Journal of School Psychology 53 (2015) 463–477

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Promoting social–emotional competence: An evaluation of the elementary version of Second Step®

Sabina Low [a,*], Clayton R. Cook [b], Keith Smolkowski [c], Jodie Buntain-Ricklefs [b]

[a] Arizona State University, Phoenix, AZ, United States
[b] University of Washington, Seattle, WA, United States
[c] Oregon Research Institute, Eugene, OR, United States

ARTICLE INFO

Article history:
Received 5 November 2014
Received in revised form 18 September 2015
Accepted 23 September 2015
Available online xxxx

Keywords:
Social-emotional learning
Social skills
Efficacy trial
Second Step®

ABSTRACT

Research has consistently linked social–emotional skills to important educational and life outcomes. Many children begin their school careers, however, without the requisite social and emotional skills that facilitate learning, which has prompted schools nationwide to adopt specific curricula to teach students the social–emotional skills that enable them to maintain optimal engagement in the learning process. *Second Step®* is one of the most widely disseminated social–emotional learning (SEL) programs; however, its newly revised version has never been empirically evaluated. The purpose of this study was to conduct a randomized controlled trial investigating the impact of the 4th Edition *Second Step®* on social–behavioral outcomes over a 1-year period when combined with a brief training on proactive classroom management. Participants were kindergarten to 2nd grade students in 61 schools (321 teachers, 7300 students) across six school districts. Hierarchical models (time × condition) suggest that the program had few main effects from teacher-reported social and behavioral indices, with small effect sizes. The majority of significant findings were moderated effects, with 8 out of 11 outcome variables indicating the intervention-produced significant improvements in social–emotional competence and behavior for children who started the school year with skill deficits relative to their peers. All the significant findings were based on teacher-report data highlighting a need for replication using other informants and sources of data. Findings provide program validation and have implications for understanding the reach of SEL programs.

1. Introduction

Although school readiness and success is most often associated with cognitive attributes and academic milestones, there is increasing evidence that social–emotional skills—in the form of understanding emotions of self and other, regulating emotions, controlling attention, problem solving, and engaging in prosocial behaviors—operate alongside and in conjunction with cognitive skills to facilitate school success (Cambourne, 2002; Denham, 2006; Denham, Bassett, & Zinsser, 2012). Social–emotional skills combine to enable social–emotional competence, which represents an overall evaluation of a child's ability to meet the social and emotional demands from the environment (Gresham, 1986; Merrell & Gueldner, 2012). A recent meta-analysis of 213 studies examining the impact of different social–emotional learning (SEL) curricula indicated that such programs are not only associated with significant improvements in students' social–emotional skills, but they were associated with improvements on end-of-the-

* Corresponding author at: Arizona State University, T. Denny Sanford School of Social and Family Dynamics, P.O. Box 873701, Tempe, AZ, 85287.
E-mail address: Sabina.low@asu.edu (S. Low).
Action Editor: Jina Yoon

http://dx.doi.org/10.1016/j.jsp.2015.09.002
0022-4405/

year academic achievement (i.e., tests and grades; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). Furthermore, research has shown that students' social–emotional skills are a better predictor of future academic performance than is their prior academic performance (Caprara, Barbaranelli, Pastorelli, Bandura, & Zimbardo, 2000; Malecki & Elliot, 2002). This is not altogether surprising, given that many scholars postulate that social interaction (with peers and teachers) is the bridge between instruction and optimized learning (Elias & Haynes, 2008; Zins, Bloodworth, Weissberg, & Walberg, 2007).

It is well documented that there is considerable variation in students' academic readiness when they begin formal schooling (Phillips & Shonkoff, 2000). In a similar vein, students vary widely in their social–emotional readiness (Fantuzzo et al., 2007). Differences in performance upon school entry often do not vanish as students progress through school. Rather, the gaps between students from advantaged and disadvantaged backgrounds tend to increase over time (Brooks-Gunn, Rouse, & McLanahan, 2007; Chatterji, 2006). Indeed, too often, children begin their school careers without the requisite social–emotional competency to facilitate learning (Rimm-Kaufman, Pianta, & Cox, 2000), which may be an increasing liability for socially or cognitively disadvantaged children, given the cumulative benefits of early cognitive-academic supports (Campbell & Ramey, 1995; Ramey & Ramey, 1998). For these reasons, early elementary represents an opportune time to deploy universal prevention efforts that promote early school success (Bernard, 2006). Thus, the focus of the current study was on the implementation of an SEL curriculum for early elementary children (K–2nd grade).

