AERA Open
January-December 2024, Vol. 10, No. 1, pp. 1–17
DOI: 10.1177/23328584241262746
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2024. https://journals.sagepub.com/home/ero

Does Universal SEL Promote Academic Success? Examining Learner Outcomes Under Routine Conditions in First-Grade Classrooms

Susan Crandall Hart
James C. DiPerna
Pui-Wa Lei
Hui Zhao
Tianying Sun
The Pennsylvania State University

Xinyue Li
Cambium Assessment

Kyle Husmann
The Pennsylvania State University

Although proponents suggest that universal social-emotional learning (SEL) programs promote academic achievement, few studies have directly tested these outcomes under routine conditions in schools. Forty first-grade classrooms participated in an effectiveness trial in which schools (N = 13) were randomly assigned to implementation of a universal SEL program (SSIS SEL CIP) or control (business-as-usual practices) conditions within sites/regions. Teachers in the treatment condition prepared for and implemented the program in accordance with their typical local practices. There were no statistically significant main effects on teacher-rated student engagement, motivation, and academic skills. Effect sizes, however, were medium to large and positive for academic engaged time and math achievement. Students with lower teacher-rated academic motivation at pretest also were more likely to improve after program exposure. These findings help identify considerations for further studying the potential academic outcomes resulting from implementing universal SEL under typical conditions.

Keywords: *achievement, program evaluation, student behavior/attitude, classroom research, elementary schools, experimental research, hierarchical linear modeling, experimental design, multi-site studies, social-emotional learning, universal program, academic outcomes, effectiveness trial, routine conditions, SSIS SEL CIP*

Whether, to what extent, and under what conditions school-based interventions improve student achievement has been a primary emphasis of research supported by the U.S. Department of Education for the past 20 years (National Academies of Sciences, Engineering, and Medicine [NASEM], 2022). The Institute for Education Science (IES), for example, has prioritized funding for large-scale causal impact evaluations in education and required the inclusion of student academic outcome measures, even when studies focus on social-behavioral interventions (U.S. Department of Education, IES, 2020). However, there have been surprisingly few randomized controlled trials (RCTs) examining the effectiveness of universal (i.e., delivered to all students in a classroom, grade level, or school as a primary prevention strategy) social-emotional learning (SEL) programs when they are taught within routine conditions in authentic educational settings (typical implementation practices in schools without additional support provided by a research team; Chhin et al., 2018). Educators tasked with selecting programs to adopt and implement in their schools might rightfully ask, Can universal SEL truly fulfill its promise to improve both student behavior and academic performance when delivered in my school, outside of a research study? To make local decisions for their setting, plan for successful implementation in their school context, and have realistic expectations for intervention outcomes, educators need comprehensive, clear, and nuanced information about the current evidence base for universal SEL programs —including their impact on academic performance under real-world conditions like their own.

Linking Student Social-Emotional Skills to Academic Achievement

The association between social-emotional competence and academic achievement is rooted in both theory and research illustrating how relational and contextual factors influence children's development (see Osher et al., 2020). Alongside supportive interactions and environments, an interrelated set of cognitive, social, and emotional skills—such as paying attention, making responsible decisions, getting along with others, and showing empathy—allow students to engage with and benefit from academic learning (S. M. Jones et al., 2019). Students with higher social-emotional skills also have higher academic skills; when students can build relationships, regulate their behavior, and understand emotions, they also do better in school (MacCann et al., 2019). Social-emotional competence predicts school connectedness and can buffer the negative impacts of mental health difficulties that would otherwise pose a risk to educational achievement (Panayiotou et al., 2019). Social-emotional skills help students successfully navigate the social and learning interactions at school, increasing their likelihood of graduating from high school and college, getting a well-paying job, and being physically and mentally healthy (D. Jones et al., 2017).

The knowledge that social-emotional skills relate to positive outcomes for students has spurred the development of universal SEL interventions, and U.S. elementary schools are increasingly adopting them (Bryant et al., 2021). In contrast to small-group or individual social-emotional or behavioral supports provided to students identified as in need of targeted (Tier 2) or intensive (Tier 3) support, universal (or Tier 1) SEL is delivered to *all* students in a classroom or school as part of a prevention approach within a school-based multitiered system of support. Such programs aim to teach students a set of foundational social-emotional skills that are theorized to facilitate positive student outcomes. One theory of change for school-based universal SEL, developed by the Collaborative for Academic, Social, and Emotional Learning (CASEL; 2020), suggests these programs initially provide the short-term benefits of improved student social-emotional skills and attitudes; over time, these changes foster the long-term outcomes of positive behavior, academic success, and mental health. Serving as a foundation for this model, a landmark meta-analysis by Durlak et al. in 2011 found that, on average, school-based universal SEL programs had large, positive effects on measures of student *SEL skills* (i.e., ratings of student cognitive, affective, and social skills demonstrated in test situations or structured tasks) collected within 6 months of implementation (g = .57, 95% confidence interval [CI] [.48, .67], n = 68 studies). Further, SEL programs were associated with increases in students' demonstration of positive behavior in daily situations (rather than hypothetical scenarios); although this effect size was smaller in magnitude (g = .24, n = 86 studies). Similarly, the effect size for the impact on student academic achievement was g = .27 (n = 35 studies), translating into a notable 11-percentile academic gain for an average student exposed to a universal SEL program (Durlak et al., 2011). Given the challenges that schools face in the wake of the COVID-19 pandemic, it is certainly appealing that programs targeting students' social-emotional competence may also contribute to eventual improvements in reading and math skills.

Understanding and Using the Mixed Evidence on Universal SEL-Academic Impact

Using evidence from SEL impact studies to not only anticipate the degree to which delivering social-behavioral program during the school day may yield academic gains for students but also subsequently make decisions about programming is an important yet daunting task for educators. Time is considered one of the most valuable commodities in schools (Zhang et al., 2022); reallocating instructional time to universal SEL could inadvertently cause unintended consequences such as reducing academic learning time or decreasing high-stakes test scores. A recent report on SEL practices in schools attempted to assuage these concerns citing the evidence from Durlak's 2011 meta-analysis, stating that "hundreds of research studies . . . suggest that SEL programs can contribute to long-term academic growth, and may be just as effective as programs designed specifically to support academic learning" (Bryant et al., 2021, p. 4).

There is emerging evidence from rigorous randomized controlled trials, however, that may temper the expectation that universal SEL substantially and consistently improves student academic outcomes (e.g., Hart et al., 2020). For example, citing some methodological limitations in the studies included in Durlak et al.'s (2011) review, Corcoran et al. (2018) conducted an updated meta-analysis of universal SEL evaluation research conducted from 1970 to 2016. They found a limited number of large-scale studies that employed rigorous randomized designs to test the impact of SEL on reading (n = 19 studies) and/or math (n = 18 studies) outcomes, and those studies produced a large range of effect sizes (g) from –.14 to .73 for reading and –.22 to .81 for math. Average main effects for both domains were positive, yet small (gs < .25), and the authors concluded that some SEL interventions "that have dominated classrooms over the past few decades might not have as meaningful [academic] effects for pre-K-12 students as once thought" (Corcoran et al., 2018, p. 56).

Several factors have been identified as potentially contributing to the wide range of academic outcome effect sizes reported in evaluations of universal SEL programs including study design, sample size, grade level, published year, type of academic outcome measure, level of measurement, timing of outcome, and inclusion of baseline academic scores in analyses (Corcoran et al., 2018; Hart et al., 2020). As SEL is a broad term that encompasses a wide range of approaches (e.g., teacher coaching, whole-school development programs, explicit skill instruction; S. M. Jones et al., 2019), the theory of change, goals, scope, target skills, instructional methods, and duration/intensity of a program are also important considerations when interpreting research evidence (Durlak et al., 2022). For example, one small quasi-experimental study from over 20 years ago included in Corcoran et al.'s (2018) review reported that the Boys and Girls Club of America had a very large positive effects on math and reading performance as assessed by student grades (*g*s = .91 and .45, respectively; Schinke et al., 2000), while a more recent large RCT of responsive classroom found negligible-to-small negative effects on state exam scores (reading *g* = –.06, math *g* = –.13; Rimm-Kaufman et al., 2014). Given all these possible sources of variability, using results of these studies to make definitive conclusions about the impact of universal SEL on student academic skills in a real-world classroom context seems extremely challenging, yet educators are faced with this task regularly. Especially given the emphasis on evidence-based interventions in the federal Every Students Succeeds Act, school decision-makers need readily interpretable evidence to make decisions about instructional time allocation, select suitable programs, and obtain funding for universal SEL implementation in their schools (Grant et al., 2017).

