This is an Accepted Manuscript of an article published by SAGE Publishing:

Hart, S. C., DiPerna, J. C., Lei, P., & Cheng, W. Nothing Lost, Something Gained? Impact of a Universal Social-Emotional Learning Program on Future State Test Performance, *Educational Researcher*, Published Online January 13, 2020. DOI: https://doi.org/10.3102/0013189X19898721

Nothing Lost, Something Gained? Impact of a Universal Social-Emotional Learning Program on Future State Test Performance

Susan Crandall Hart, James Clyde DiPerna, Puiwa Lei, and Weiyi Cheng

The Pennsylvania State University

Author Note

Susan Crandall Hart, James Clyde DiPerna, Puiwa Lei, & Weiyi Cheng, Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University.

Weiyi Cheng is now at the Professional Testing Corporation, New York, NY.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A090438 and R305B150033 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.

Correspondence concerning this article should be addressed to Susan Crandall Hart, 208G Rackley Building, The Pennsylvania State University, University Park, PA, 16802. E-mail: susan.hart@psu.edu. Phone number: (814) 863-2485. James DiPerna can be reached at jdiperna@psu.edu and 814-863-2813. Puiwa Lei can be reached at puiwa@psu.edu, and Weiyi Cheng can be reached at wuc131@gmail.com.

Abstract

Although the promise of universal social-emotional learning (SEL) programs enhancing student academic outcomes has captured public attention, there has been limited research regarding such programs' impact on students' state test scores. We used multilevel modeling of follow-up data from a multiyear, multisite cluster-randomized efficacy trial to investigate the impact of a brief universal SEL program on students' subsequent state test performance. Although somewhat smaller in magnitude than those reported in previous SEL meta-analyses (e.g., Durlak et al., 2011), observed effect sizes generally were positive and consistent with other studies employing similar designs (i.e., randomized trial, state test outcome, baseline academic covariate). These findings may assuage concerns about the program negatively impacting state test scores due to lost instructional time; however, they also temper expectations about large academic gains resulting from its implementation.

Keywords: social-emotional learning, academic achievement, state tests, methodology, SSIS-CIP

Nothing Lost, Something Gained? Impact of a Universal Social-Emotional Learning Program on Future State Test Performance

A growing body of research has linked children's social-emotional skills and prosocial behavior to a variety of positive life outcomes (e.g., Heckman & Kautz, 2012; Jones, Greenberg, & Crowley, 2015; Moffitt et al., 2016). Studies have demonstrated that these skills and behaviors in early childhood strongly predict future academic achievement (e.g., Caprara, Barbaranelli, Pastorelli, Bandura, & Zimbardo, 2000; Malecki & Elliott, 2002). Perhaps in response, there has been widespread interest in the implementation of programming in schools intended to promote students' social and emotional learning (SEL; Collaborative for Academic, Social, and Emotional Learning, 2019). SEL has been hypothesized to promote student achievement given observed relationships between student social-emotional competencies (i.e., social skills, emotional processes, cognitive regulation; Jones & Bouffard, 2012) and attitudes/behaviors (e.g., approaches to learning, such as engagement and motivation) that facilitate academic learning (DiPerna, Volpe, & Elliot, 2002).

The logic model linking SEL to academic outcomes (i.e., CASEL, 2019) suggests that these programs improve prosocial skills, attitudes, and approaches to learning in the short-term, which subsequently increase the likelihood of students benefiting from classroom instruction over time. Results from multiple research syntheses (e.g., Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Sklad, Diekstra, De Ritter, Ben, & Gravesteijn, 2012; Taylor, Oberle, Durlak, & Weissberg, 2017) indicate that SEL promotes the development of students' social-emotional skills. Less is known, however, about the impact of universal (i.e., taught to all students in a given classroom, grade level, or school) SEL programming on students' academic outcomes in the short- and long-term (Schonfeld et al., 2015).

Impact of SEL on Academic Achievement

Durlak et al.'s (2011) widely cited meta-analysis reported an 11-percentile average gain in achievement (Hedges' g = .27) for participants in SEL programs, but the authors noted a lack of available studies as a primary limitation. Only 16% of the 213 reviewed studies reported academic achievement outcomes, and only 15% included follow-up data in any outcome domain. Furthermore, only some of these studies used randomized designs. In Sklad et al.'s (2012) meta-analytic review of 75 published studies, universal programs were associated with positive effects on students' academic achievement (d = .46 for immediate outcomes; d = .26 for follow-up outcomes) based on the subsample of studies (n = 10 for immediate & 7 for follow-up) that included these outcomes. A meta-analysis of the follow-up effects of SEL programs (collected 6 months to 18 years post-intervention) reported an almost 13-percentile gain in student achievement (g = .33), but less than 10% (n = 8) of the included studies reported academic outcomes of any kind (Taylor et al., 2017). Measurement of student achievement outcomes varied across studies in these meta-analyses with some using standardized tests of achievement and others using GPA, course grades, and/or teacher ratings of academic competence.

A more recent meta-analysis of universal SEL focused exclusively on academic achievement impacts. Corcoran, Cheung, Kim and Xie (2018) identified 611 SEL evaluations conducted between 1970 and 2016 and determined that 40 of them met their inclusionary criteria (e.g., control condition, random assignment or matching technique, baseline equivalency testing, and achievement outcomes). Across these 40 studies, they found positive effects on academic achievement (reading g = .25, math g = .26). The authors also explored whether methodological and substantive features of individual SEL evaluation studies (e.g., research design, sample size, grade level, program intensity) could be systematically related to variations in effect size. They identified 19 high quality randomized controlled trials (RCTs) with larger sample sizes that reported reading outcomes, and effect sizes ranged from -.14 to .73 across these studies. Given this variation, the authors suggested that some popular programs may not have as meaningful academic effects as previously assumed (Corcoran et al., 2018).

Jones and Doolittle (2017) also noted differences between the magnitude of observed effects reported in large meta-analyses and those from individual RCTs of SEL programs which tend to be more variable and modest. As one example, a large-scale multisite RCT of seven SEL programs provided no evidence that the programs improved student outcomes (Social and Character Development Research Consortium [SACD], 2010). Further, in their review of 13 RCTs evaluating 11 popular SEL programs published between 2004 and 2015, Jones, Barnes, Bailey, and Doolittle (2017) concluded that program impacts on student-level outcomes were mixed, with both statistically significant (and mostly small to moderate in magnitude) effects and nonsignificant findings reported. With respect to academic outcomes, positive student-level effects were most often based on teacher report rather than direct measures of student academic skills. As Jones and Doolittle (2017) noted, the evidence of SEL efficacy from "gold-standard" RCTs is ambiguous, often demonstrating mixed or null findings.

Universal SEL and State Test Performance

One area that has received limited attention to date is the impact of universal SEL on students' state test scores. Surprisingly, few studies have specifically examined U.S. students' state test performance after exposure to universal SEL programming, despite these scores being a policy-relevant, easily obtainable, and suitable indicator of students' academic achievement (Somers, Zhu, & Wong, 2011). State test performance not only has the potential to shed light on the impact of SEL on distal student outcomes but also represents an area of emphasis for educators across the country. Although a majority of teachers support teaching social-emotional skills in schools (Bridgeland, Bruce, & Hariharan, 2013), there are important concerns about allocating instructional time to non-academic programming given resource constraints and emphasis on test score improvement.

Many schools have begun actively restricting classroom time that does not directly prepare students for high-stakes tests (Schonfeld et al., 2015), and some researchers (e.g., DiPerna, Lei, Bellinger, & Cheng, 2016; Rimm-Kaufman et al., 2014) have suggested that the possibility of decreased academic instructional time resulting from implementation of universal SEL curricula should be explored. Others (e.g., Whitehurst, 2019) even question whether SEL may be a “distracting fad that comes with high opportunity costs.” (p. 68). Given the finite instructional time present in a school day, this fear may not be unfounded; teaching SEL requires replacing time that would otherwise be spent on a different type of instruction or activity. In this way, a “lost opportunity cost” (i.e., benefit that is missed when a specific program is implemented) of universal SEL may be that students receive less academic instruction (Hunter, DiPerna, Hart, & Crowley, 2018; Levin, McEwan, Belfield, Bowden, & Shand, 2017).

School professionals tasked with making choices about SEL implementation face two seemingly diametrically opposed questions. Does SEL facilitate *gains* in future student test scores by improving student social-emotional skills and behaviors that enhance classroom learning over time? Or, does SEL implementation create *losses* in future test score performance as a result of reallocating academic instructional time to focus on SEL skills? Few published studies are currently available to explore these questions, likely in large part due to this being a concern that has emerged since the mandates of the No Child Left Behind Act (NCLB) in 2001. As an example, of the eight studies in Taylor et al. (2017) reporting follow-up academic outcomes, only two were published since the enactment of NCLB, and both were the only studies to employ district or state test scores.