2. *Second Step*® program

One of the most widely disseminated SEL curricula in schools is the *Second Step*® program, which was developed by Committee for Children (CfC), a non-profit organization in Seattle. *Second Step*® is a skills-focused SEL curriculum that emphasizes directly teaching students' skills that strengthen their ability to learn, have empathy, manage emotions, and solve problems. The *Second Step*® logic model (see Fig. 1) stipulates that students who are provided direct instruction in social–emotional skills and opportunities to practice those skills, and receive reinforcement for exhibiting them are likely to experience a range of improved intermediate outcomes, and result in a cascade of positive distal outcomes. Previous studies have found support for the underling logic model of the original *Second Step*® program, though other smaller or less rigorous studies have found mixed or null effects (see Gottfredson et al., 2010 for review). For example, Grossman et al. (1997) conducted a randomized controlled trial of the *Second Step*® program to examine its impact on aggression and positive social behavior among elementary school students. Findings from this study indicated that physical aggression decreased among students in the *Second Step*® classrooms when compared to students in the control classrooms. This improvement was maintained at a 6-month follow-up assessment. Other studies have shown that students receiving *Second Step*® lessons had improved social skills at posttest when compared to children in control classrooms, based on teacher reports (Holsen, Iversen, & Smith, 2009; Holsen, Smith, & Frey, 2008). However, a recent school randomized trial ($n = 12$ schools) by Gottfredson et al. (2010; 3rd Edition *Second Step*®) found no positive or negative effects of Second Step on school achievement or positive behaviors. In the case of this study, however, the control schools were, on average, found to be implementing a fairly high level of SEL programming/supports, making it difficult to clearly differentiate dosage between intervention and control schools.

Recently, CfC has developed and released the 4th Edition of the *Second Step*® program (2012). The new *Second Step*® program includes revised content and materials designed to further enhance student success in school. The most significant change to *Second Step*® is the new content related to teaching students *Skills for Learning*. Specifically, three aspects of self-regulation are addressed in the lessons in the first unit at each grade: attention, working memory, and inhibitory control. Attention refers to the ability to direct, focus, and shift attention while screening out or ignoring distractions (Barkley, 1997; Rueda, Rothbart, McCandliss, Saccomanno, & Posner, 2005). Working memory involves the ability to remember and use information, such as a teacher's directions or the instructions for an activity (Demetriou, Christou, Spanoudis, & Platsidou, 2002). Inhibitory control, also referred to as effortful control, helps children stop automatic but inappropriate responses or actions and remember appropriate behaviors such as raising a hand before speaking (Blair, 2002; Rennie, Bull, & Diamond, 2004). These skill domains are assumed to be important contributors to classroom success, but further research is needed to empirically validate these associations.

There are separate curricula for each grade to enable teachers to deliver instruction that is developmentally appropriate and relevant for their students. The program includes scripted, teacher-friendly lesson cards; posters that outline learned skills; DVDs that illustrate particular skills; brain builder games designed to increase retention and use of skills; and a material binder that includes lessons for teaching and reinforcing skills, skills for learning cards, and home links for families. There are a total of 22 lessons that are organized across four units: (a) Skills for Learning, (b) Empathy, (c) Emotion Management, and

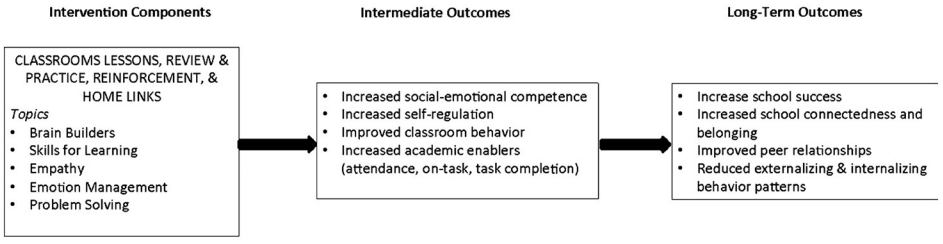

Fig. 1. Logic model for the *Second Step*® program.