Academic Effectiveness Under Routine Conditions: Studying the SSIS Program

Other important factors to consider when interpreting the findings of universal SEL-academic impact studies are the conditions under which research was conducted relative to the everyday conditions in schools. Interventions are often first evaluated in initial *efficacy* trials conducted under ideal conditions, where a research team typically has a great deal of control over or involvement in implementation supports and resources such as training, coaching, and monitoring. If interventions are found to be successful with optimal implementation in such a "best-case scenario," they then often undergo testing under "real world" or routine conditions as part of an *effectiveness* trial (Wigelsworth et al., 2016). Many implementation science frameworks suggest that when programs are not implemented well, they are less effective (Domitrovich et al., 2008); and programs that do not match with the strengths, needs, and realities of a real-world setting are unlikely to be implemented as intended by the developers (S. M. Jones et al., 2019). When causal impact studies are conducted under the realistic conditions found in a broad range of school contexts, their findings are likely to be helpful, relevant, and useful for educators who must use this information in their setting (NASEM, 2022). There is currently a lack of research, however, on the impact of universal SEL on academic performance under routine implementation conditions. Less than 2% of studies funded by IES have been effectiveness, replication, or scale-up trials; of those, the majority have focused on academic, rather than social-behavioral, interventions (U.S. Department of Education, IES, 2022).

The achievement impact of one popular universal program, the SSIS (Social Skills Improvement System) SEL Classwide Intervention Program (SSIS SEL CIP; Elliott & Gresham, 2017), has been evaluated in an efficacy trial (i.e., a more highly controlled study involving researcher support for training and implementation), but its current published evidence base lacks academic outcome data from an effectiveness trial (i.e., evaluation under routine conditions including the everyday practices, staff, and resources that would be normally available in schools; Wigelsworth et al., 2016). The SSIS SEL CIP is a brief, manualized classwide program that focuses on children's mastery of 10 core social skills. Influenced by social learning, operant, and cognitive behavioral theories of learning and development, the program uses direct instruction, modeling, practice, and reflection to teach social behaviors that teachers have identified as essential for classroom learning. The previous randomized controlled efficacy trial in first grade (DiPerna et al., 2018) was conducted with researcher-provided training and support and examined the medial (approaches to learning) and distal outcomes (academic performance) of the program. In a sample of 59 first-grade classrooms, statistically significant ($p < .05$) positive and small effects on teacher ratings of both academic motivation and engagement (*g*s both = .17) were found. Results for direct observations of engagement and assessments of academic achievement, however, were not statistically significant ($p > .05$); effect sizes were all positive, small, and confidence intervals included zero (observed engaged time *g* = .13, reading *g* = .07, math *g* = .04; DiPerna et al., 2018). The impact of this program on learning-related behaviors and achievement under real-world conditions has yet to be evaluated; this information would be useful to educators who need to make decisions within local contexts and under authentic conditions.

Given this, the current study examines four research questions related to the learning-related and academic outcomes of a brief universal program implemented in first grade classrooms under routine implementation conditions in authentic educational settings. Using data from a preregistered IES-funded randomized effectiveness trial in the early elementary grades, we examined the real-world medial impact of the program, i.e., whether students in classrooms implementing the

SSIS SEL CIP program under routine conditions demonstrate different levels of academic engagement and motivation as rated by teachers (Aim 1) and observed engagement in instruction (Aim 2) compared to children in nonimplementing classrooms. We further evaluated the distal impact of the program on reading and math achievement measures (Aim 3). Finally, consistent with our preregistration, we evaluated if effects are different based on student- and class-level variables, such as initial skill levels (Aim 4).

Method

Participants

Thirteen elementary schools within seven school districts participated in this study. Schools were in three states in the West North Central, East North Central, and South Atlantic regions of the United States. A variety of census locales were represented including remote rural, large suburb, midsized city, and large city. In total, 40 first-grade classrooms participated. All teachers were female. Most teachers were White (80%) and spoke English as their primary language (88%); approximately 15% were Hispanic/Latinx, 13% spoke Spanish as their primary language, and 5% were Black. Teachers had an average of approximately 15 years of experience in the classroom, and 10 years of experience in their current school. The majority had a bachelor's degree as their highest level of education (63%), while the remaining teachers had master's degrees. Classroom size ranged from 11 to 26 students ($M = 19$), and 65% of teachers taught in large-sized schools (i.e., over 400 students). Over half of teachers (58%) taught in classrooms comprised of predominately racial-ethnic minoritized students, and 60% taught in schools where at least three-quarters of the students were eligible for free or reduced-price lunch.

The student analysis sample (Table 1) included 344 students, of which 337 were included in the teacher rating and direct assessment outcome analyses. From these participating students, a direct observation subsample of 215 students was randomly selected (approximately 6 students per classroom, stratified by gender).[1] Overall, 18 cases were missing by the end of the study (12 students moved and 6 students had missing data; see detailed description of participant flow in the Recruitment and Data Collection section). Fifty-two percent of students in the analysis sample were male, 48% were female; 44% were White, 24% Black, 23% Hispanic/Latinx, 4% more than one race, and 3% Asian. Students' primary language was approximately 94% English, 5% Spanish, and 2% another language; about 25% of students also spoke a secondary language. Approximately 5% of students received special education services at the beginning of the school year, and 23% received other supplemental academic services (e.g., language support, Title 1, response to intervention).

TABLE 1
Sample Descriptive Characteristics by Treatment Condition in Grade 1

Variable	Control $N = 182$	SSIS SEL CIP $N = 162$
Gender[a]		
Male	57.69	46.30
Female	42.31	53.70
Special education services (≥1)	6.04	4.32
Speech/language impairment	3.30	2.47
Learning disability	3.85	1.85
Emotional behavior disorder	1.10	0.00
Attention deficit/hyperactivity disorder (ADHD)	0.55	0.00
Intellectual/Cognitive Disability	0.55	0.00
Autism	1.10	0.00
Supplemental service	24.73	21.60
English language learners	9.34	3.09
Race-ethnicity		
Asian	6.59	0.00
Black/African American	15.93	33.33
White	55.49	30.86
Hispanic/Latinx	17.58	29.63
Multiracial and other	4.40	6.18
Student skills		
Social skills composite	2.37 (0.47)	2.13 (0.45)
Problem behavior composite[a]	0.30 (0.30)	0.48 (0.40)
Class environment		
Classroom organization	5.67 (0.76)	5.41 (0.92)
Emotional support	5.64 (1.06)	5.54 (0.70)
Instructional support[a]	2.96 (1.14)	1.82 (0.51)
School characteristics		
Large schools[b]	66.48	74.07
Predominantly Black/Hispanic schools[c]	36.81	83.95

Note. Table entries are percentages within condition for categorical variables, mean (*SD*) for continuous variables.
[a]Baseline differences were statistically significant ($p < .05$).
[b]Dichotomized: 1 = large (401–756), 0 = small (129–400).
[c]Dichotomized: 1 = high (60.01%–88.02%), 0 = low (5.43%–60%).

Measures

Student outcome data were collected from both the treatment and control classrooms at the beginning of the year (pretest) and end of year (posttest). Students' approaches to learning were assessed with teacher ratings (academic engagement and motivation) for all students in the sample; direct observation was also used to measure a subsample of students' engagement in instruction. Academic achievement (reading and math) was measured through direct assessment. Measures of classroom environment and students' overall social skills and problem behaviors, as well as student-, class-, and school-level demographic variables, were used as covariates.

Approaches to Learning

Teacher ratings. The Academic Competence Evaluation Scales (ACES; DiPerna & Elliott, 2000) were used to assess students' approaches to learning. Teachers used a 5-point frequency scale ranging from *never* to *almost always* to complete items. The Academic Motivation Scale (11 items) evaluates students' approach, persistence, and interest in learning with items such as "is motivated to learn*"* and "prefers challenging tasks." The Academic Engagement Scale (eight items) includes items such as "speaks in class when called upon" and "asks questions about tests or projects" to evaluate students' attention and active participation in classroom instruction. Evidence from prior studies with early elementary students suggests that scores obtained from these scales are reliable and valid indicators of social and behavioral factors that influence students' classroom learning (DiPerna & Elliott, 2000; DiPerna et al., 2002, 2005). Cronbach's alpha from the current sample was high (.95–.97) across scales and waves.[2]

Direct observation. The Cooperative Learning Observation Code for Kids-2 (CLOCK-2; Volpe & DiPerna, 2018), a direct observation protocol, was used to measure students' instructional engaged time. Informed by several other observation systems with evidence to support their use (e.g., Behavior Observation of Students in Schools; Shapiro, 2004), the CLOCK-2 uses a momentary time sampling method to capture student engagement at 15-second time intervals across observation periods of 10 to 12 minutes. Active engagement assesses when a student is actively attending to an assigned task, such as raising a hand to answer a question, asking a relevant question, writing, or taking out relevant materials for a classroom activity. Passive engagement is when a child passively attends to instruction or an assigned task, such as listening to a teacher talk or looking at the board. In the current study, scores were calculated (proportion of intervals in which a student was observed to be either actively or passively engaged across the 40 to 48 intervals observed) and averaged across the observations conducted (i.e., an average of five observations were conducted per observed student across each data collection period). Interrater reliability (Kappa) for observations conducted by two observers simultaneously was .78 at pretest and .90 at posttest (prevalence and bias adjusted Kappa; Sim & Wright, 2005).