Results from the handful of SEL RCTs that have reported effects on state tests are mixed. In one study (Snyder et al., 2010) of the Positive Action program (Flay, Allred, & Orway, 2001), medium-to-large effect sizes were reported for school-level Grade 4 state test proficiency; another study of the same program (Bavarian et al., 2013) reported small-to-medium effect sizes for an aggregated school-level measure of state test proficiency for students in Grades 3-8. Schonfeld et al.'s (2015) study of the Promoting Alternative THinking Strategies (PATHS; Kusche & Greenberg, 1994) program yielded mostly small positive effect sizes for students' state test proficiency in Grades 3-6. Recent RCTs of Second Step (Committee for Children, 2008), 4Rs (Morningside Center for Teaching Responsibility, 2001), and Responsive Classroom (Northeast Foundation for Children, 2007) reported mostly small effect sizes - some positive and others negative (Espelage, Rose, & Polanin, 2016; Jones, Brown, & Aber, 2011; Jones, Brown, Hoglund, & Aber, 2010; Rimm-Kaufman et al., 2014). Across all of these state test studies, however, most differences were found to be statistically nonsignificant (ps >.10) or marginally significant (.05 < ps < .10). Collectively, existing evidence suggests that the links between universal SEL and state test gains are not clear.

Several of these studies, though, reported differential effects in relation to student skill level prior to program implementation. Jones et al. (2010, 2011) found that the 4Rs program may be most efficacious for students identified by teachers as having the highest behavioral risk. Similarly, achievement effects of Responsive Classroom were most pronounced for students with lower initial academic skills (Rimm-Kaufman et al., 2014). Examining the impact of SEL on specific subgroups of students—such as baseline skill levels as an indicator of risk—has been noted as an important direction for the field in advancing understanding of which students benefit from universal programs (Greenberg & Abenavoli, 2017).

Variations in SEL-State Test Impact Studies

Adding to the complexity, a number of methodological features vary across studies investigating the impact of SEL on state test scores, which in turn may explain at least some of the variability in the magnitude of observed effects (see Cheung & Slavin, 2016). For example, some RCTs have reported continuous test scores at the student level (Espelage et al., 2016; Jones et al., 2010, 2011; Rimm-Kaufman et al., 2014), while others have analyzed proficiency scores at the student level (Schonfeld et al., 2015) or school level (Bavarian et al. 2013; Snyder et al., 2010). Similarly, some have used student-level academic (Espelage et al., 2016; Rimm-Kaufman et al., 2014) or behavioral (Jones et al., 2010, 2011) skills to examine baseline equivalency between groups and/or control for pretest differences in outcome analyses. Others (Bavarian et al., 2013; Schonfeld et al., 2015; Snyder et al., 2010) did not collect student-level academic data to examine the equivalence of academic skills between conditions at baseline. At present, the impact of these particular variations has not been systematically examined between or within studies.

Intervention-related factors also are important to consider as the broad umbrella of universal SEL has grown to include many program foci and outcomes (Jones & Doolittle, 2017). Conceptual differences in the characterization of SEL has led to the term being used to describe a number of educational interventions targeting individuals, classrooms, and entire schools. Program targets run the gamut from discrete social skills, to cognitive/emotion regulation and prosocial behavior, to broad personality traits and dispositions (Whitehurst, 2019). In the Taylor et al. (2017) review, for example, 63% of studies reporting follow-up academic outcomes evaluated comprehensive whole-school reform efforts (e.g., Yale Child Study Center Prevention Model, Child Development Project) rather than more narrowly defined SEL programs. While several unifying frameworks have been developed to help clarify definitions and theories of change associated with SEL (i.e., CASEL, 2019; Jones & Bouffard, 2012), variability remains regarding what constitutes an SEL program.

Universal SEL programs also can vary in terms of their guiding theory, instructional approach, integration with academic curricula and instruction, target grades/ages, content, format/method, instructional time, intensity, duration, training, and cost. Some programs (e.g., Positive Action, 4Rs) utilize over 35 hours of direct instruction with students, while others like Responsive Classroom embed SEL into teacher instructional practice. PATHS and Second Step focus on cognitive-behavioral approaches to SEL, emphasizing areas like self-regulation, emotion management, and problem solving, while 4Rs uses literacy-based lessons and requires extensive teacher training and ongoing coaching by program developers. Although some have suggested that the dosage of educational programming may have an impact on both the magnitude and direction of expected outcomes (Cheung & Slavin, 2012), Corcoran et al. (2018) reported the reading effect sizes for low intensity vs. high intensity SEL interventions were .32 and .12, respectively. Those authors concluded that increased SEL dosage does not necessarily lead to larger academic impacts. This point may be particularly relevant given aforementioned concerns that SEL implementation may actually displace academic instruction in schools.

Social Skills Improvement System Classwide Intervention Program (SSIS-CIP)

One universal social-emotional program developed for elementary classrooms is the Social Skills Improvement System Classwide Intervention Program (SSIS-CIP; Elliott & Gresham, 2007). The program and is grounded in operant, social learning, and cognitive-behavioral theories of student learning and behavior. The SSIS-CIP is focused on the promotion of positive social behavior and includes 10 instructional units targeting skills such as cooperation, self-control, responsibility, assertion, and empathy. Scripted lesson plans each require approximately 20-30 minutes to complete. Lessons are taught via instructional strategies such as reinforcement, modeling, role-playing, and problem-solving, and they include brief videos and student role-play exercises. The program is relatively brief (10-12 hours of instructional time), does not require extensive preparation or formal training, and has been used in schools across the United States.

A multiyear, multisite cluster-randomized trial evaluated the efficacy of the SSIS-CIP with respect to several outcome domains. Reported effect sizes for social skills were g = .18 (7.14% improvement index) in first grade and g = .36 (14.06 % improvement index) in second grade (DiPerna, Lei, Bellinger, & Cheng, 2015; DiPerna, Lei, Cheng, Hart, & Bellinger, 2018). Effect sizes were similar in magnitude for academic motivation and engagement at both grade levels (DiPerna, Lei, Bellinger, & Cheng, 2016; DiPerna et al., 2018), and moderation analyses indicated that students who were in classrooms with lower levels of initial social and learning-related behavior outcomes made the greatest gains (DiPerna et al., 2015; 2016). In both grade levels, the program had a negligible impact on immediate academic outcomes as measured by reading and math computerized adaptive tests given within the same year as program exposure (ES range -.04 - .07; DiPerna et al., 2015, 2018); however, long-term effects of the SSIS-CIP on state test scores have not been examined to date.

As such, the primary purpose of this study was to examine the association between exposure to the SSIS-CIP, a universal SEL program, and students' state test performance in subsequent grades. Specifically, we investigated the impact of implementation of the SSIS-CIP in second grade on student achievement in Grades 3-5 as measured by state test scores in math and reading. Given that previous RCTs of other SEL programs (e.g., Jones et al., 2010, 2011; Rimm-Kaufman et al., 2014) have found treatment interactions with baseline student skills on state test outcomes, we also tested for interactions between SSIS-CIP and students' academic skills prior to program implementation. Finally, in light of recent work exploring methodological considerations with respect to SEL-academic achievement impact studies (e.g. Corcoran et al., 2018), we systematically explored the use of different types of scores (i.e., continuous vs. proficiency status) and inclusion of baseline academic skills to see what, if any, impact these methodological variations had on observed outcomes.

Method

Participants

Figure 1 displays the flow of participants through study phases (implementation in Grade 2, outcome data collection in Grades 3-5) by condition. Three cohorts of second grade classrooms (*N* = 51) across seven schools in two school districts (one small urban, one small rural) participated in the study. The classroom sample comprised 95% of all second grade classrooms across the participating schools (*N* = 7). Prior to classroom randomization to condition (SSIS-CIP or business-as-usual control), the parents of all second grade students were invited to have their child participate in the study. Although all students in treatment classrooms participated in the SSIS-CIP lessons, only students whose parents provided consent participated in the data collection. The majority of the student participants were female (56%) and white (75%). Approximately 10% received special education services, and 20% received supplemental services (e.g., Title 1 academic support). All participating second grade teachers were white, and most (82%) were female. All participants were treated in accord with APA ethical guidelines.

Measures

State test scores. The Pennsylvania System of School Assessment (PSSA; Pennsylvania Department of Education [PADOE], 2012), a criterion-referenced achievement test based on the Pennsylvania Academic Standards in math and reading. Students in Grades 3-8 take the PSSAs each spring, and performance is reported as both scaled scores and performance levels (i.e., Below Basic, Basic, Proficient, and Advanced). Educators and administrators often focus on performance levels when communicating with key stakeholders (parents, students, and colleagues) rather than relying on continuous scores as these latter scores have little meaning without providing additional contextual information (e.g., score distribution for the child's grade level). Given the use of performance levels in school practice, and the use of continuous scores in many recent SEL studies, we examined continuous scores and proficiency status (i.e., meeting Proficient or Advanced cut-off score as determined by the PADOE for that year, grade level, and subject area) in the current study.