(d) Problem Solving. These four units cover a range of skills and behaviors such as being respectful learners, planning to learn, identifying others feelings, showing compassion, making friends, and managing disappointment. Teachers implement the student-focused lessons as part of their normal classroom activities, and each lesson takes roughly 25–40 min, depending on grade level, one time per week.

To date, the newly revised *Second Step*® program has not been evaluated for its ability to produce improved social–emotional and learning outcomes for students. A study evaluating the efficacy of the revised *Second Step*® program is important for a number of reasons. First, schools are increasingly being held accountable to adopt and implement evidence-based programs or interventions (Cook, Tankersley, & Landrum, 2009; Slavin, 2002). Given that *Second Step*® has been purchased by over 25,000 schools, it remains important for schools to know whether they have adopted an efficacious program. Second, although prior research on the original *Second Step*® program has shown its promise for use in schools, evidence has been mixed. Based on the adaptations made to the program, both rigorous and contemporary studies are needed to determine internal and ecological validity on a wider range of targeted outcomes.

3. Current study

In line with the above discussion, the purpose of the proposed study was to conduct a rigorous, large-scale trial across geographically and ethnically diverse schools. The primary aim of this study was to evaluate the main effects of *Second Step*® on early elementary students' social–emotional competencies. Researchers have advocated for the examination of moderators of treatment effectiveness in order to better understand for whom and under what conditions particular programs like *Second Step*® may produce effects (Flay et al., 2005). Thus, our secondary aim was to examine heterogeneity in response to the program, based on baseline student competencies, classroom management, and school-level characteristics. Lastly, we examined basic descriptive information about the influence of implementation on outcomes. Implementation factors are getting increasing attention given their influence on program outcomes (Hanson, Dietsch, & Zheng, 2012; Low, Van Ryzin, Brown, Smith, & Haggerty, 2014; Proctor & Brownson, 2012). The student outcomes were assessed through direct observations in students' learning environment as well as teacher ratings (i.e., surveys) of both problem behaviors and strengths/assets. Contextual data on classroom climate was obtained through teacher self-report, and data on classroom management were obtained through direct observation.

3.1. Hypotheses

The specific hypotheses that guided this study were consistent with prior literature on the positive effects of SEL curricula and the anticipated outcomes produced by the *Second Step*® program. Specifically, the hypotheses were predicated on the notion that social–emotional competencies can be taught and enhanced through instruction, modeling, reinforcement, and opportunities for generalization/repetition. Drawing upon Fig. 1, it was hypothesized that early elementary students who participated in *Second Step*® would demonstrate (a) improvements in teacher-reported social–emotional skills and behaviors, and (b) lower levels of observed disruptive behavior.

Although a universal program in delivery, we hypothesized that some children may respond differently to the program. We first hypothesized that children's response to the program may vary depending on their baseline level of competency. We expected larger gains among children with lower skills (relative to same-age peers); conversely, we did not expect large improvement from students that already possess strong social–emotional skills. We also postulated that teachers' ability to manage their classrooms could influence the implementation quality and effectiveness of any classroom-delivered curriculum, including *Second Step*®. In particular, we hypothesized that students in classrooms with better teacher management skills would be more responsive to Second Step (i.e., realize more benefits due to broader classroom supports). At the exploratory level, we wanted to rule out any potential moderation of study effects by demographic factors (e.g., state, ethnicity, grade level). Because this study took place in two states, with differing laws and environments, and because schools differed in the proportion of ethnic minorities, we examined both state and ethnicity (i.e., percent of white students in the school) as moderators. Lastly, due to developmental differences in kindergarten, first- and second-grade students, it is possible that the grade level could influence the effects of *Second Step*®, and thus, it is important to rule out potential moderating effects.