Academic Achievement

Reading skills. The Star Early Literacy and Reading computerized adaptive tests (Renaissance Learning, 2018a, 2018b) were used to assess students' reading skills. Star Early Literacy consists of 27 items designed to measure the early literacy and beginning reading skills of pre-kindergarten to Grade 3 children, including word knowledge and skills, comprehension strategies and constructing meaning, and numbers and operations. Star Reading is a 34-item comprehensive reading test, aligned to national curriculum standards in reading and language arts, for students in Grades K through 12. It assesses five reading domains: word knowledge and skills, comprehension strategies and constructing meaning, analyzing literary text, understanding author's craft, and analyzing argument and evaluating text. Both tests yield scores on the Star Unified Scale, a common scale developed with item response theory that links and spans the entire range of knowledge and skills measured by the two tests. Early Literacy unified scores range from 200 to 1100, while Reading scores range from 600 to 1400. In their norming sample, internal consistency for unified scores were .91 and .98 and test-retest was .85 and .94 for Early Literacy and Reading tests, respectively. In addition, evidence supports their use as a valid measure of student reading skill proficiency (Renaissance Learning, 2018a, 2018b).

Math skills. Star Math (Renaissance Learning, 2018c) was used as a direct assessment of students' math skills. A computer-adaptive standards-based assessment, Star Math includes 34 multiple-choice items that assess K–12 students' skills in four domains: numbers and operations, algebra, geometry and measurement, and data analysis, statistics, and probability. The Unified Score Scale for Star Math ranges from 600 to 1400. Internal consistency for Star Math is reported in the technical manual as .97, and test-retest was .94. Evidence of test validity, including concurrent, predictive, and construct, supports its intended use as a measure of math achievement (Renaissance Learning, 2018c).

Classroom Environment. The Classroom Assessment Scoring System (CLASS K-3; Pianta et al., 2008), a widely used structured observational tool focused on the classroom instructional environment, was used to measure general instructional practices across classrooms in the study. The CLASS K-3 assesses 10 dimensions (positive climate, negative climate, teacher sensitivity, regard for student perspectives, behavior management, productivity, instructional learning formats, concept development, quality of feedback, and language modeling); observers use a 7-point scale ranging from *low* (1–2), *middle* (3–5), to *high* (6–7). These dimensions combine to form three broad domains: emotional support (i.e., teachers' warmth and sensitivity toward students), classroom organization (i.e., teachers' use of effective behavior management and varied learning modalities), and instructional support (i.e., teachers' use of strategies that develop higher-order thinking and language skills). The CLASS has strong theoretical and conceptual underpinnings, and research has generally supported its three-factor structure in early elementary student samples (e.g., Sandilos et al., 2016). Internal consistency was adequate in the current study (Cronbach's alpha ranged .88–.90 across domains). In addition, interrater agreement (within-1-point) was .87 to .92 across domains.

Study Procedures

Recruitment and Data Collection. Given our goal of evaluating the effectiveness of the SSIS SEL CIP in routine conditions, we focused our recruitment efforts on districts that were already considering new adoption of a universal program to promote positive behavior (i.e., typical end-users of a program like the SSIS SEL CIP). Although not an exclusionary criterion for participation, none of the recruited schools reported current use of a published SEL curriculum in their elementary classroom(s). Seven districts were invited to enroll. All first-grade teachers within the participating schools were invited to join the project; all but one teacher provided active consent to participate (Figure 1). Participating teachers then sent home consent forms with their first-grade students; of the 518 families that returned forms, 80% provided active consent. An average of nine students with consent from each classroom were then randomly selected to participate in the data collection associated with the project. Almost all students (99%) provided verbal assent to participate in the study. A total of 362 students were enrolled in the study. The total attrition rate was 5%, with a differential attrition of 2.7% across conditions, which falls within What Works Clearinghouse's (2022) "green" region of tolerable threat of bias under both cautious and optimistic assumptions. All participants were treated in accordance with APA's ethical principles for human subjects' research.

Random assignment was performed at the school-level blocked by sites (i.e., matched schools by region and school demographic variables) such that all first-grade classrooms within a school either implemented the program or continued business-as-usual (BAU) practices. To assess approaches to learning, teachers from both conditions completed online questionnaires to rate their students' academic engagement and motivation at baseline (e.g., pretest—prior to any program implementation) and at the end of the school year (posttest). Teachers also completed additional questionnaires to assess other constructs including student social behaviors, teacher experiences and classroom characteristics, and reports on program implementation. Teachers were compensated for their time spent completing measures.

To directly assess approaches to learning through observation of engagement in instruction, trained field staff members conducted CLOCK-2 an average of five observations at both pretest and posttest on a randomly selected subset of students (approximately five to six students per classroom). Observations were between 10 and 12 minutes long and were conducted during literacy and math academic instruction. Observers attended a 2-day in-person training prior to data collection and reached 80% mastery before conducting observations. During data collection, about 14% of all observations were conducted with two observers to monitor interrater reliability. Field staff members also administered the Star assessments to students to collect academic skill information at both pretest and posttest. Almost all students completed the Star Early Literacy assessment (10 minutes) at the beginning of the year; those that reached criterion (851 unified score) then switched to the Star Reading assessment at the end of the year (20 minutes). All students completed the Star Math assessment at both time points, which took approximately 25 minutes. In addition, a CLASS observation (two cycles of 20 minutes each) was conducted at pretest by a trained and certified CLASS observer. Observers completed an additional mastery activity with a certified CLASS trainer before conducting observations. In addition, a subsample (21%) of classrooms were rated by two observers to monitor interrater reliability.

It should be noted that, although we initially planned to collect data for the trial across multiple school years and cohorts, the current study sample is limited to participants from just one year of data collection (2018–2019 school year). During the 2019–2020 school year, an additional six schools and 21 first-grade classrooms were enrolled in the trial; however, they were not able to complete participation due to the outbreak of the COVID pandemic. As the pandemic continued into the subsequent school year (2020–2021), we were unable to resume data collection with a new cohort of schools and classrooms.

SSIS SEL CIP Implementation. The materials to teach the SSIS SEL CIP were provided to teachers assigned to the treatment condition at the beginning of the year and to the teachers assigned to the control condition after posttest at the end of the year. Materials included a teacher manual, three lesson scripts per unit, digital presentation slides, brief video clips, and role play scenarios. Each core unit covers a foundational social skill aligned with CASEL's SEL framework e.g., ask for help (self-awareness), do nice things for others (social awareness), stay calm with others (self-management), say please and thank you (relationship skills), and do the right thing (responsible decision-making). In addition to the 10 core units, materials were provided for an additional 13 supplemental units that covered more sophisticated social skills. Each core and supplemental unit included three lessons. The program lessons typically require approximately 20 to 25 minutes to teach, and each one includes six instructional steps: tell, show, do, practice, monitor progress, and generalize. Consistent with the goal of evaluating the program outcomes under real-world conditions, research staff did *not* provide any training, coaching, or implementation support, or fidelity feedback to teachers but instead encouraged schools to follow their typical practices for rolling out new curricular programs. Of the teachers who reported their