The PSSA math test includes 72 multiple-choice questions and 4 open-ended items, and it assesses content in five domains: (a) numbers and operations, (b) algebraic concepts, (c) geometry, (d) measurement, and (e) data analysis and probability. The reading test includes 58 multiple choice items and 3-5 open-ended questions assessing two content domains: (a) comprehension and reading skills, and (b) interpretation and analysis of fictional and nonfictional text. Overall reliability estimates (Cronbach's alpha) were .93-.95 in math and .91-.92 in reading across Grades 3-5. (See PADOE [2012, 2013, 2014] for further details regarding content, reliability and validity evidence, and other technical characteristics of the PSSAs.)

Pretest academic skills. The STAR Reading and Math computerized adaptive tests (Renaissance Learning, 2009, 2010) were used to assess students' baseline academic skills in the fall of second grade. STAR Reading uses vocabulary-in-context test items to measure students' skills in constructing meaning from text, while STAR Math assesses numeration and computation objectives through multiple-choice items. Overall reliability of the STAR scores are reported to be high (.95), and validity evidence supports their use for their intended purpose as a measure of student academic skill proficiency (Renaissance Learning, 2009, 2010). The STAR assessments were administered by trained research assistants during the 4-week period before exposure to the SSIS-CIP in Grade 2.

Intervention Implementation

Second grade teachers from classrooms randomly assigned to the implementation condition participated in a 1-day training regarding the SSIS-CIP Early Elementary Version (Elliott & Gresham, 2007). These teachers then implemented the SSIS-CIP during a 12-week period from early November through mid-February. To monitor implementation fidelity, teachers completed weekly standardized checklists and research staff completed independent observations of approximately 20% of the SSIS-CIP lessons in each classroom. Fidelity was high based on ratings by teachers (98%) and independent observers (97%).

Data Analyses

Overall and differential rates of attrition were calculated according to the guidelines in the most recent *What Works Clearinghouse Procedures Handbook - Version 4.0* (WWC; U.S. DOE, 2017b). By design, some cohorts of students did not reach the later grade levels by the end of the project; their outcome data are therefore "absent by design" and considered "ignorable" (i.e., not counted as attrition and not compromising; U.S. DOE, 2017b). All other sources of missing outcome data, however, were included in attrition calculations for each sample (math and reading outcomes in Grades 3-5). With one exception, these combinations of overall (2% - 22%) and differential (1% - 5%) attrition resulted in tolerable levels of potential bias for both optimistic and cautious sets of assumptions, as defined by the WWC (U.S. DOE, 2017b), meeting the criterion for *low attrition*. The exception was Grade 4 math, in which the combination of overall (19%) and differential (8%) attrition could result in a “tolerable” threat of bias under optimistic assumptions, but an “unacceptable” threat of bias under cautious assumptions according to WWC. Across all grade levels and subject areas, all chi-square tests for differential attrition fell above the .05 statistical significance threshold. After attrition calculations, missing cases of outcome data were deleted, and the resulting dataset (with imputed missing baseline data) comprised the analytic sample.

Missing baseline data ranged from 1% to 12% of the analytic samples. Baseline data were missing completely at random based on Little’s MCAR test in all samples ($ps > .05$) except for the Grade 4 math sample ($p = .046$). There were no significant associations, however, between missingness and any of the demographic or score variables in the Grade 4 math sample. As listwise deletion would have resulted in the loss of 10% or more of the analytic sample in some grades/subjects, missing baseline data were imputed per the WWC’s recommended approaches for addressing missing baseline data (U.S. DOE, 2017b).

Specifically, multiple imputation was conducted using the Blimp Multilevel Imputation Software 1.0 application (Keller & Enders, 2017). Blimp performs imputation by implementing a fully conditional specification algorithm (i.e., chained equations and sequential regression) with a latent variable formulation for incomplete categorical variables (Enders, Keller, & Levy, 2017). Blimp 1.0 for Windows was used to perform 5 sets of multilevel imputations for cases of missing baseline variables in Grade 3 (42 for math sample, 34 for reading), Grade 4 (37 for math, 30 for reading), and Grade 5 (7 for math, 1 for reading). All variables included in the analyses (indicator variable for intervention status, all of the covariates used in the impact model, and the outcome data) were used in the imputation procedure. The nested data structure was accounted for using cluster identifiers. WWC guidelines for analyses with missing data in low-attrition RCTs do not require assessing baseline equivalence (U.S. DOE, 2017b, p. 36-46). However, we calculated differences in baseline characteristics (in standard deviation units) between the intervention and control group prior to exposure to the SSIS-CIP. The absolute value of baseline difference effect sizes were less than .25 for the majority of variables (math pretest, gender, race, and special education) in both the math and reading analytic samples. They were larger than .25, however, for the reading pretest and the supplemental services variable in both the math and reading samples. In our primary analyses, we controlled for all available baseline characteristics, cohorts, and schools in all impact models to minimize possible bias.

Multilevel modeling was used to estimate the effect of SSIS-CIP implementation in Grade 2 on later elementary (Grades 3-5) state test performance to account for students being nested within classrooms and schools. We used unconditional three-level models to estimate intraclass correlation (ICC) coefficients, which indicate the degree to which the assumption of independence was violated due to clustering (Raudenbush, 1997). For state test scores, ICCs ranged from .06 - .54 at the class level and from .00 - .19 at the school level. Given the size of these ICCs (Raudenbush, Spybrook, Liu, & Congdon, 2005), two-level models (students nested in classrooms), with school modeled as a fixed effect due to the relatively small number of schools, were analyzed to provide proper standard error estimates. We adjusted for classroom and school variations in all analyses.

Student math and reading state test scores in Grades 3-5 were the dependent variables, analyzed (separately) as continuous scaled scores and proficiency status (1=proficient, 0=nonproficient). Classroom assignment to condition (1=SSIS-CIP intervention, 0=business-as-usual control) in Grade 2 was the independent variable of interest for this study. Student-level covariates in the multilevel model included pretest academic skills, students' sex (1 =male, 0 = female), race (1 = white, 0 = non-white), receipt of supplementary services (1 = yes, 0 = no) and receipt of special education (1 = yes, 0 = no) in Grade 2. All those variables were grand-mean centered within grade level/subject area. Dummy coded cohort and school variables were also included as covariates in the model. To explore if SSIS-CIP treatment effects depended on prior academic skill levels, interaction effects between SSIS-CIP treatment and student-level academic pretest were tested by adding product terms to the model. When product terms were statistically significant at the .05 level, the pattern of interaction was further examined by plotting the adjusted means. We estimated multilevel models using the Mixed procedure of SAS (version 9.3) for continuous scaled scores and the Glimmix procedure with Bernoulli distribution and logit link for proficiency status.

Given concerns about the limitations of relying only on statistical significance testing for interpretation of study findings and the growing consensus about the need for reporting multiple indices of results for valid scientific reasoning (e.g., Wasserstein & Lazar, 2016), we also report effect sizes for each main effect outcome. Following the WWC guidelines (U.S. Department of Education, 2017a), Hedges' *g* (i.e., the adjusted group mean difference divided by the pooled within-group student-level standard deviation) was calculated when analyzing continuous state test scores. An improvement index (U.S. DOE, 2017a; the expected percentile rank improvement for an average student in the control group had the student received the treatment) also was calculated as a more practical indicator of SSIS-CIP impact. For proficiency status, effects were reported as odds ratios (i.e., the estimated odds of reaching state test proficiency for students exposed to the SSIS-CIP compared with the odds for those not exposed). Additionally, 95% confidence intervals were calculated and reported to provide insight into the uncertainty associated with the effect sizes. Finally, effect sizes were calculated for models in which the student-level academic pretest was not included to elucidate the impact of controlling for baseline student academic achievement on the magnitude and pattern of observed effects.

Results

Primary Analyses

Math. Table 1 reports descriptive statistics for demographic variables, STAR pretest math scores, and state test math scores for the students who completed the math state tests in each intermediate grade. SSIS-CIP intervention effect was estimated using 2-level models controlling for baseline student characteristics, cohort, and school variables. As shown in Table 2, for both continuous scores and proficiency status, SSIS-CIP exposure did not yield any statistically significant differences on state test performance (all *p*s >.05). As shown in Table 3, controlling for the math pretest, effect sizes were positive, small, and similar in magnitude for continuous scores (*g*s = .13 - .15) across the three grade levels. In terms of an improvement index, an average comparison group student would have demonstrated an approximate 5% increase in percentile rank of math continuous scores in Grades 3-5 if the student had received the SSIS-CIP. Odds ratios (Table 3) were small in magnitude and similar across Grades 3 and 4 (ORs = 1.15 and 1.17), but somewhat larger in magnitude in Grade 5 (OR = 2.05).

