10	-0.18	-0.07	0.04	-0.19	-0.20	0.22†	-0.15	-0.14	–[2]	-.38*
Number	175	237	58	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	1.94	1.92	2.08	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.76	0.89	1.03	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Teacher Observation of Classroom Adaptation–Revised – Aggression														
Baseline	–	.75**	.56**	.17*	-0.05	-0.09	0.06	-0.11	-0.12	0.12	-0.08	-0.01	-0.06	-.25**
Post-Test	–	–	.61**	0.04	-0.03	-0.07	0.07	0.06	-.17*	.25**	-.18**	-0.08	-0.01	-.20**
Follow-Up	–	–	–	0.04	-0.13	0.04	0.01	-0.18	-.26*	0.24†	-0.19	-0.10	–[2]	-.49**
Number	175	236	58	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	1.35	1.54	1.69	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.61	0.78	0.91	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Teacher Observation of Classroom Adaptation–Revised – Inattention/Hyperactivity														
Baseline	–	.83**	0.31	-0.04	.18*	-0.12	.18*	0.05	-0.14†	.22**	-0.11	-0.05	-0.07	-.30**
Post-Test	–	–	.54**	0.04	0.00	-.27**	.15*	.18**	-0.05	.15*	-0.08	-0.07	-.16*	-.33**
Follow-Up	–	–	–	0.11	-0.19	-0.15	0.06	-0.09	-0.11	0.14	-0.08	-0.15	–[2]	-0.27†
Number	176	236	60	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	2.63	2.36	2.57	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	1.26	1.32	1.42	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38

	Base-line	Post-Test	Follow-Up	Inter	Grade	Stu Gen	FRPL	Spec Ed	LEP	Black	Hisp	White	Teach Gen	Yrs Exp
mCLASS Text Reading and Comprehension[3]														
Baseline	–	.76**	–	-0.04	.17**	-0.03	-.33**	-.28**	-0.05	-.12**	-0.02	.20**	0.06	.29**
Post-Test	–	–	–	-.18**	-0.02	-0.02	-.34**	-.23**	-.20**	0.06	-.19**	.17**	.09*	.21**
Number	786	675	–	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	1.23	1.23	–	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	1.24	1.30	–	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Illinois Assessment of Readiness – English Language Arts														
Baseline	–	.70**	.68**	0.09†	–[4]	0.10†	-.31**	-.27**	-.24**	-.12*	-0.02	.23**	-0.12†	0.09
Post-Test	–	–	.74**	-0.04	.07**	.09**	-.31**	-.27**	-.06*	-.26**	.15**	.21**	-0.01	.12**
Follow-Up	–	–	–	-0.02	.17**	.09**	-.26**	-.26**	-.07**	-.21**	.12**	.18**	0.00	.09**
Number	380	1457	2054	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	702.7	707.88	715.57	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	34.48	34.30	34.22	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Academic Competence Evaluation Scale – Reading Skills														
Baseline	–	.84**	.77**	-0.09	0.07	0.08	-.29**	-.19*	-0.06	-0.12	-0.02	0.14†	.20*	.23*
Post-Test	–	–	.68**	-.16*	0.02	0.05	-.35**	-.24**	0.02	-.17**	0.07	.17**	.21**	.37**
Follow-Up	–	–	–	-.26*	-0.03	0.12	-0.25†	0.10	0.08	-.29*	0.23†	0.02	–[2]	.46**
Number	184	236	61	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	2.38	2.74	2.85	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	1.02	1.11	1.10	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Illinois Assessment of Readiness – Mathematics														
Baseline	–	.71**	.72**	0.05	–[4]	-0.02	-.36**	-.20**	0.04	-.37**	.20**	.28**	-.17**	.19**
Post-Test	–	–	.77**	-0.03	-.14**	-0.04	-.34**	-.22**	0.02	-.35**	.21**	.26**	-.07*	.16**
Follow-Up	–	–	–	0.02	-.06**	-.05*	-.30**	-.21**	0.01	-.28**	.16**	.23**	-.09**	.10**
Number	388	1443	2053	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658