4. Method

4.1. Participants

This study included students in kindergarten through second grade enrolled in school districts in both Arizona and Washington state. Five school districts participated in Washington state and one district in Arizona (Mesa). School districts ranged from rural to urban settings and were recruited in the Spring of 2012 after approval from the institutional review boards (IRB). Participating school districts, teachers, and parents of the students provided passive consent—in accordance with IRB procedures and district policies.

4.1.1. Recruitment and retention

The Washington site was able to secure and maintain the participation of 41 schools across five school systems (see Fig. 2). On average, 6 randomly selected classrooms participated in data collection from each school. A total of 224 teachers agreed to

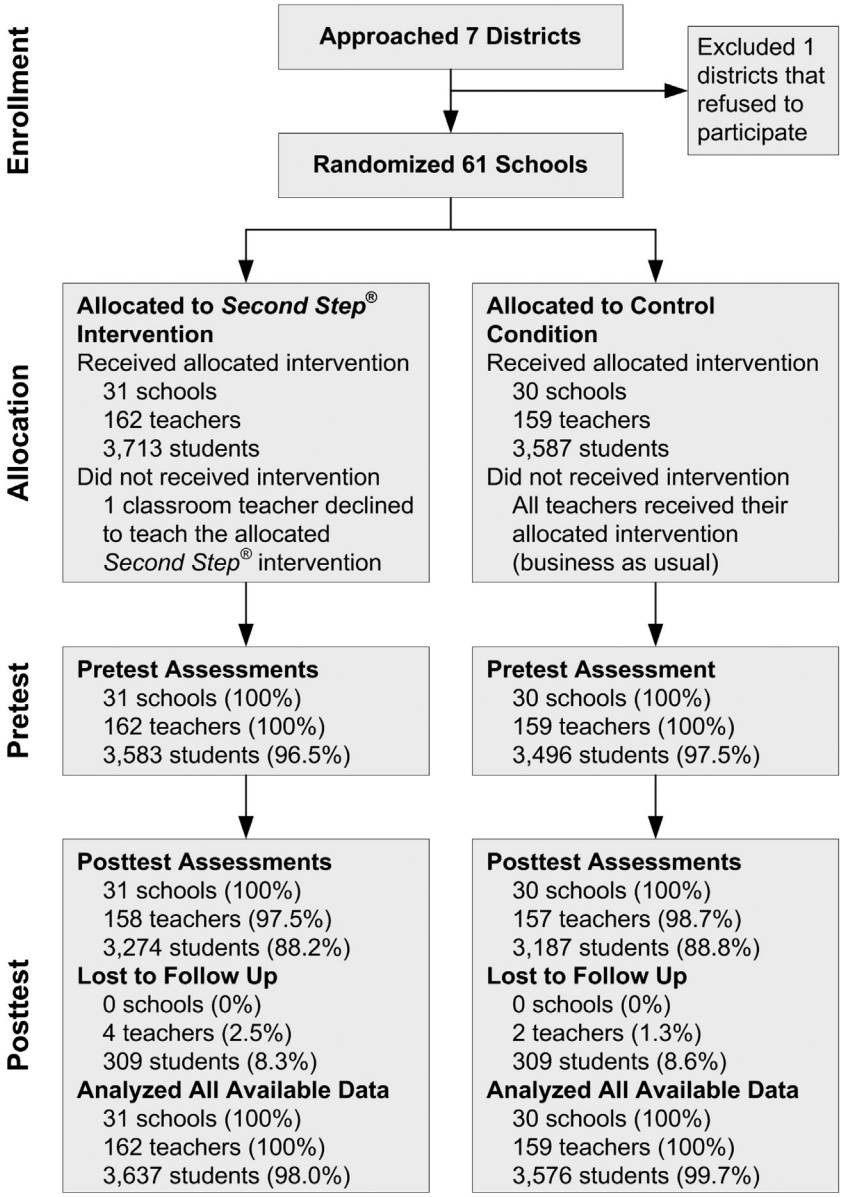

Fig. 2. Research design and participant flow for schools, teachers, and students throughout the cluster-randomized evaluation of *Second Step*®. Percentages use the sample sizes at allocation as the denominator.

participate and passive parental permission was obtained for 4891 students. The Arizona site was able to secure and maintain participation from 20 schools from the Mesa School District. An average of 5 classrooms per school were randomly selected to participate in data collection, with a total of 97 teachers. Passive parental permission was obtained for 2409 students. Approximately, 1% of parents declined across both sites.