Two statistically significant interactions were observed between SSIS-CIP exposure and baseline math skills (Figure 2). Students in the SSIS-CIP condition who had lower pretest math scores demonstrated a greater probability of reaching proficiency status on Grade 3 (p = .04) and Grade 5 (p = .048) math state tests relative to their peers in the control condition. For continuous math scores, however, the only statistically significant (p = .002) interaction was in Grade 3 where SSIS-CIP participation was associated with higher math state test scores for students with lower initial math skills (Appendix Figure A1).

Reading. Table 4 reports descriptive statistics for demographic variables, STAR pretest reading scores, and state test reading scores for students in Grades 3-5. Controlling for all other variables, SSIS-CIP participation was not associated with any statistically significant differences in students' reading state test continuous scores (Table 5). Differences in probabilities of reaching proficiency on the reading state test between students in SSIS-CIP and control classrooms were only statistically significant in Grade 5 (p = .03). In that grade, controlling for student reading skills at pretest, SSIS-CIP exposure was associated with higher odds of reaching proficiency in reading (OR = 3.60; Table 3), though the relationship between reading proficiency status and SSIS-CIP participation was positive in all three grades. In contrast, the observed effect size for continuous reading scores was negative and small in Grade 3 (g = -.09) but positive and small in Grades 4 and 5 (gs = .05 and .10, respectively). Improvement indices ranged from -3.59% to 3.98% across grade levels. There were no statistically significant interactions between SSIS-CIP and baseline reading skills on state test reading performance across the two scores types.

Removal of Academic Pretest

Given the aforementioned inconsistent findings —as well as methodological variations with respect to accounting for students' baseline academic skills—across previous studies of universal SEL academic outcomes, we conducted follow-up analyses to examine the potential impact of omitting the academic pretest on observed effects. Specifically, we re-ran the previous models without the baseline academic covariate. Resulting model estimates appear in Appendix

Table A1 (Math) and Table A2 (Reading). It is important to note that these “no-pretest” models are not reported as evidence of the SSIS-CIP’s impact on academic outcomes. On the contrary, these models are reported to highlight the potential impact of not controlling for a baseline measure of student academic achievement in studies of long-term academic outcomes resulting from SEL implementation.

As shown in Table 3, there were no statistically significant differences for state math test continuous scores or proficiency levels (all *p*s > .05). Effect sizes in models omitting the math pretest were positive and small in magnitude (*g*s = .13 – .17, ORs = 1.04 – 2.03), and improvement indices ranged from approximately 5 – 7%. As shown in Table 3, without controlling for students’ reading skills at baseline, the differences in reading state test performance were statistically significant for Grade 5 continuous scores (*p* = .01) and proficiency status (*p* = .006). Effect sizes in the no-pretest models in reading were all positive (*g*s = .10 - .17; ORs = 1.26 - 4.87) with improvement indices ranging from approximately 4% – 7%.

Discussion

The primary purpose of this study was to investigate the effects of implementing the SSIS-CIP, a universal social-emotional learning curriculum, in second grade on students’ subsequent state test performance in Grades 3-5. Almost all observed effects were positive; however, the majority of 95% confidence intervals extended into the negative range. In addition, the majority of differences were not statistically significant (i.e., *p*s > .05). While overall observed differences were consistent with, and in several cases more positive than, findings from other SEL studies with comparable designs, findings from the current study suggest some possible variations in results by score type (continuous scores vs. proficiency), skill area, grade level, and student baseline skills.

In math, observed effect sizes from continuous score main effects models were similar across samples in Grades 3 – 5, though the magnitude of the effect appeared somewhat larger for Grade 5 proficiency status. In contrast to the fairly consistent results across grade levels in math, observed main effect sizes from both continuous score and proficiency status models in reading demonstrated an increasing pattern across grades. Overall, the pattern of observed effects from continuous score models yielded mean effect sizes that were somewhat larger in magnitude for math relative to reading (95% CIs overlapped in Grades 4 and 5, however). In proficiency models, though, mean odds ratios appeared similar in math and reading in Grade 3 and 4, but smaller in math as compared to reading in Grade 5. Tests of interactions by pretest academic skill indicated statistically significant interactions for some score types in Grade 3 and 5 math, but none in reading. Controlling for students' baseline academic skill appeared to make the most difference for continuous reading scores – reducing the magnitude by more than 50%, for example, in Grades 3 and 4. Some caution should be taken when making inferences based on these comparisons, however, given many overlapping CIs and the need for replication.

In all continuous score models, main effect sizes would be considered small in both math and reading ($g < .2$) according to Cohen's 1988 criterion. According to criterion by Chen, Cohen, and Chen (2010), math and reading odds ratios in Grades 3 and 4 would be considered small. Grade 5 odds ratios would be considered small-to-medium in math and medium-to-large in reading. Researchers, however, have cautioned against interpretation of effect size estimates without contextualization relative to previous studies with similar interventions and methodologies (Ferguson, 2009), including comparable approaches to statistical controls (Hedges, 2008). As such, it is important to consider observed effect sizes from the current study relative to those from recent longitudinal efficacy trials employing randomized designs, student-level state test outcome data, and statistical controls for baseline student skill.

Situating SSIS-CIP State Test Findings

Overall, the pattern and magnitude of observed effects in the current sample appear consistent with results reported in methodologically-similar previous studies (though current results tend to be more positive in several cases). For example, Rimm-Kaufmann et al. (2014) did not find statistically significant differences in Grade 5 state test performance of students (controlling for baseline math achievement) between students in Responsive Classroom schools for 3 years and students in control schools (g = –.13 in math and –.06 in reading). However, a mediation analysis revealed that fidelity of Responsive Classroom implementation was positively and significantly related to achievement (g = .27 in math and .30 in reading). Similarly, Jones et al. (2010, 2011) reported nonsignificant small and negative main effects on Grades 3 and 4 state test scores after 1-2 years of exposure to 4Rs (after controlling for baseline aggression). Finally, Espelage et al. (2016) reported that Grade 8 state test effects for a sample of students with disabilities exposed to Second Step over multiple years were small, mixed, and not statistically significant (g = -.09 in math and .08 in reading).

Using statistical controls to account for students' pre-intervention achievement can improve the precision of impact estimates (Bloom, Richburg-Hayes, & Black, 2007; Somers et al., 2011). Not surprisingly, results from the current samples indicated that the magnitude of many effect sizes increased (particularly for continuous reading scores) when pretests were omitted from the model. Reporting the outcome as proficiency levels rather than continuous scores also affected interpretation in some cases. As an example, only one model specification for Grade 5 reading yielded a difference between treatment and control that did not fall below the

.05 threshold of statistical significance – and that model is the specification that we believe best addresses threats to internal validity. When student academic skill is controlled and continuous scores are used, SSIS-CIP exposure does not appear to impact student achievement in Grade 5 reading. However, when the academic pretest is omitted and/or a proficiency level outcome is used, statistically significant differences are found. Similarly, in some recent RCTs of universal SEL programs in which no student-level skill covariate was included in the analyses and proficiency outcomes were utilized, effects on state test outcomes appeared more likely to have at least some statistically significant, positive, and/or larger effects. For example, Schonfeld et al. (2015) reported statistically significant main effects on state test proficiency status for the treatment condition (at least 2 years of PATHS + teacher training & coaching) in Grade 4 (math OR = 1.91, reading OR = 1.72), but not in Grade 5 (math OR = 1.21, reading OR = 1.06) or Grade 6 (math OR = 1.38, reading OR = 1.13). In that study, the lack of baseline data for statistical controls was noted as a limitation.

Similarly, level of analysis may also be worth consideration. In a study of the Positive Action program in which the average of 2 years of school-level aggregated proficiency data was used as a baseline covariate, analyses indicated a school-level program effect (aggregated across Grades 3-8) for math at the $p = .07$ level, but no significant main effect was found in reading. (Standardized difference between pre and posttest in between-group differences was .38 in math and .22 in reading [Bavarian et al., 2013]). In another study (Snyder et al., 2010), Positive Action was found to have statistically significant main effects on percentage of students proficient at the school level (N=20, g = .69 in math and .72 in reading), according to matched-paired *t* tests. The study did not employ a repeated measures design, however, and no student-level information was available for the analyses (aggregated school-level archival baseline and outcome state test data were used).