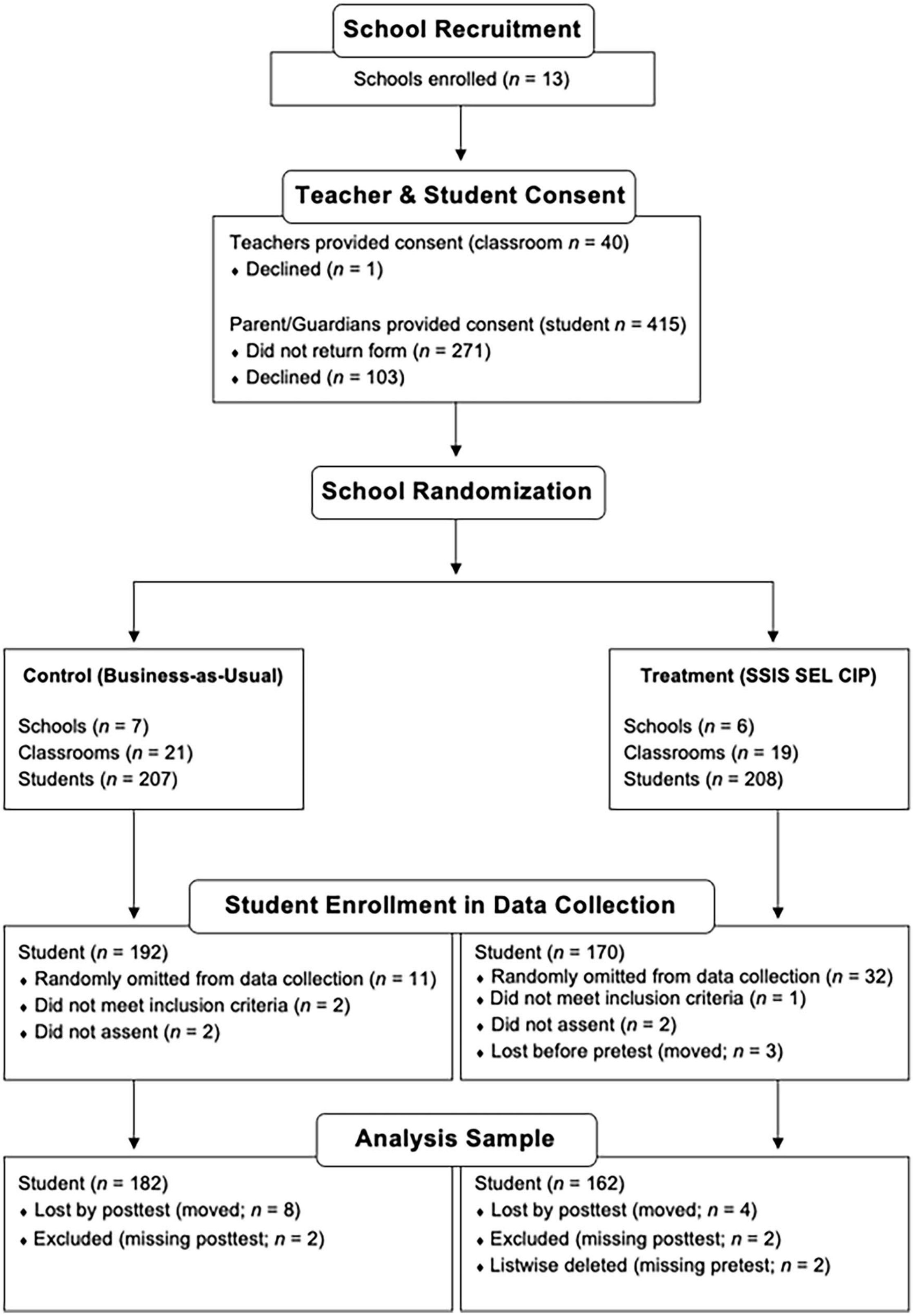
School Recruitment
Schools enrolled (n = 13)
Teacher & Student Consent
Teachers provided consent (classroom n = 40)
♦ Declined (n = 1)
Parent/Guardians provided consent (student n = 415)
♦ Did not return form (n = 271)
♦ Declined (n = 103)
School Randomization
Control (Business-as-Usual)
Schools (n = 7)
Classrooms (n = 21)
Students (n = 207)
Treatment (SSIS SEL CIP)
Schools (n = 6)
Classrooms (n = 19)
Students (n = 208)
Student Enrollment in Data Collection
Student (n = 192)
♦ Randomly omitted from data collection (n = 11)
♦ Did not meet inclusion criteria (n = 2)
♦ Did not assent (n = 2)
Student (n = 170)
♦ Randomly omitted from data collection (n = 32)
♦ Did not meet inclusion criteria (n = 1)
♦ Did not assent (n = 2)
♦ Lost before pretest (moved; n = 3)
Analysis Sample
Student (n = 182)
♦ Lost by posttest (moved; n = 8)
♦ Excluded (missing posttest; n = 2)
Student (n = 162)
♦ Lost by posttest (moved; n = 4)
♦ Excluded (missing posttest; n = 2)
♦ Listwise deleted (missing pretest; n = 2)

FIGURE 1. *Flow of participants in Grade 1 SSIS SEL CIP effectiveness trial.*

planning and preparation activities ($n = 18$), 83% planned on their own, 39% planned with colleagues, 39% watched the publisher-provided online training video, and 17% attended a training conducted by their school or district. Of these four planning/preparation approaches, 39% of teachers used just one, 44% used two, and 17% used three. No teachers used all of them. Of the teachers who planned individually or with colleagues, 88% used existing planning time at school, 29% used additional planning time provided by their school/district, and 18% planned at home.

Of the 30 core unit lessons and 39 advanced unit lessons available to them, implementing teachers reported teaching and average of 24 total lessons (*SD* = 8, range = 7–36). Unit completion appeared variable across the sample: about one-quarter of the implementing classrooms (26%) reported teaching fewer than three complete SSIS SEL CIP units (defined as teaching all three lessons in a unit), while the same percentage of teachers (26%) were on the higher end of implementation dosage, teaching at least 9 complete core social skill units (out of 10 available). To collect information about the implementation of lessons that were delivered, research field staff conducted an average of five observations per teacher. According to these observations, teachers completed approximately 75% of the lesson steps on average (*SD* = 14%). Both observers and teachers themselves used a scale of *not implemented* (1) to *completely implemented* (5) to report on their program adherence; on average, they rated a similar level of fidelity (*M* = 3.83 and 3.86 for observers and teachers, respectively).

Business-as-Usual Practices. Teachers assigned to the control condition in the current sample completed a questionnaire to report their familiarity and use of SEL under routine conditions. Approximately 19% of the teachers reported "very little" familiarity with SEL, 57% reported being "somewhat" familiar, and 24% indicated "extensive" familiarity. Only 14% of teachers, however, reported that they had previously taught a published SEL program. As part of their daily instructional practice, 100% of control teachers reported teaching SEL skills in their classroom. About 90% reported informal instruction (i.e., teachable moments, reinforcement), 67% reported holding morning meetings, and 52% reported formal instructional (i.e., explicit/direct instruction, modeling, practice, feedback) of these skills in their classroom. Furthermore, most control teachers reported teaching/emphasizing the following SEL skills as part of their regular classroom instruction: self-management (90%), responsible decision-making (90%), relationship skills (81%), self-awareness (62%), and social awareness (57%).

Data Analysis

The project was preregistered with the Registry of Efficacy and Effectiveness Studies prior to data analysis (DiPerna & Lei, 2020). Missing data analysis was conducted first. In addition to the above-mentioned attrition, seven students missed one or both direct academic assessments at pretest. The total missing data amount was small (percentage of the enrolled sample with missing data ranged from 1.1% to 4.7% across variables). Missing was not completely at random (Little's MCAR chi-square = 204.2, *df* = 135, $p < .001$). Missing posttests appeared to be related to participant's language status (the missing group did *not* have English language learners). The four participants missing math pretest were racial-ethnic minority students who spoke English as their primary language from schools with high percentages of Black and Hispanic/Latinx students. The four students missing reading pretest also tended to have higher social skill composite scores, did not receive special education, and came from classrooms with higher emotional support and class organization. Missing data were deleted listwise in subsequent analyses due to the small missing amount overall (<5%), and the variables related to missing were included as covariates to mitigate potential bias (e.g., Graham, 2009). The deleted sample also did not differ significantly from the remained sample on any of the baseline or demographic variables.

Baseline equivalence across experimental conditions was assessed on the teacher rating and direct assessment analysis sample (*N* = 337) using a three-level random intercepts regression model to account for the nesting of students in classrooms in schools for student-level variables (gaussian distribution for continuous variables and binominal distribution for dummy demographic variables; multiple categories were collapsed for special education services and racial-ethnic groups). A two-level random intercepts regression model was used for class-level variables (i.e., CLASS classroom organization, emotional support, and instructional support), and logistic regression model was used for school-level demographic variables. Compared to the BAU control (see Table 1), the treatment group had significantly higher baseline scores on Problem Behavior composite (*b* = 0.16, *SE* = 0.05, *p* = .002, ES = 0.46), fewer male students (*b* = –0.44, *SE* = 0.22, *p* = .047, ES = –0.28), and lower CLASS Instructional Support (*b* = –1.03, *SE* = 0.36, *p* = .013, ES = –1.09). These variables were included as covariates in the treatment effect analyses to minimize bias. Baseline difference between conditions in engaged time for the observation subsample was also statistically significant (*b* = 0.10, *SE* = 0.04, *p* = .025, ES = 0.91) and was controlled for in the subsample analyses.

A multisite clustered randomized control trial was conducted to evaluate the effectiveness of the universal SSIS SEL CIP under routine conditions in schools. Schools were randomized into treatment conditions within sites (regions). We used three-level random-intercepts only hierarchical linear models (HLM) to estimate treatment effects for each of the student outcome variables (motivation, engagement,

TABLE 2

Student-Level Descriptive Statistics of Measures by Treatment Condition and Time Point in Grade 1

	Pretest		Posttest	
Variable	Control (*N*=175)	SSIS SEL CIP (*N*=162)	Control (*N*=175)	SSIS SEL CIP (*N*=162)
Approach to learning				
Academic motivation	50.95 (10.38)	48.07 (9.48)	51.02 (11.06)	49.80 (8.58)
Academic engagement	50.65 (9.65)	46.01 (8.89)	52.92 (11.14)	50.14 (8.86)
Engaged time	0.74 (0.12)	0.86 (0.11)	0.81 (0.15)	0.89 (0.11)
Academic skills				
Reading	808.59 (80.43)	801.53 (66.92)	867.57 (95.33)	843.45 (72.66)
Math	830.22 (49.97)	821.67 (44.59)	862.68 (62.01)	857.54 (54.07)

Note. Seven control group students were dropped from primary analyses due to missing pretest or posttest academic skill scores. Academic Motivation and Engagement scores are *T* scores (mean=50, *SD*=10 across time and condition). Engaged Time is proportion of total intervals in which active or passive engaged time was observed (*N*=117 for control and 98 for SSIS SEL CIP).