All schools continued in the study from baseline through Spring of 2013, and only two teachers had to suspend participation (health or personal reasons). Specifically, the Washington site was able to maintain the participation of 223 teachers and 4232 students from Fall to Spring data collection across the 41 participating schools, for an overall student attrition rate of 13%. The Arizona site was able to maintain the participation of 96 teachers and 2326 students across 20 schools by end of Year 1, for an average student attrition rate of 3%. Attrition primarily encompassed students who moved out of district or to a non-participating school.

With regard to socioeconomic status, 50% and 78% of participating students in Washington and Arizona, respectively, received free and reduced lunch. The racial and ethnicity breakdown of the students was as follows: 45.8% (WA) and 40.1% (AZ) Caucasian, 18.2% (WA) and 0.3% (AZ) Asian, 8.1% (WA) and 5.9% (AZ) African American, 14.7% (WA) and 47.1% (AZ) Latino/a, 1.6% (WA) and 6.3% (AZ) Native American, 1.7% (WA) and 0.3% (AZ) Native Hawaiian or Pacific Islander, 9.9% (WA) and 0% (AZ) reported more than one race, and 20.4% (WA) and 10.1% (AZ) were unknown. This sample of students was relatively representative of the

ethnicity distribution of school-aged children in the United States (US Census, 2011). Teachers' average age and years teaching experience were 43.78 (SD = 12.33) and 15.24 (SD = 9.97), respectively, and 88% were Caucasian, 0.6% black or African American, 2.8% Asian, 0.9% Native Hawaiian or Pacific Islander, 0.6% American Indian or Alaska Native, 4.3% more than one race, and 2.2% other. In addition, 6% of teachers reported they were Hispanic or Latino/a.

4.2. Procedures and design

4.2.1. Overview

The study used a large-scale, matched, randomized-control design with 61 elementary schools randomly assigned within their district to either the early-start (treatment; n = 31) or delayed-start (control; n = 30) conditions (see Fig. 2). Schools within Washington and Arizona were matched on free and reduced lunch and percent of non-white students for design purposes (Murray, 1998). Results from the matching process indicated that there were no significant differences between treatment and control groups on baseline measures (see results section). The present study includes data from only the fall (T1) and spring (T2) assessments in Year 1.

This study represents the middle of the continuum from efficacy to effectiveness (Flay et al., 2005; Smolkowski, Strycker, & Seeley, 2013). It has some features consistent with an efficacy trial, such as efforts to ensure full delivery of program, efforts to ensure all participating teachers have foundation in proactive classroom management, and bi-weekly implementation tips for *Second Step*® and classroom management. However, consistent with effectiveness research, there was (a) limited involvement from the researchers in the day-to-day implementation of the interventions, and (b) professional development training activities consistent with natural ways in which schools would access the material, receive training in how to implement the intervention, and obtain support during implementation, allowing for greater generalization to real-world educational conditions..

4.2.2. Training participation

Two separate, brief trainings were provided to participating early-start schools: the *Second Step*® curriculum (1 h) and Proactive Classroom Management (PCM; 3 h). *Second Step*® was consistent with standard support operations provided by Community for Children, and intended to increase motivation to implement the program, allow teachers to become familiar with the content, and provide specific examples of how to deliver program with fidelity. All early-start schools participated in the training, and all kindergarten, 1st- and 2nd-grade teachers involved in data collection participated in the webinar, as determined by attendance sheets collected by school personnel.