While similar to (if not more positive than) findings reported by recent SEL-state test impact studies employing similar methodologies, the current results (i.e., continuous score gs $\leq$ .15 and improvement indices below 6%) are smaller than those in SEL meta-analyses reporting SEL achievement effects. In Durlak et al. (2011) and Sklad et al. (2012), immediate achievement outcomes from universal SEL programs were reported as g = .27 (improvement index = 11%) and d = .46, respectively. Similarly, follow-up outcomes have been reported as d = .26 in Sklad et al. (2012) and g = .33 (improvement index = 12.9%) in Taylor et al. (2017). There are several possible reasons for differences in these meta-analytic findings as compared to results of the current study and other similar RCTs. We employed a randomized design, used state test scores as our measure of achievement, and controlled for a measure of baseline academic skill at the student-level. In contrast, less than half (47%) of the included studies (conducted from 1955-2007) in Durlak et al.'s (2011) meta-analysis employed randomized designs, and 56% of studies (from 1995-2008) in Sklad et al. (2012) included some form of random assignment. Studies included in Durlak et al. (2011) and Taylor et al. (2017) used student grades, GPAs, or nationally-standardized tests to measure student academic performance. In the meta-analysis by Sklad et al. (2012), teacher ratings of academic competence also were included. Although not explicitly reported as a design feature in any of these meta-analyses, of the eight studies in the Taylor et al. (2017) review that reported follow-up academic outcomes, only four used measures of baseline student achievement in analyses. Notably, without controlling for baseline academic skill, the 95% CI ranges for all of our outcomes (Grade 3-5 reading and math) include the mean effect size (g = .27) reported in Durlak et al. (2011).

Students participating in the current study were exposed to the SSIS-CIP for a relatively brief amount of time (approximately 10-12 hours over 12 weeks) at a cost of approximately $19 per student (see Hunter et al., 2018) whereas some other programs require a much larger resource investment. It is possible that the SSIS-CIP lacked sufficient dosage to substantially move the needle on student achievement; however, the state test results observed in this study (all but one positive effect size) and relatively small amount of instructional time required for implementation may be seen as acceptable to stakeholders given its efficacy in improving social behavior in the classroom (e.g., DiPerna et al., 2015). The current study reports findings from one sample and one program, and results require replication (e.g., Makel & Plucker, 2014). However, collective findings across similar RCTs suggest some convergence around the possibility that SEL programs, on average, have small impacts on state test scores across students, math and reading domains, and grade levels. For schools considering the SSIS-CIP, the current findings may help address potential concerns about possible negative state test outcomes resulting from reallocation of academic instructional time; however, they also may temper conclusions from earlier meta-analytic research (e.g., Durlak et al., 2011) about large academic gains resulting from SEL programs – at least as measured by state test scores.

Limitations

Although this study addresses a gap in the literature, it does have several limitations. The state test subsample sizes decreased at each subsequent follow-up point given the cohort design and duration of the study. The current results are based upon data collected during follow-up to an RCT that was powered to detect immediate student outcomes. As a result, the current analyses were underpowered in detecting very small longitudinal effects (i.e., ES < .1). While baseline equivalence between treatment and control groups was demonstrated for most student variables, it was not present for the reading pretest and supplemental services. However, the study appears to meet WWC guidelines for a low-attrition RCT, and we controlled for all available baseline characteristics in impact models to minimize possible bias. Resources did not allow us to measure student impact in later grades (i.e., middle and high school), so long-term "sleeper effects" (whereby childhood interventions may promote or offset behaviors in future developmental periods) were unable to be explored. Differential academic achievement effects based on implementation factors such as dosage and fidelity remain unaddressed in the context of this study, given a lack of variability in these attributes. However, all SSIS-CIP classrooms in the current sample were exposed to the SSIS-CIP for approximately 10-12 hours, and as the intervention was straightforward to implement and relatively brief, fidelity was observed to be high across the sample. Finally, the study took place in small urban and rural school communities, and the majority of students and teachers were white, potentially limiting generalizability of findings to classrooms in larger districts in urban or suburban settings.

Future Directions for SEL-Achievement Research

Jones and Doolittle (2017) described sometimes contradictory findings in the current SEL research base, pointing to lack of precision in both the conceptualization and measurement of SEL as a challenge for the field. As one example, they noted misalignment between specific social-emotional targets of SEL interventions and outcomes that are measured as indicators of success (such as broad academic achievement). Future research efforts can advance our understanding of the academic implications of universal SEL by becoming more precise and aligned – conceptually, linguistically, and methodologically. Our review of the extant SEL impact research and results of the current study reveal some preliminary methodological considerations; however, reviews (e.g., Cheung & Slavin, 2016) and/or meta-analyses (e.g., Corcoran et al., 2018; Wigelsworth et al., 2016) that more systematically investigate the relationship between methodology and study effect sizes are an important next step for the SEL research community. Nonetheless, SEL researchers should include measures of baseline student-level academic skills and control for these characteristics in future studies, even those with high-quality randomized designs, to improve precision, reduce bias, and provide more accurate impact estimates (U.S. DOE, 2017a). We also encourage researchers to continue collecting follow-up academic outcomes (see Somers et al., 2011) and being planful with respect to incorporating state test scores into research studies (i.e., at which level of analysis, using what metric).

While the impact of SEL on achievement reported in large meta-analyses has helped legitimize the importance of fostering student social-emotional competence among both school-based practitioners and policymakers, it may have had an unintended consequence of narrowing public perceptions of success toward a focus on academic outcomes. Instead, it would be worth re-emphasizing SEL as a gain in and of itself, without the need to qualify its value through links to academic achievement. In this vein, future research should continue to consider not only specific social-emotional outcomes (i.e., those targeted by the intervention's theory of change) but also broad indicators of student success and well-being that go beyond test score metrics. For example, indices of positive developmental trajectories may include peer relationships, attendance, reduced need for specialized services, degree attainment, and mental health (Taylor et al., 2017). Emerging research suggests that SEL may have impacts beyond student-level outcomes, such as classroom climate (Gregory et al., 2016) and teacher beliefs (Domitrovich et al., 2016), so such effects should continue to be explored by researchers and considered by practitioners and policymakers.

Conclusions

In sum, the current study represents a first step in understanding the impact of the SSIS-CIP on students' subsequent state test performance. Findings mirror other recent RCTs suggesting that the impact of universal SEL programming on state test performance does not yield statistically significant differences with generally small observed differences. While observed main effect sizes for the SSIS-CIP condition were mostly positive, confidence intervals extended in both the positive and negative directions. The differences between results from high-quality meta-analyses on the impact of SEL on academic achievement and recent evidence from several RCTs raises several important considerations regarding the conceptualization of SEL, universal SEL program attributes, and academic skill measurement (baseline and outcome). While results from the current study require replication, they may help assuage the concerns of stakeholders regarding the lost academic instruction resulting from SSIS-CIP implementation in primary classrooms. At the same time, current findings help inform the ongoing conversation about the value of SEL in schools, including how to best conceptualize, measure, interpret, and communicate effectiveness in both research and practice.

References

Bavarian, N., Lewis, K. M., Dubois, D. L., Acock, A., Vuchinich, S., Silverthorn, N., … Flay, B. R. (2013). Using social-emotional and character development to improve academic outcomes: A matched-pair, cluster-randomized controlled trial in low-income, urban schools. *Journal of School Health, 83*, 771–779. doi:10.1111/josh.12093

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for students that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29,* 30-59. doi:10.3101/0162373707299550

Bridgeland, J., Bruce, M., & Hariharan, A. (2013). *The Missing Piece: A National Teacher Survey on How Social and Emotional Learning Can Empower Children and Transform Schools.* Chicago, IL: Civic Enterprises, Hart Research Associates, and the Collaborative for Academic, Social, and Emotional Learning.

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

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitude of odds ratios in epidemiological studies. *Communications in Statistics – Simulation and Computation, 39,* 860-864. doi:10.1080/03610911003650383

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*, 283–292. doi:10.3102/0013189X16656615

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,* 56–72. doi:10.1016/j.edurev.2017.12.001

Cohen, J. (1988). *Statistical power for analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collaborative for Academic, Social, and Emotional Learning (2019). *Approaches.* Retrieved from http://www.casel.org/what-is-sel/approaches/

Committee for Children. (2008). *Second Step: Student success through prevention program.* Seattle, WA: Author.

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

DiPerna, J. C., Lei, P., Bellinger, J., & Cheng, W. (2015). Efficacy of the Social Skills Improvement System Classwide Intervention Program (SSIS-CIP) Primary Version, *School Psychology Quarterly*, *30*, 123–141. doi:10.1037/spq0000079

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*, 189–203. doi: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,* 1-16. doi:10.1037/edu0000191

Domitrovich, C. E., Bradshaw, C. P., Berg, J. K., Pas, E. T., Becker, K. D., Musci, R., … Ialongo, N. (2016). How do school-based prevention programs impact teachers? Findings from a randomized trial of an integrated classroom management and social-emotional Program. *Prevention Science*, 325–337. doi:10.1007/s11121-015-0618-z

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. doi:10.1111/j.1467-8624.2010.01564.x

Elliott, S. N., & Gresham, F. M. (2007*). Social Skills Improvement System: Classwide Intervention Program*. Minneapolis: Pearson.