reading, math, and engaged time[3]). In this three-level structure, students were nested in classrooms in schools while sites were treated as fixed effects. Relevant baseline and demographic variables, including those that showed statistical nonequivalence between treatment conditions, or association with missing data were controlled for in the models. Specifically, we included student gender as reported by teachers (1=male student, 0=female student), race/ethnicity (1=White student, 0=racial-ethnic minority student), language status (1=English language learner student, 0=student who spoke English as a primary language), special education indicator (1=received special education services, 0=did not receive special education services), supplementary service other than special education indicator (1=received supplementary services, 0=did not receive supplementary services), baseline social skills composite, problem behavior composite, and student-level baseline outcome in question (group mean-centered) at the student level.[4] Classroom observation scores on emotional support, organization, and instructional support, as well as class mean baseline outcome (group mean-centered) were entered at the classroom-level. School-level covariates included school enrollment size (1=large [>401 students], 0=small [≤400 students]), percentage of Black and Hispanic/Latinx students (1=high [>60%], 0=low [≤60%]), region/site indicators, and school-level baseline outcome (grand mean-centered). Treatment (1=SSIS SEL CIP, 0=BAU control) effect was tested at the school-level controlling for all covariates listed above (all covariates were grand mean centered unless otherwise specified).

Treatment effect sizes were calculated by dividing the estimated treatment coefficients from the main-effects model by the pooled within treatment group student-level standard deviation of the baseline outcome measures in question (Hedges' *g*). We used baseline student-level standard deviations to standardize the effect estimates because they were not affected by treatment. CIs (95%) for treatment effect sizes were estimated using Hedges' (2011) approximation for three-level models (Equation 32). To further facilitate interpretation of practical importance, the estimated effect sizes were converted to improvement indexes, indicating the average percentile rank change expected had an average control student received the treatment (i.e., assuming standard normal distribution for the control group, the proportion of the area for the difference between the median and the percentile corresponded to the effect size or simply the percentage of the area under the standard normal curve below the effect size minus 50; see What Works Clearinghouse, 2022). In addition, we tested whether treatment effects were moderated by student demographic variables and baseline outcome differences at each of the three levels. Statistically significant interactions were plotted to facilitate interpretation of the patterns.

Results

Average outcome scores were largely similar across treatment groups at both pretest and posttest except observed engaged time at pretest (Table 2). The treatment group on average had slightly (albeit statistically nonsignificant) lower scores on motivation, engagement, reading, and math than the control group at pretest. For the subsample with the direct observation data, however, the treatment group had statistically significant higher engaged time than the control group at pretest. Although both groups gained higher scores at posttest, the between-group differences remained the same in direction but varied somewhat in magnitude.

Intraclass correlations (ICCs) for reading and math were negligible at class-level but not at school-level (>.10 at posttest; Table 3). In contrast, ICCs for teacher ratings of motivation and engagement at posttest were small at school level but moderately large at class level. Engaged time ICCs

TABLE 3

Reliability Indices and Intraclass Correlation for Approach to Learning, Academic Skills, and Engaged Time

	Reliability Indices		ICC (school)		ICC (class)	
Variable	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Approach to learning						
Academic Motivation	.97	.97	.04	.01	.05	.11
Academic Engagement	.95	.96	.11	.06	.05	.19
Engaged Time	.59	.53	.37	.33	.22	.15
Academic skills						
Reading	—	—	.10	.17	<.001	<.001
Math	—	—	.04	.11	.03	.03

Note. κ agreement index reported for Engaged Time and Cronbach's α for all others. ICC = intraclass correlation. Reliability for Reading/Math not reported because item-level data is not available.

were large at class level and even larger at school level at both time points. As such, we report results from three-level random-intercepts models. Among the covariates, pretest scores were strongly predictive of posttest scores of the same outcome at student- and class-levels (Table 4). Holding all else constant, classroom organization, English language learners, and baseline student-level social skills composite were positively associated with academic engagement. Students with higher baseline problem behavior composite scores tended to have lower adjusted academic motivation and math scores. On average, male students had lower adjusted engaged time than female students. Students who received supplemental services also had lower adjusted academic motivation and engagement scores than their peers.

Although students who received SSIS SEL CIP instruction tended to have slightly higher adjusted posttest outcome scores after controlling for relevant covariates than the BAU control group, the differences were not statistically significant. Estimated effect sizes (standardized adjusted differences) were small for medial outcomes (Hedges' g = .09 and .17 for academic motivation and engagement, respectively; Table 5). Improvement index was 3.59% for motivation and 6.75% for engagement, indicating that average students in the control condition would have ranked four to six percentile points higher on these measures had they experienced the SSIS SEL CIP lessons. Engaged time had the largest effect size (g = .47), which was equivalent to about 18% on improvement index. Effect sizes for the distal outcomes varied by domain, negligible for reading (g = .00 with an improvement index of 0%) but larger for math (g = .29 with an improvement index of 11.41%).

We also tested whether the effect of SSIS SEL CIP was moderated by student demographic variables and baseline outcome differences at student, class, and school levels. Only one interaction reached statistical significance at the .05 level (with Benjamini-Hochberg correction for two measures of teacher rating), the moderation of treatment by student-level baseline on academic motivation ($b = –0.24$, $SE = 0.07$, $p = .001$). Students with lower baseline motivation scores tended to have higher adjusted motivation scores after exposure to the SSIS SEL CIP program under routine conditions than their peers in the control condition (Figure 2). Conversely, students with already high baseline scores (i.e., beyond 0.5 *SD* above the mean) tended to have lower adjusted motivation scores than their BAU counterparts.

Discussion

The purpose of this randomized effectiveness trial was to investigate four research aims concerning the medial (learning-related behavior) and distal (academic achievement) outcomes of a brief universal program implemented in 40 first grade classrooms across 13 U.S. schools under authentic, everyday implementation conditions (without researcher support or oversight). Outcomes measures included teacher ratings of academic motivation and engagement, tests of early reading and math skills, and direct observation of students' engaged time in instruction.

With respect to Aim 1, the observed differences in academic motivation and engagement did not reach our a priori threshold of statistical significance. The main effect size for academic engagement was roughly double that of academic motivation, though both confidence intervals included zero. Compared to findings in the previous efficacy trial of the SSIS program in first-grade classrooms that featured researcher-led training and monitoring of implementation (DiPerna et al., 2018), the current effect sizes for these outcomes were similar; however, the between-group differences were statistically significant in the efficacy trial. It is worth noting that the sample size for the efficacy trial was roughly double the size of the current effectiveness study, affording more statistical power. In addition, DiPerna et al. (2018) used class-level random assignment rather than school level, which also increased statistical power.

TABLE 4

Grade 1 Mixed Model Main Effect Model Estimates (SEs) for Approach to Learning, Academic Skills, and Engaged Time

	Teacher Rating		Direct Measure		Observation
Predictor	Academic Motivation	Academic Engagement	Reading	Math	Engaged Time
Fixed effects					
Intercept	51.95*** (6.56)	49.75*** (6.56)	855.82*** (52.39)	844.18*** (21.09)	–0.10 (0.11)
School size	–0.29 (5.36)	–1.51 (6.24)	23.42 (24.29)	12.34 (16.67)	0.08 (0.09)
Black/Hispanic enrollment	–0.91 (5.78)	–1.88 (6.43)	–29.45 (61.24)	–16.41 (23.07)	–0.06 (0.11)
Student-level pretest	0.60*** (0.05)	0.52*** (0.05)	0.67*** (0.05)	0.84*** (0.05)	0.24* (0.09)
Class-level pretest	0.30† (0.17)	0.63* (0.23)	0.80** (0.21)	0.71** (0.21)	0.44 (0.28)
School-level pretest	0.06 (0.97)	0.58 (0.73)	0.78 (0.65)	0.81 (0.48)	–0.35 (0.51)
Student-level motivation pretest	—	0.24*** (0.06)	1.85*** (0.51)	1.22*** (0.31)	—
Classroom organization	2.86 (1.90)	5.58* (2.18)	–3.61 (12.95)	–4.50 (9.34)	0.03 (0.05)
Emotional support	–0.39 (1.37)	–1.86 (1.64)	5.41 (9.17)	7.34 (6.91)	–0.01 (0.03)
Instructional support	–1.50 (1.14)	–2.39† (1.28)	5.00 (7.83)	4.25 (5.41)	–0.02 (0.02)
Social skill composite	1.72 (1.34)	2.95* (1.44)	–2.05 (10.98)	–5.60 (6.82)	–0.02 (0.02)
Problem behavior composite	–3.40* (1.38)	–2.22 (1.47)	–15.97 (13.34)	–20.08* (8.03)	–0.06* (0.03)
Male	–1.16† (.68)	–0.46 (0.72)	–6.72 (6.67)	3.05 (4.07)	–0.04** (0.01)
White	–0.39 (0.90)	0.54 (0.96)	3.82 (8.90)	7.14 (5.34)	0.004 (0.02)
Special education	0.41 (1.63)	–3.27† (1.73)	–2.39 (15.94)	5.66 (9.63)	0.01 (0.03)
Supplemental services	–2.97*** (0.89)	–2.62** (0.93)	–4.35 (8.41)	–9.16† (5.25)	0.02 (0.02)
English learners	2.27 (1.61)	3.82* (1.71)	–4.38 (14.6)	1.07 (9.32)	0.02 (0.03)
Treatment effect					
SSIS SEL CIP	0.91 (2.56)	1.71 (3.23)	1.19 (22.48)	14.75 (11.20)	0.06 (0.07)
	p = .734	*p* = .615	*p* = .960	*p* = .238	*p* = .398
Random effects					
Class-level variance	9.67	11.97	138.86	161.33	0.004 (0.002)
School-level variance	10.15	17.70	390.58	163.13	0.006 (0.005)
Residual variance	33.80	37.97	3359.58	1196.48	0.006 (0.001)