	Base-line	Post-Test	Follow-Up	Inter	Grade	Stu Gen	FRPL	Spec Ed	LEP	Black	Hisp	White	Teach Gen	Yrs Exp
Mean	703.97	706.83	712.16	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	32.93	29.41	30.54	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Academic Competence Evaluation Scale – Mathematics Skills														
Baseline	–	.78**	.68**	-.15*	-0.01	0.07	-.28**	-.15*	-0.01	-.16*	-0.01	0.13†	.21*	.26**
Post-Test	–	–	.67**	-.23**	-0.05	0.00	-.37**	-.20**	0.07	-.28**	0.13†	.23**	.27**	.37**
Follow-Up	–	–	–	-0.24†	0.04	-0.06	-0.26†	0.07	0.03	-0.25†	0.16	0.12	–[2]	0.29†
Number	177	236	57	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	2.51	2.89	2.95	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.92	0.97	1.11	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Academic Competence Evaluation Scale – Motivation														
Baseline	–	.77**	.48**	-0.05	-.19**	.19*	-.25**	-0.12	-0.07	-0.04	-0.08	0.10	0.02	0.15†
Post-Test	–	–	.55**	-0.09	-0.02	.18**	-.22**	-.18**	-0.01	-0.10	0.01	0.10	.17*	.32**
Follow-Up	–	–	–	-0.13	0.20	0.15	-0.06	0.06	0.12	-0.11	0.06	0.05	–[2]	.38*
Number	177	236	57	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	2.51	2.89	2.95	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.92	0.97	1.11	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Academic Competence Evaluation Scale – Engagement														
Baseline	–	.76**	.44*	0.06	-.36**	0.07	-.19*	-0.14†	-.17*	0.07	-.21**	0.12	0.06	.25**
Post-Test	–	–	.44**	-0.07	-0.01	.15*	-.20**	-.16*	-0.10	-0.03	-0.06	.13*	.15*	.24**
Follow-Up	–	–	–	-0.02	0.10	0.11	0.12	-0.11	0.02	0.00	-0.01	-0.04	–[2]	0.20
Number	182	234	62	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	3.62	3.86	3.60	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	1.11	1.06	1.16	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
Social Skills Improvement System														
Baseline	–	.79**	0.32†	-0.05	-.16*	.23**	-0.10	-0.03	-0.02	-0.08	-0.03	0.06	0.07	.22*

	Base-line	Post-Test	Follow-Up	Inter	Grade	Stu Gen	FRPL	Spec Ed	LEP	Black	Hisp	White	Teach Gen	Yrs Exp
Post-Test	–	–	.50**	-0.08	0.00	.28**	-0.10	-0.09	0.07	-.19**	0.11†	0.07	.15*	.33**
Follow-Up	–	–	–	-0.14	0.09	0.13	-0.05	0.11	0.16	-0.25†	0.21	0.01	–[2]	.40*
Number	184	236	62	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	3.21	3.32	3.15	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.57	0.62	0.61	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38
TRSSEC														
Baseline	–	.55**	.34	.02	-.22**	.22**	.03	-.12	-.02	-.03	.00	-.07	.03	.00
Post-Test	–	–	.54**	-.03	-.04	.21**	-.09	-.11†	.09	-.20**	.15*	.03	.05	.20**
Follow-Up	–	–	–	-.05	.00	.15	.19	.20	.10	-.29*	.22†	.11	–[2]	.34*
Number	181	235	61	3866	3866	3253	3444	3441	3444	3443	3443	3443	3223	2658
Mean	3.19	3.23	3.16	0.52	2.43	0.50	0.81	0.12	0.15	0.53	0.39	0.06	1.07	15.78
SD	0.82	0.90	0.89	0.50	1.13	0.50	0.39	0.33	0.36	0.50	0.49	0.24	0.29	9.38

Note. *$p < .05$ **$p < .01$ †$p < .10$; Inter = Intervention status (0 = PATHS Only, 1 = CARE+PATHS); Grade = Grade; Stu Gen = Student gender (0 = male, 1 = female); FRPL = Free/reduced-price lunch status (0 = no, 1 = yes); Spec Ed = Special education status (0 = no, 1 = yes); LEP = Limited English proficiency status (0 = no, 1 = yes); Black = Black or African American race/ethnicity dummy code (0 = no, 1 = yes); Hisp = Hispanic or Latino/a/e/x race/ethnicity dummy code (0 = no, 1 = yes); White = White race/ethnicity dummy code (0 = no, 1 = yes); Teach Gen = Teacher gender (0 = male, 1 = female); Yrs Exp = Teacher years of experience.