Enders, C. K., Keller, B. T., & Levy, R. (2017). A chained equations imputation approach for multilevel data with categorical and continuous variables. *Psychological Methods*, Advance online publication. doi:10.1037/met0000148.

Espelage, D. L., Rose, C. A., & Polanin, J. R. (2016). Social-emotional learning program to promote prosocial and academic skills among middle school students with disabilities. *Remedial & Special Education*, *37*, 323–332. doi:10.1177/0741932515627475

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology, Research and Practice, 40*, 532– 538. doi:10.1037/a0015808

Flay, B. R., Allred, C. G., & Ordway, N. (2001). Effects of the Positive Action Program on achievement and discipline: Two matched-control comparisons. *Prevention Science, 2*, 71-89. doi:10.1023/A:1011591613728

Greenberg, M., & Abenavoli, R. Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness, 10,* 40-67. doi: 10.1080/19345747.2016.1246632

Gregory, A., Hafen, C. A., Ruzek, E., Allen, J. P., & Pianta, R. C. (2016). Closing the racial discipline gap in classrooms by changing teacher practice, *School Psychology Review, 45*(2), 171–191.

Heckman, J. J., & Kautz. T. (2012). Hard evidence on soft skills. *Labour Economics, 19,* 451-464.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*, 167–171. doi:10.1111/j.1750-8606.2008.00060.x

Hunter, L. J., DiPerna, J. C., Hart, S. C., & Crowley, M. (2018). At what cost? Examining the cost effectiveness of a universal social-emotional learning program. *School Psychology Quarterly, 33,* 147-154. doi:10.1037/spq0000232

Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health*, *105*, 2283–2290. doi:10.2105/AJPH.2015.302630

Jones, S. M., Barnes, S. P., Bailey, R., & Doolittle, E. J. (2017). Promoting social and emotional competencies in elementary school. *The Future of Children, 27*(1), 49-72.

Jones, S. M., & Bouffard, S. M. (2012). Social and emotional learning in schools: From programs to strategies. *Social Policy Report 26*(4), 1-33.

Jones, S. M., Brown, J. L., Hoglund, W. L. G., & Aber, J. L. (2010). A school-randomized clinical trial of an integrated social-emotional learning and literacy intervention: Impacts after 1 school year. *Journal of Consulting and Clinical Psychology*, *78*, 829–842. doi:10.1037/a0021383

Jones, S. M., Brown, J. L., & Lawrence Aber, J. (2011). Two-year impacts of a universal school-based social-emotional and literacy intervention: An experiment in translational developmental research. *Child Development, 82*, 533–554. doi:10.1111/j.1467-8624.2010.01560.x

Jones, S. M., & Doolittle, E. J. (2017). Social and emotional learning: Introducing the issue. *The Future of Children, 27*(1), 3-12. Retrieved from http://www.futureofchildren.org.

Keller, B. T., & Enders, C. K. (2017). Blimp User's Manual (Version 1.0). Los Angeles, CA.

Kusche, C.A., & Greenberg, M. T. (1994). *The PATHS Curriculum.* Seattle, WA: Developmental Research and Programs.

Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2017). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*. SAGE.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*, 304–316.

Malecki, C. K., & Elliot, S. N. (2002). Children's social behaviors as predictors of academic achievement: A longitudinal analysis. *School Psychology Quarterly, 17*, 1–25.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Caspi, A. (2011). "A gradient of childhood self-control predicts health, wealth, and public safety," *Proceedings of the National Academy of Sciences of the United States of America*, 2693–98. doi:10.1073/pnas.1010076108

Morningside Center for Teaching Social Responsibility. (2001). *The 4Rs (Reading, Writing, Respect, & Resolution): A teaching guide.* New York, NY: Author.

Northeast Foundation for Children (2007). *Responsive Classroom.* Turner Falls, MA: Author.

Pennsylvania Department of Education (2012-2014). *Technical report for the Pennsylvania System of School Assessment.* Data Recognition Corporation.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173–185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2005). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software [Computer program] (Version March 09, 2005).

Renaissance Learning (2009). *STAR Math Technical Manual.* Wisconsin Rapids, WI: Renaissance Learning Inc.

Renaissance Learning (2010). *STAR Reading Technical Manual.* Wisconsin Rapids, WI: Renaissance Learning Inc.

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

Schonfeld, D. J., Adams, R. E., Fredstrom, B. K., Weissberg, R. P., Gilman, R., Voyce, C., … Speese-Linehan, D. (2015). Cluster-randomized trial demonstrating impact on academic achievement of elementary social-emotional learning. *School Psychology Quarterly*, *30*, 406–420. doi:10.1037/spq0000099

Sklad, M., Diekstra, R., De Ritter, M., Ben, J., & Gravesteijn, C. (2012). Effectiveness of school-based universal social, emotional, and behavioral programs: Do they enhance students' development in the area of skill, behavior, and adjustment? *Psychology in the Schools, 49*, 892-909. doi:10.1002/pits

Snyder, F., Flay, B., Vuchinich, S., Acock, A., Washburn, I., Beets, M., & Li, K.-K. (2010). Impact of a social-emotional and character development program on school-level indicators of academic achievement, absenteeism, and disciplinary outcomes: A matched-pair, cluster-randomized, controlled trial. *Journal of Research on Educational Effectiveness, 3*, 26-55. doi:10.1080/19345740903353436

Social and Character Development Research Consortium. (2010, October). *Efficacy of schoolwide programs to promote social and character development and reduce problem behavior in elementary school children* (NCER 2011-2001). Washington, DC: National Center for Education Research, Institute of Education.

Somers, M., Zhu, P., & Wong, E. (2011). *Whether and how to use state tests to measure student achievement in a multi-state randomized experiment: An empirical assessment based on four recent evaluations*. (NCEE Reference Report 2012-4015). Washington DC: Institute of Education Sciences, U.S. Department of Education.

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*, 1156–1171. doi:1111/cdev.12864

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2017a). *What Works Clearinghouse: Procedures handbook* (Version 4.0).

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2017b). *What Works Clearinghouse: Standards handbook* (Version 4.0).

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician, 70*, 129–133. doi:10.1080/00031305.2016.1154108

Whitehurst, G. J. (2019, Summer). Should schools embrace social-emotional learning? A prevalence of ‘policy-based evidence-making.’ *Education Next*, 68-73.

Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., ten Bokkel, I., Tate, K., & C. Emery (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*, 347-376. doi:10.1080/0305764X.2016.1195791

Table 1

Demographic Variables and Math Test Scores by Condition and Grade of State Test Administration

	Grade 3		Grade 4		Grade 5	
	SSIS-CIP (*N*=244)	Control (*N*=187)	SSIS-CIP (*N*=158)	Control (*N*=147)	SSIS-CIP (*N*=72)	Control (*N*=67)
Demographic Variable [a]						
Gender (male)	44.26	41.71	43.04	40.82	45.83	34.33
Race (white)	73.36	79.89	76.58	81.63	72.22	88.06
Special education	10.16	7.49	8.48	7.07	9.72	8.96
Supplemental services	15.08	24.06	12.91	24.08	19.44	17.91
Math Test Scores						
STAR (pretest)[b]	448.33	444.26	458.65	441.12	460.62	443.01
	(113.76)	(88.58)	(103.97)	(91.05)	(118.65)	(74.61)
State Test - Continuous Score[c]	1332.03	1306.74	1488.59	1414.15	1513.40	1400.40
	(179.28)	(173.73)	(249.10)	(218.15)	(262.69)	(221.99)
State Test - Proficiency Status[d]	79.51	78.61	82.91	76.87	77.78	62.69

Note. Analysis samples include imputed baseline data when missing.

[a]Average percentages across all imputed samples for each demographic variable.

[b]Mean (*SD)* averaged across all imputed samples.

[c]Mean (*SD*).

[d]Percentage of sample meeting proficiency cutoff.