Note. Fixed effect estimates for Region are not presented to conserve space. School Size dichotomized: 1 = large (401–756), 0 = small (129–400). Black/Hispanic Enrollment dichotomized: 1 = high (60.01%–88.02%), 0 = low (5.43%–60%). *N* for teacher ratings and direct measure: Control = 175, SSIS-CIP = 162. For Direct Observation, Poisson model estimates on log scale *(N*: Control = 117, SSIS SEL CIP = 98).

$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

TABLE 5

Standardized Group Differences, 95% Confidence Interval and Improvement Indices

	Effect Size	95% Confidence Interval	Improvement Indices (%)
Approaches to learning			
Academic Motivation	.09	[–0.23, 0.41]	3.59
Academic Engagement	.17	[–0.28, 0.64]	6.75
Engaged Time[a]	.47	[–0.30, 1.24]	18.08
Academic skills			
Reading	.00	[–0.56, 0.56]	0.00
Math	.29	[–0.19, 0.77]	11.41

Note. Effect size is calculated by standardizing coefficient for treatment from the main effects HLM model by student-level pooled within group standard deviation of pretest scores.

[a] Effect size is calculated for the % Engaged Time scale at the means.

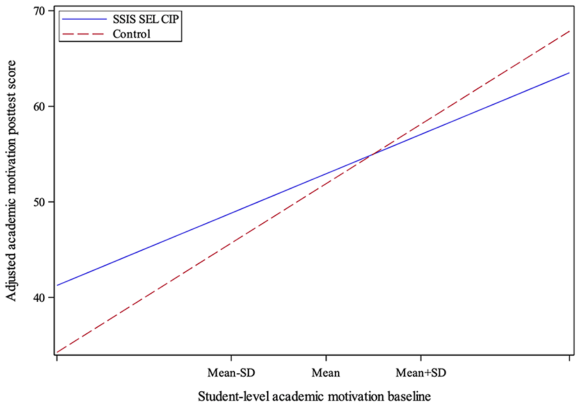

FIGURE 2. *Interaction between student-level baseline and experimental condition on academic motivation.*

Because of disruptions to implementation and data collection caused by the COVID-19 pandemic, the current sample size is roughly one-third of the target specified in the preregistration of our effectiveness trial, and the study is underpowered to detect smaller effect sizes. (Obtained power for academic motivation and engagement was .09 and .14, respectively.[5]) However, the consistent effect sizes under both "ideal" (i.e., the efficacy study by DiPerna et al., 2018) and routine implementation conditions (i.e., the current effectiveness study) suggest that SSIS SEL CIP implementation in first-grade classrooms potentially yields small, positive effects on teacher-rated approaches-to-learning outcomes on average. While the wide confidence bands for the effect sizes in the current study prevent drawing definitive conclusions about program effectiveness, differences in power may explain why effect size estimates are similar across the two studies but only those in DiPerna et al. (2018) efficacy trial were statistically significant.

In Aims 2 and 3, the main effects for engaged time and academic achievement were again not statistically significant in our sample, and all effect size confidence intervals included zero. Although the effect size was negligible for reading, it was medium-to-large for math and observed engaged time (*g*s = .29 and .47, respectively). The magnitude of these latter effects is notable, particularly because these outcomes were assessed using measurement methods (e.g., direct observation from independent observer and standardized achievement tests) that tend to be less susceptible to bias and yield more modest effect sizes than teacher ratings or researcher-developed assessments (Lipsey et al., 2013). In addition, teachers in both conditions reported spending the same amount of time and emphasis on math instruction, which suggests that these uncontrolled variables are likely not explanatory factors for the observed difference in math. In the previous first-grade efficacy trial (DiPerna et al., 2018), the results for these outcomes were similarly not statistically significant; the main effect size was similar for reading but smaller for math and observed engaged time than those in the current study. As noted previously, the reduced sample size due to disruption of the effectiveness trial during the pandemic resulted in limited power (obtained power was .05, .23, and .48 for reading, observed engaged time, and math, respectively) and must be considered when interpreting the current results.

Finally, we used interaction models to understand whether the program had any moderated effects on learning-related or academic achievement of specific groups of first-grade students (Aim 4). We found that the program was more effective for motivation when students started with lower motivation levels. In the previous efficacy trial, there were no interactions between student- or class-level variables and treatment condition that were statistically significant in first grade (DiPerna et al., 2018); however, there were significantly moderated effects for similar outcomes in a sample of second-grade students. Specifically, the program was found to be more efficacious for improving engagement and motivation when second-grade students had lower levels of these skills at the beginning of the year; students receiving support services also fared better in math outcomes (DiPerna et al., 2016).

In previous research on the INSIGHTS program, McCormick et al. (2015) found that intervention effects in reading and math depended on school-level variability: effect sizes were larger for students enrolled in schools with lower levels of leadership and accountability. Effects for academic achievement and student engagement were also larger for students in schools with lower levels of perceived physical and emotional safety (as reported by teachers). Given these results, future research on the SSIS SEL CIP could explore potential moderators beyond student-level pretest scores and demographic variables, including characteristics of the teacher, classroom, and school. Such work could help our understanding of the types of settings and contexts in which this program, and other similar classwide SEL curricula, are most likely to have meaningful effects on learning and achievement outcomes (Durlak et al., 2022). Additionally, differences in implementation practices present another layer of variability that is particularly salient to understanding the effectiveness of the SSIS SEL CIP and similar programs considering the constraints, resources, and staffing typically present in schools. In our study, in which implementation decisions were made locally, teachers assigned to the treatment condition reported a range of approaches to lesson delivery, and program completion also varied across the sample (see Hunter et al., 2022; Neugebauer et al., 2023). While beyond the scope of the preregistered design and analyses for this study, understanding the impact of differing levels of implementation factors (e.g., dosage, fidelity, delivery, engagement) on medial and distal outcomes is an important future direction for research on the effectiveness of universal SEL in routine conditions.

For school-based decision-makers and potential program adopters attempting to anticipate the degree to which the SSIS SEL CIP and similar programs can promote learner outcomes in their local setting, the current results may at first seem inconsistent with Durlak et al.'s (2011) meta-analysis of studies of universal SEL programs that found an average small-to-moderate ($g = .27$) effect on academic achievement. However, Corcoran et al. (2018) reported the largest effect sizes on SEL-academic achievement outcomes were found in studies published during the 1990s and 2000s compared to studies conducted in the past 10 to 15 years that showed smaller or even null effects. For example, a 2020 study by Hennessey and Humphrey found that Promoting Alternate Thinking Strategies (PATHS; Greenberg & Kusché, 2006) had no discernable effect on students' reading, writing, and/or math achievement; all main effect estimates were slightly negative and statistically nonsignificant. The authors suggested these findings may point to important differences in timing, context, design, and intervention between their study and earlier ones (Hennessey & Humphrey, 2020).

Similarly, Jacob et al. (2019) suggested that declining average main effect sizes found in contemporary randomized trials of educational interventions may be in part due limited contrast between the treatment and counterfactual that has occurred in response to changes in the educational policy and practice landscape over time. As such, it could be that BAU practices have evolved in recent years to incorporate facets of SEL instruction and/or emphasize skills that overlap with skills that are covered in universal SEL programs. This certainly seems possible in our sample, in which 100% of teachers assigned to the control condition reported teaching SEL skills in some format (e.g., self-made lessons, classroom meetings, informal interactions) as part of their routine classroom instruction. The existing instructional practices present under routine conditions in schools, which in the past decade may have become more responsive to, and focused on, student SEL needs, may present another source of variability to consider when interpreting research results and making decisions about universal programming.