[1]Only four students had a reported disciplinary incident in the 2020-21 school year. Therefore, we used teacher-reported Teacher Observation of Classroom Adaptation–Revised scores as the baseline measure for disciplinary incident outcomes.

[2]Participation in the follow-up reports on student well-being was very low, resulting in no variation in teacher gender at that timepoint. Because of low sample size, we did not test follow-up models for teacher reports of student well-being.

[3]Chicago Public Schools stopped using the TRC in SY2022-23 and did not report a new formative assessment in the administrative records for that year, so we do not have TRC follow-up data.

[4]Our sample included grade 1–4 students in SY2021–22, and the IAR is only administered to grades 3 and above, meaning our baseline sample only included grade 3 students.

Exhibit A9. Student Continuous Outcome Model Fit Statistics

Measure	χ^2	df	*p*-value	RMSEA	CFI	TLI	SRMRw	SRMRb
Attendance rate								
Post-Test	44.80	32	.066	.010	0.997	0.993	.019	.059
Follow-Up	44.17	32	.075	.010	0.997	0.992	.019	.063
Number of Disciplinary Incidents								
Post-Test	46.80	34	.071	.010	0.997	0.994	.025	.060
Follow-Up	43.98	34	.118	.009	0.997	0.994	.031	.056
TOCA-R – Overall								
Post-Test	21.48	26	.716	.000	1.000	1.000	.047	.109
TOCA-R – Aggression								
Post-Test	22.81	29	.785	.000	1.000	1.000	.046	.103
TOCA-R – Inattention/Hyperactivity								
Post-Test	24.08	26	.572	.000	1.000	1.000	.056	.103
IAR – ELA								
Post-Test	9.15	38	1.000	.000	1.000	1.000	.057	.097
Follow-Up	15.56	26	.946	.000	1.000	1.000	.023	.058
ACES – Reading Skills								
Post-Test	23.33	28	.716	.000	1.000	1.000	.046	.105
IAR – Mathematics								
Post-Test	37.56	37	.443	.003	1.000	0.999	.082	.081
Follow-Up	15.51	27	.962	.000	1.000	1.000	.011	.059
ACES – Mathematics Skills								
Post-Test	21.34	28	.811	.000	1.000	1.000	.043	.107
ACES – Motivation								
Post-Test	17.55	26	.892	.000	1.000	1.000	.034	.110
ACES – Engagement								
Post-Test	35.79	27	.120	.033	0.968	0.942	.060	.183
SSIS								
Post-Test	24.93	29	.682	.000	1.000	1.000	.047	.128
TRSSEC								
Post-Test	25.40	29	.657	.000	1.000	1.000	.049	.112

Note. df = degrees of freedom

Exhibit A10. Student Categorical Outcome Model Fit Statistics

Measure	AIC	BIC	aBIC
Chronic Absenteeism			
Post-Test	33329.89	33755.57	33539.50
Follow-Up	29108.60	29553.05	29372.45
One or More Disciplinary Incidents			
Post-Test	25773.95	26293.53	26029.79
Follow-Up	22922.46	23442.04	23178.30
TRC			
Post-Test	16139.25	16423.03	16264.17

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**

Notice of Trademark: "American Institutes for Research" and "AIR" are registered trademarks. All other brand, product, or company names are trademarks or registered trademarks of their respective owners.

Copyright © 2023 American Institutes for Research®. All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, website display, or other electronic or mechanical methods, without the prior written permission of the American Institutes for Research. For permission requests, please use the Contact Us form on AIR.ORG.

XXXXX_MO/23