Table 2

Model Estimates (Standard Errors) for SSIS-CIP Treatment Effect on Math State Test Continuous Scores and Proficiency Status

	Continuous Scores			Proficiency Status		
	Grade 3 (*N*=431)	Grade 4 (*N*=305)	Grade 5 (*N*=139)	Grade 3 (*N*=431)	Grade 4 (*N*=305)	Grade 5 (*N*=139)
Intercept	1302.54** (45.67)	1367.45** (89.37)	1373.69** (78.76)	1.58† (.93)	.98 (1.41)	1.24 (.92)
Student-level covariates						
Academic pretest	1.06** (.07)	1.21** (.12)	.96** (.16)	.02** (.00)	.01** (.00)	.01** (.00)
Gender (male)	-3.74 (12.71)	-8.45 (19.16)	34.47 (27.97)	-.09 (.34)	.44 (.39)	.05 (.50)
Race (white)	28.84 (17.76)	14.18 (29.63)	-6.40 (44.38)	-.31 (.45)	-.20 (.59)	.18 (.75)
Supplemental services	-46.26** (17.11)	-85.78** (28.12)	-20.46 (38.92)	-.88* (.39)	-.74 (.46)	-1.33* (.62)
Special education	29.22 (24.86)	-82.68† (44.94)	-6.66 (54.26)	-.04 (.57)	-.82 (.73)	-.33 (.88)
Treatment effect						
SSIS-CIP	22.28† (13.03)	32.54 (20.46)	38.51 (29.97)	.14 (.35)	.16 (.40)	.72 (.51)
p value	.09	.11	.20	.69	.70	.16

Note. Cohort and school indicators are controlled for in the model but not reported.
†$p < .10$; *$p < .05$; **$p < .01$.

Table 3

Effect Sizes for Math and Reading State Test Scores when Estimated with and without Academic Pretest Covariate

	Grade 3 (*N*=431/432)		Grade 4 (*N*=305/306)		Grade 5 (*N*=139/138)	
	Pretest	No Pretest	Pretest	No Pretest	Pretest	No Pretest
			Math			
Continuous scores						
Hedges' *g*	.13†	.17†	.14	.16	.15	.13
[95% CI]	[-.06, .32]	[-.02, .36]	[-.09, .36]	[-.06, .39]	[-.18, .49]	[-.21, .46]
Improvement index	5.17	6.75	5.57	6.36	5.96	5.17
Proficiency status						
Odds ratio	1.15	1.04	1.17	1.19	2.05	2.03
[95% CI]	[.58, 2.27]	[.60, 1.80]	[.53, 2.56]	[.60, 2.36]	[.76, 5.52]	[.68, 6.07]
			Reading			
Continuous scores						
Hedges' *g*	-.09	.10	.05	.17†	.10	.15*
[95% CI]	[-.28, .10]	[-.09, .30]	[-.17, .28]	[-.05, .40]	[-.24, .43]	[-.19, .48]
Improvement index	-3.59	3.98	1.99	6.75	3.98	5.96
Proficiency status						
Odds ratio	1.01	1.26	1.34	1.77†	3.60*	4.87**
[95% CI]	[.57, 1.81]	[.76, 2.08]	[.64, 2.80]	[.90, 3.45]	[1.10, 11.82]	[1.39, 17.08]

Note. Effect sizes adjusted for covariates. (*N* = Math/Reading).
†$p < .10$; *$p < .05$; **$p < .01$.

Table 4

Demographic Variables and Reading Test Scores by Condition and Grade of State Test Administration

	Grade 3		Grade 4		Grade 5	
	SSIS-CIP (*N*=245)	Control (*N*=187)	SSIS-CIP (*N*=161)	Control (*N*=145)	SSIS-CIP (*N*=72)	Control (*N*=66)
Demographic Variable [a]						
Gender (male)	44.49	41.71	42.86	40.69	45.83	33.33
Race (white)	72.41	80.64	77.89	82.21	72.22	87.88
Special education	10.53	7.70	8.57	7.03	9.72	9.09
Supplemental services	15.02	24.28	12.67	22.48	19.44	18.18
Reading Test Scores						
STAR (pretest)[b]	252.16	217.98	267.22	220.14	278.83	222.92
	(133.76)	(102.59)	(128.52)	(99.58)	(137.61)	(86.85)
State Test - Continuous Score[c]	1323.75	1314.13	1372.06	1311.43	1365.50	1285.17
	(166.56)	(138.22)	(216.42)	(197.54)	(215.49)	(180.10)
State Test - Proficiency Status[d]	71.02	69.52	72.67	62.76	69.44	51.52

Note. Analysis samples include imputed baseline data when missing.

[a]Average percentages across all imputed samples for each demographic variable.

[b]Mean (*SD)* averaged across all imputed samples.

[c]Mean (*SD*).

[d]Percentage of sample meeting proficiency cutoff.

Table 5

Model Estimates (Standard Errors) for SSIS-CIP Treatment Effect on Reading State Test Continuous Scores and Proficiency Status

	Continuous Scores			Proficiency Status		
	Grade 3 (*N*=432)	Grade 4 (*N*=306)	Grade 5 (*N*=138)	Grade 3 (*N*=432)	Grade 4 (*N*=306)	Grade 5 (*N*=138)
Intercept	1345.09** (34.04)	1417.70** (71.94)	1253.91** (62.13)	2.25** (.82)	2.65 (1.62)	-.34 (1.09)
Student-level covariates						
Academic pretest	.80** (.05)	1.00** (.08)	.60** (.13)	.01** (.00)	.02** (.00)	.01** (.00)
Gender (male)	-39.86** (10.37)	-87.97** (15.89)	-60.03* (24.63)	-.62* (.29)	-1.44** (.39)	-1.32* (.57)
Race (white)	20.36 (14.76)	24.46 (25.09)	20.65 (38.21)	.04 (.40)	.39 (.54)	.91 (.79)
Supplemental services	-12.44 (14.66)	-79.79** (23.84)	-19.37 (34.58)	-.52 (.36)	-1.04* (.48)	.25 (.71)
Special education	-22.05 (19.92)	-108.09** (32.54)	-42.09 (46.28)	-.49 (.50)	-1.37† (.73)	-.77 (1.04)
Treatment effect						
SSIS-CIP	-14.17 (10.93)	10.66 (16.51)	19.60 (26.31)	.01 (.30)	.29 (.38)	1.28* (.61)
p value	.19	.52	.46	.97	.43	.03

Note. Cohort and school indicators are included in the model but not reported.
†$p < .10$; *$p < .05$; **$p < .01$.

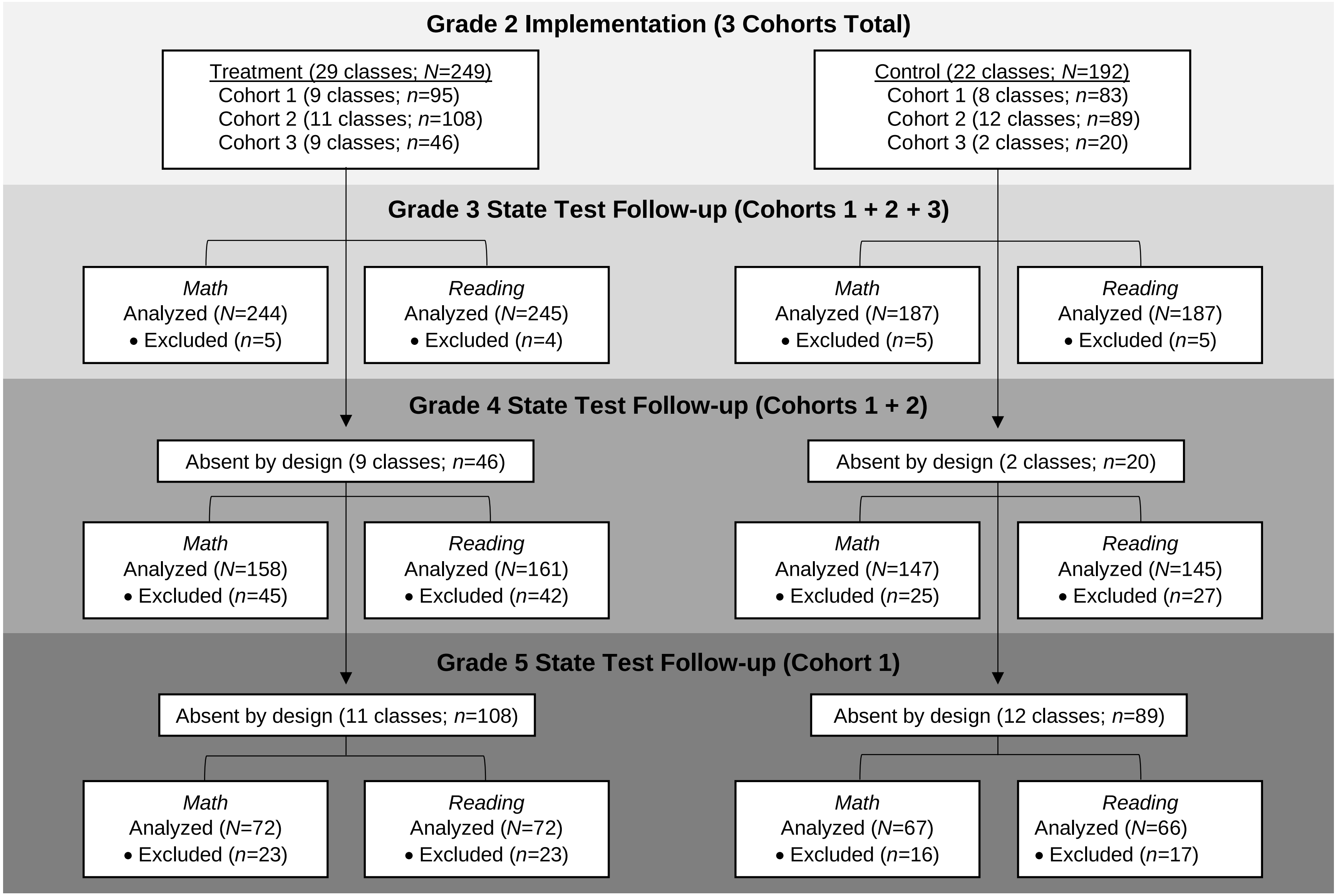

Figure 1. Participant flow through study phases by condition. Analyzed *N*s included cases with imputed baseline data when missing. Excluded cases were attrition (missing outcome data). Absent by design were due to planned loss of cohorts over time.