Limitations

Although this study used a cluster randomized design, multilevel analyses that accounted for student skills at baseline, a sample of geographically and demographically diverse schools, and several methods for measuring student outcomes (direct assessment, observation, and teacher rating), there are several limitations. Schools considering new adoption of a universal SEL program to implement under routine conditions were recruited for the study because they were considered to be typical "end-users"; however, we did not employ a probability sampling plan using theories of moderatorsl nor did we intentionally oversample for subgroups of interest in order to adequately power moderation tests. Furthermore, because only students who had active parent/guardian consent could participate in data collection for the study, we do not have any information for the nonconsenting students and therefore cannot rule out the possibility of baseline differences between students whose families provided consent and those who declined. In addition, consistent with the aims of an effectiveness trial, we intentionally did not provide schools with implementation training, monitoring, or supports. In general, teachers did not appear to receive such supports at their schools as only 3 out of 18 responding teachers indicated that there were efforts at their school to provide training, coaching, monitoring, and/or feedback to them. Although teachers cited multilevel factors that influenced their implementation and program adaptation decisions at the conclusion of data collection (e.g., Hunter et al., 2022; Neugebauer et al., 2023), our current analyses were preregistered to only examine overall main effects and moderation based on student-level characteristics. Therefore, we are not able to rule out the effect that implementation factors or variability may have had on obtained outcomes.

In addition, as implementers of the program, classroom teachers knew their randomly assigned condition, so their ratings may have been susceptible to bias. Relatedly, some field staff members conducted student observations and (separate) observations of teachers' lesson implementation, so these observers were aware of condition assignment as well. Our direct assessment of academic skills and observation of engaged time both represent relatively broad outcomes assessed at the end of the school year after a relatively brief implementation period. Assessments of more fine-grained academic skills and observations more closely aligned with the specific behavioral skills covered in the program lessons may be more indicative of impacts observable by teachers following implementation at the end of the school year.

Toward Understanding and Expecting Heterogeneity

Although the current findings were mixed and differed from registered hypotheses, they raise important considerations for future research and school-based practice. Perhaps most notably, although universal SEL is universally *delivered*, it is not necessarily universally *received*. As such, the value of universal programs for improving student outcomes may not be fully appreciated by examining main effect sizes (as supported by the statistically significant interaction found in this current study). As explained by Greenberg and Abenavoli (2017), prevention programs may offer little improvement for an individual while still garnering population-level benefits. A universal program offered to all students in a classroom or school likely "works" best for certain needs, contexts, and students (e.g., those who start with lower skill levels), however, assuming it causes little adverse effect (i.e., is low-cost, low-risk, and time-efficient), the benefit may still be “worth it.” Schools may be better served—and have more realistic expectations for program effectiveness—if they are given program selection guidance beyond the main effect size found in one sample in an RCT; the goal of future universal SEL research efforts should be to provide schools with actionable information that allows them to make decisions that benefit the most students in their context with the least costs and negative consequences.

Relatedly, Bryan et al. (2021) encouraged researchers to consider the need for a *heterogeneity revolution* and suggested that main effect analyses from large-scale evaluation efforts may be misguided if the intention is to take findings from samples with vast amounts of heterogeneity and generalize these to inferences about the intervention’s effect in other populations. Instead, they and others (e.g., Jacob et al., 2019; NASEM, 2022) have called for a paradigm shift in evaluation research that *expects* intervention effect to be context-dependent and variable, seeks to measure and understand this context and variability, creates infrastructure to support shared research, and incentivizes efforts that partner with practitioners and communities. Along the lines of recent efforts undertaken by Dong et al. (2023), future research on the SSIS SEL CIP specifically, and school-based universal SEL programs more generally, would benefit from integrative data analyses (e.g., the simultaneous analyses of multiple data sets; Curran & Hussong, 2009) that could allow adequately powered investigations that capture contextual and implementation heterogeneity and allow for generalizations to specific populations.

For education stakeholders, the findings of this study may temper expectations about large academic achievement gains resulting from implementation of a classroom program such as the SSIS SEL CIP. However, the consistency in some effect size estimates between this study and the previous first grade efficacy trial suggest that teachers may observe some positive changes in students’ learning-related behavior and/or academic skills following exposure to the program. In addition, the magnitude of effect sizes in this study conducted in routine implementation conditions were similar and, in some cases, larger than in the previous efficacy trial, which may indicate this program’s feasibility for being scaled up into the context of typical school practice (e.g., everyday realities of resource and support constraints).

Nonetheless, the current findings also underscore the challenges faced by decision-makers who are attempting to choose, advocate for, and implement school-based interventions for their students, classrooms, and schools; determining what constitutes an evidence-based universal SEL program for their own school setting and existing BAU practices with the current universal SEL evidence available seems daunting at best and nearly impossible at worst. Future research on the academic achievement outcomes of universal SEL should further prioritize practitioner and community voices about what research questions and outcomes matter most to them, what costs and unintended side effects are incurred during and after routine implementation, and what decisions and adaptations are made to match universal SEL programs to local and authentic contexts.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A170047 to The Pennsylvania State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Open Practices Statement

The larger effectiveness trial from which these data were drawn was pre-registered (#1833.1v1) in the *Registry of Efficacy and Effectiveness Studies*. Dataset and analysis code are available upon request.

ORCID iDs

Susan Crandall Hart https://orcid.org/0000-0003-3762-8232
James C. DiPerna https://orcid.org/0000-0002-4663-9286
Hui Zhao https://orcid.org/0000-0001-9863-5730
Tianying Sun https://orcid.org/0009-0005-9232-9237
Kyle Husmann https://orcid.org/0000-0001-9875-8976

Notes

1. Seven students were only included in the observation outcome analyses due to missing academic pre or posttest data.
2. A shorter five-item version of the engagement and motivation scales with comparable Cronbach’s alphas to the full scales were initially used for pretest while the full scales were used for posttest. To enhance comparability of scores across time, IRT concurrent calibration was used to link the pretest and posttest scores for each scale using the Graded Response Model (Samejima, 1969) based on the common-items nonequivalent groups design (i.e., distribution of participant trait scores for posttest was estimated relative to that for pretest). The estimated latent trait scores were then converted to T-scores (i.e., mean = 50, SD = 10 across pre- and posttests) to facilitate interpretation.
3. The mixed procedure of SAS was used to estimate the HLM models for motivation, engagement, reading, and math outcomes because their score distributions did not depart significantly from normal. The glimmix procedure of SAS was used to estimate Poisson regression models for engaged time because the proportion scores were highly skewed.

4. As suggested by reviewers, we also included student-level baseline academic motivation in the model given a significant moderated effect was found (see Results).

5. Obtained power for this study was calculated using Optimal Design (Spybrook et al., 2005–2011).

References

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bryant, G., Mainelli, A., Crowley, S., Glennen, C., & Edzie, K. (2021). *Finding your place 2021: Social emotional learning takes center stage in K-12*. https://tytonpartners.com/finding-your-place-2021-social-emotional-learning-takes-center-stage-in-k-12/

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a culture of replication: An examination of education and special education research grants funded by the Institute of Education Sciences. *Educational Researcher*, *47*(9), 594–605. https://doi.org/10.3102/0013189X18788047

Collaborative for Academic, Social, and Emotional Learning. (2020). *Evidence-based social and emotional learning programs: CASEL criteria updates and rationale*. https://casel.org/11_casel-program-criteria-rationale/

Corcoran, R. P., Cheung, A. C. K., Kim, E., & Xie, C. (2018). Effective universal school-based social and emotional learning programs for improving academic achievement: A systematic review and meta-analysis of 50 years of research. *Educational Research Review*, *25*, 56–72. https://doi.org/10.1016/j.edurev.2017.12.001

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*, 81–100. https://doi.org/10.1037/a0015914

DiPerna, J. C., & Elliott, S. N. (2000). *Academic competence evaluation scales*. The Psychological Corporation.

DiPerna, J. C., & Lei, P. (2020, January). Effectiveness of the Social Skills Improvement System - SEL Edition Classwide Intervention Program (SSIS-CIP SEL) in the primary grades. *Registry of Efficacy and Effectiveness Studies* (Registry ID: 1833.1v1). https://sreereg.icpsr.umich.edu/sreereg/subEntry/2421/pdf?section=all&action=download

DiPerna, J. C., Lei, P., Bellinger, J., & Cheng, W. (2016). Effects of a universal positive classroom behavior program on student learning. *Psychology in the Schools*, *53*(2), 189–203. https://doi.org/10.1002/pits.21891

DiPerna, J. C., Lei, P., Cheng, W., Hart, S. C., & Bellinger, J. (2018). A cluster randomized trial of the Social Skills Improvement System-Classwide Intervention Program (SSIS-CIP) in first grade. *Journal of Educational Psychology*, *110*(1), 1–16. https://doi.org/10.1037/edu0000191

DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2002). A model of academic enablers and elementary reading/language arts achievement. *School Psychology Review*, *31*(3), 298–312. https://doi.org/10.1080/02796015.2002.12086157

DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2005). A model of academic enablers and mathematics achievement in the elementary grades. *Journal of School Psychology*, *43*(5), 379–392. https://doi.org/10.1016/j.jsp.2005.09.002

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., Romanelli, L. H., Leaf, P. J., Greenberg, M. T., & Ialongo, N. S. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion*, *1*(3), 6–28. https://doi.org/10.1080/1754730X.2008.9715730