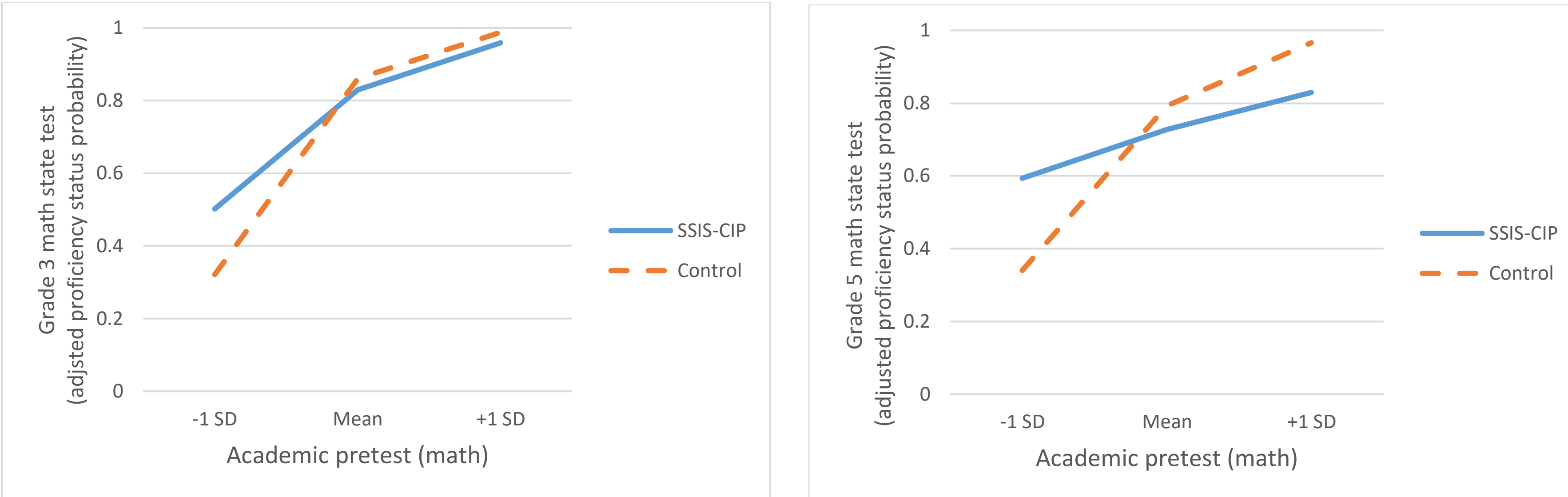

Figure 2. Interactions between treatment condition and baseline skills on Grade 3 and Grade 5 math state test proficiency scores.

Appendix

Table A1

Model Estimates (Standard Errors) for SSIS-CIP Treatment Effect on Math State Test Continuous Scores and Proficiency Status without Math Pretest Covariate

	Continuous Scores			Proficiency Status		
	Grade 3 *N*=431	Grade 4 *N*=305	Grade 5 *N*=139	Grade 3 *N*=431	Grade 4 *N*=305	Grade 5 *N*=139
Intercept	1386.63** (53.95)	1424.67** (105.86)	1349.57** (93.04)	2.54** (.78)	1.64 (1.48)	.67 (.96)
Student-level covariates						
Gender (male)	-21.01 (15.77)	-15.11 (22.34)	41.66 (31.50)	-.37 (.27)	.22 (.35)	.03 (.48)
Race (white)	38.27† (22.50)	21.74 (34.33)	-28.23 (49.79)	-.02 (.39)	.11 (.51)	-.17 (.74)
Supplemental services	-109.74** (21.59)	-181.85** (31.22)	-67.75 (43.01)	-1.55** (.35)	-1.69** (.41)	-1.70** (.60)
Special education	-102.30** (29.06)	-270.17** (44.47)	-110.21† (57.51)	-1.66** (.42)	-2.62** (.56)	-1.32† (.79)
Treatment effect						
SSIS-CIP	30.28† (16.35)	37.94 (23.50)	40.30 (33.82)	.04 (.28)	.18 (.35)	.71 (.50)
p value	.06	.11	.24	.89	.61	.16

Note. Cohort and school indicators are controlled for in the model but not reported.
†$p < .10$; *$p < .05$; **$p < .01$.

Table A2

Model Estimates (Standard Errors) for SSIS-CIP Treatment Effect on Reading State Test Continuous Scores and Proficiency Status without Reading Pretest Covariate

	Continuous Scores			Proficiency Status		
	Grade 3 *N*=432	Grade 4 *N*=306	Grade 5 *N*=138	Grade 3 *N*=432	Grade 4 *N*=306	Grade 5 *N*=138
Intercept	1355.24** (42.03)	1337.30** (89.82)	1176.95** (55.30)	1.93** (.71)	.44 (1.55)	-1.15 (1.03)
Student-level covariates						
Gender (male)	-42.52** (13.27)	-83.51** (19.11)	-37.54** (10.86)	-.63* (.25)	-1.21** (.34)	-.95† (.51)
Race (white)	21.43 (18.78)	21.38 (29.99)	8.54 (16.87)	.09 (.34)	.28 (.50)	.74 (.75)
Supplemental services	-90.90** (17.67)	-193.50** (26.72)	-52.11** (14.67)	-1.53** (.31)	-2.33** (.45)	-.70 (.62)
Special education	-106.12** (24.58)	-229.33** (37.77)	-102.03** (19.81)	-1.56** (.41)	-2.60** (.63)	-1.91* (.94)
Treatment effect						
SSIS-CIP	16.18 (13.74)	36.62† (20.00)	29.79* (11.66)	.23 (.26)	.57† (.34)	1.58** (.57)
p value	.24	.07	.01	.38	.096	.006

Note. Cohort and school indicators are included in the model but not reported.
†$p < .10$; *$p < .05$; **$p < .01$.

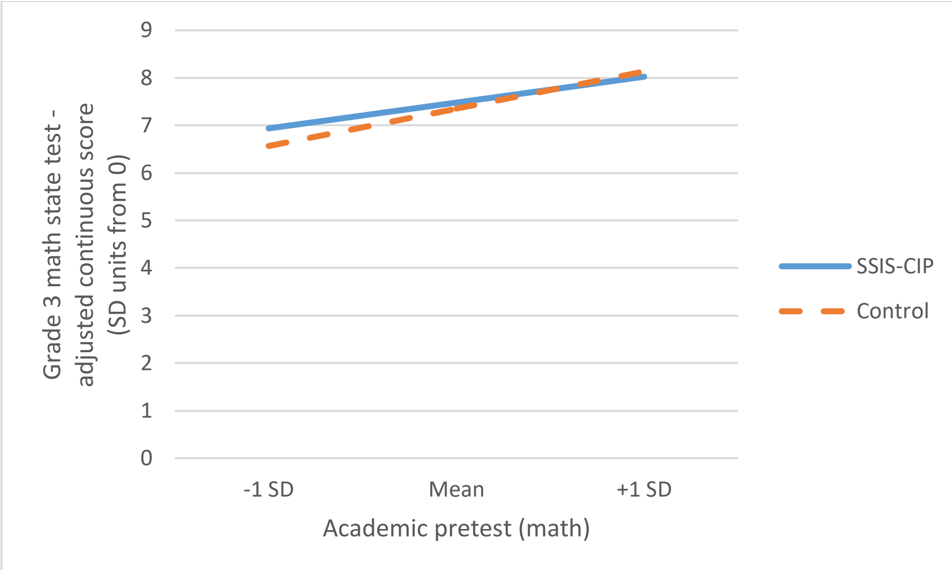

Figure 1A. Interaction between treatment condition and baseline math score on Grade 3 math state test continuous scores.