Dong, N., Herman, K. C., Reinke, W. M., Wilson, S. J., & Bradshaw, C. P. (2023). Gender, racial, and socioeconomic disparities on social and behavioral skills for K-8 students with and without interventions: An integrative data analysis of eight cluster randomized trials. *Prevention Science*, *24*(8), 1483–1498. https://doi.org/10.1007/s11121-022-01425-w

Durlak, J. A., Mahoney, J. L., & Boyle, A. E. (2022). What we know, and what we need to find out about universal, school-based social and emotional learning programs for children and adolescents: A review of meta-analyses and directions for future research. *Psychological Bulletin*, *148*(11–12), 765–782. https://doi.org/10.1037/bul0000383

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. 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

Elliott, S. N., & Gresham, F. M. (2017). *SSIS SEL Edition Classwide Intervention Program manual*. Pearson.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*(6), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Grant, S., Hamilton, L. S., Wrabel, S. L., Gomez, C. J., Whitaker, A., Leschitz, J. T., Unlu, F., Chavez-Herrerias, E. R., Baker, G., Barrett, M., Harris, M., & Ramos, A. (2017). *Social and emotional learning interventions under the Every Student Succeeds Act: Evidence review*. RAND Corporation. https://doi.org/10.7249/RR2133

Greenberg, M. T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness*, *10*(1), 40–67. https://doi.org/10.1016/j.future.2015.08.005

Greenberg, M. T., & Kusché, C. A. (2006). Building social and emotional competence: The PATHS curriculum. In S. R. Jimerson, & M. J. Furlong (Eds.), *Handbook of school violence and school safety: From research to practice* (pp. 395–412). Lawrence Erlbaum.

Hart, S. C., DiPerna, J. C., Lei, P., & Cheng, W. (2020). Nothing lost, something gained? Impact of a universal social-emotional learning program on future state test performance. *Educational Researcher*, *49*(1), 5–19. https://doi.org/10.3102/0013189X19898721

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380. https://doi.org/10.3102/1076998610376617

Hennessey, A., & Humphrey, N. (2020). Can social and emotional learning improve children's academic progress? Findings from a randomised controlled trial of the Promoting Alternative Thinking Strategies (PATHS) curriculum. *European Journal of Psychology of Education*, *35*(4), 751–774. https://doi.org/10.1007/s10212-019-00452-6

Hunter, L. J., DiPerna, J. C., Hart, S. C., Neugebauer, S., & Lei, P. (2022). Examining teacher approaches to implementation of a classwide SEL program. *School Psychology*, *37*(4), 285–297. https://doi.org/10.1037/spq0000502

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M. A. (2019). A framework for learning from null results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Jones, D., Crowley, D. M., & Greenberg, M. T. (2017). *Improving social emotional skills in childhood enhances long-term well-being and economic outcomes*. Edna Bennet Pierce Prevention Research Center, The Pennsylvania State University. https://www.rwjf.org/en/library/research/2017/07/improving-social-emotional-skills-in-childhood-enhances-long-term-well-being.html

Jones, S. M., Farrington, C. A., Jagers, R., Brackett, M., & Kahn, J. (2019). *National Commission on Social, Emotional, and Academic Development: A research agenda for the next generation*. The Aspen Institute. http://nationathope.org/wp-content/uploads/aspen_research_final_web_optimized.pdf

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2013). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.

MacCann, C., Jiang, Y., Brown, L. E. R., Double, K. S., Bucich, M., & Minbashian, A. (2019). Emotional intelligence predicts academic performance: A meta-analysis. *Psychological Bulletin*, *146*(2), 1580–186. https://doi.org/10.1037/bul0000219

McCormick, M. P., Cappella, E., O'Connor, E. E., & McClowry, S. G. (2015). Context matters for social-emotional learning: Examining variation in program impact by dimensions of school climate. *American Journal of Community Psychology*, *56*(1–2), 101–119. https://doi.org/10.1007/s10464-015-9733-z

National Academies of Sciences, Engineering, and Medicine. (2022). *The future of education research at IES: Advancing an equity-oriented science*. The National Academies Press. https://doi.org/10.17226/26428

Neugebauer, S. R., Sandilos, L. E., DiPerna, J. C., Hunter, L. J., Hart, S. C., & Ellis, E. (2023). 41 teachers, 41 different ways: Exploring teacher implementation of a universal social-emotional learning program under routine conditions. *Elementary School Journal*, *124*(1), 157–192.

Osher, D., Cantor, P., Berg, J., Steyer, L., & Rose, T. (2020). Drivers of human development: How relationships and context shape learning and development. *Applied Developmental Science*, *24*(1), 6–36. https://doi.org/10.1080/10888691.2017.1398650

Panayiotou, M., Humphrey, N., & Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. *Contemporary Educational Psychology*, *56*, 193–204. https://doi.org/10.1016/j.cedpsych.2019.01.009

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

Renaissance Learning. (2018a). *Star Assessments for Early Literacy Technical Manual*. Author.

Renaissance Learning. (2018b). *Star Assessments for Math Technical Manual*. Author.

Renaissance Learning. (2018c). *Star Assessments for Reading Technical Manual*. Author.

Rimm-Kaufman, S. E., Larsen, R. A. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., Merritt, E. G., Abry, T., & DeCoster, J. (2014). Efficacy of the Responsive Classroom approach: Results from a 3-year, longitudinal randomized controlled trial. *American Educational Research Journal*, *51*(3), 567–603. https://doi.org/10.3102/0002831214523821

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph, No. 17). Psychometric Society.

Sandilos, L. E., Shervey, S. W., DiPerna, J. C., Lei, P., & Cheng, W. (2016). Structural validity of CLASS K-3 in primary grades: Testing alternative models. *School Psychology Quarterly*, *32*(2), 226–239. https://doi.org/10.1037/spq0000155

Schinke, S. P., Cole, K. C., & Poulin, S. R. (2000). Enhancing the educational achievement of at-risk youth. *Prevention Science*, *1*(1), 51–60. https://doi.org/10.1023/A:1010076000379

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use interpretation and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Shapiro, E. S. (2004). *Academic skills problems workbook* (3rd ed). The Guilford Press.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Liu, X., Martinez, A., & Raudenbush, S. (2005–2011). *Optimal design plus empirical evidence (Version 3.01)* [software program]. https://wtgrantfoundation.org/optimal-design-with-empirical-information-od

U.S. Department of Education, Institute of Education Sciences. (2020). *Education research grants program: Request for applications*. https://ies.ed.gov/funding/pdf/2021_84305A.pdf

U.S. Department of Education, Institute of Education Sciences. (2022). *Search funded research grants and contracts*. https://ies.ed.gov/funding/grantsearch/

Volpe, R. J., & DiPerna, J. C. (2018). *Cooperative Learning Observation Code for Kids-2*. Unpublished observation code.

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). https://ies.ed.gov/ncee/wwc/Handbooks

Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., ten Bokkel, I., Tate, K., & Emery, C. (2016). The impact of trial stage, developer involvement and international transferability on universal social and emotional learning programme outcomes: A meta-analysis. *Cambridge Journal of Education*, *46*(3), 347–376. https://doi.org/10.1080/0305764X.2016.1195791

Zhang, Y., Cook, C. R., & Lyon, A. R. (2022). A simple matter of time? School-level analysis of the relationship between time allocation, treatment integrity, and student outcome. *School Mental Health*, *14*, 73–87. https://doi.org/10.1007/s12310-021-09412-2

Authors

SUSAN CRANDALL HART is a postdoctoral scholar at The Pennsylvania State University, University Park, 125 CEDAR Building, PA 16802; e-mail: susan.hart@psu.edu. Susan's research interests include implementation, evaluation, and equity of school-based social-emotional health prevention and intervention.

JAMES C. DIPERNA is a professor of education at The Pennsylvania State University, University Park, PA; e-mail: jdiperna@psu.edu. Jim's research interests include social-emotional learning, academic competence, school-based prevention programs, health promotion, and the use of technology to facilitate learning.

PUI-WA LEI is a professor of education at The Pennsylvania State University, University Park, PA; e-mail: puiwa@psu.edu. Pui-Wa's research interests are methodological issues of multivariate statistical analyses, particularly structural equation modeling, and in applications of item response theory.

HUI ZHAO is a doctoral student at The Pennsylvania State University, University Park, PA; e-mail: hzz23@psu.edu. Hui's research interests are text structure strategy, academic self–efficacy, reading comprehension, and application of structural equation modeling.

TIANYING SUN is a doctoral student at The Pennsylvania State University, University Park, PA; e-mail: tianyings@psu.edu. Tianying's research interests are social-emotional learning and multilevel modeling.

XINYUE LI is a psychometrician at Cambium Assessment, Washington DC; e-mail: lixiao0614@gmail.com. Xinyue's research focuses on educational assessment and item response modeling.

KYLE HUSMANN is a postdoctoral scholar at The Pennsylvania State University, University Park, PA; e-mail: kdh38@psu.edu. Kyle's research interests focus on developmental science, human-computer interaction, and implementation science.