The PCM trainings are not standard practice but were a response to district needs at the time of recruitment. Thus, a very brief overview of classroom strategies were presented so as to meet the needs of schools without providing a sufficiently strong dosage that one would anticipate having a strong impact on classroom behaviors. Specifically, PCM strategies were delivered either via DVD or in-person and focused on skills that would help support, reinforce, and facilitate the engagement in lessons and use of skills covered in *Second Step*®. In particular, the PCM training focused on reviewing and modeling five strategies: (a) positive greetings at the door to precorrect problem behavior, (b) opportunities to respond, (c) effective cueing system to regain attention, (d) strategically and intentionally establishing relationships with all students, and (e) teach, model, and reinforce expected behaviors. These strategies were selected based on prior research demonstrating their efficacy to improve classroom behavior and student engagement (Simonsen, Fairbanks, Briesch, Myers, & Sugai, 2008; Sutherland & Wehby, 2001).

The DVD was a recording of an in-person training delivered to Arizona schools by the second author to maintain consistency across the two training options in format and content. All 10 early-start schools in Arizona were offered in-person training by the second author, and 83 teachers participated. A total of 10 early-start schools and 61 teachers in Washington participated in the PCM training by watching the DVD, and 11 early-start schools and 55 teachers participated in an in-person presentation of the PCM training. Every teacher involved in implementation and data collection participated in the PCM training.

4.2.3. Program implementation

Teachers were asked to complete weekly self-report ratings of implementation (via computer survey on Datstat) to record adherence to the curriculum, engagement, and dosage. Adherence had two components: adherence to the key lesson components (5 items; yes/no) and adaptations/modifications (4 items on 4-point scale, 0 = *Never* to 3 = *Always*; e.g., "to what extent did you leave out parts of the lesson"). Engagement had two components: ratings of the degree of student engagement (3 items on a 4-point scale, 0 = *Not at All* to 3 = *A Lot*; e.g., "to what extent were students following along with the lesson") and estimated percentage of students who were engaged in the lesson (0–100%). The self-report of implementation was predicated on recommendations from Sanetti and Kratochwill (2011), who have demonstrated that it is possible to develop valid self-report measure of implementation. In addition, teachers were asked how many lessons they completed at the end of the year by school liaisons as an indicator of dosage. Of all indicators of implementation, dosage varied the most within schools. The average number of lessons completed across sites was 17.42 (SD = 3.72, range = 7–25). The school-level (unconditional) intraclass correlation (ICC) = .32 suggests that teachers completed lessons at a pace more similarly within schools than between schools. Data suggest that most teachers delivered the program with fidelity: on average, 85% of lesson components were reportedly delivered (SD = .17, ICC < .01), and most teachers made only a few adaptations, (Sum = 1.92, SD = 1.28, range = 0.00–6.55, ICC = .02). Further, most students were engaged (Sum = 7.78, SD = .87, range = 4.40–9.00, ICC = .12).

4.2.4. Supporting implementation

To support the integrity of implementation of the *Second Step*® curriculum and PCM strategies, monthly tips and reminders were developed and disseminated to teachers. Two tips and reminders were sent per month: one for the *Second Step*® curriculum and the other for PCM strategies. This process began at the beginning of December 2012 and continued until May 2013 and totaled 12 tips and reminders. The tips and reminders were distributed to early-start teachers in two ways: (a) email with attachment and (b) school liaisons putting printed copies in teachers' mailboxes.

4.2.5. Compensation

Participating schools were given a financial stipend for their involvement in the study, and school liaisons were given $250 a year for their support in communicating with teachers, distributing materials, coordinating data collection times, and tracking implementation. Liaisons served as the point person within each school to coordinate research activities and monitored implementation but did not directly implement the *Second Step*® curriculum. Teachers were compensated $5 per student per online survey, with a $25 bonus for completing the survey on all the students within a 3-week window of time. Teachers were also compensated $75 for completion of implementation logs. Early-start schools were provided the curricula at no-cost, and delayed schools were scheduled to receive the free curricula at the end of data collection.

4.3. Measures

The fall data were collected between October 10 and November 6, 2012. Reports indicated that roughly 93% of all teachers across sites completed the online surveys within the allotted time frame. Spring data collection occurred between April 22 and May 31, and reports indicate that 93% of teachers completed the online surveys.

4.3.1. School demographic and archival data

We collected school-level data from publicly available online sources (e.g., NCES website, school district websites) on the type of school (e.g., public vs. private), number of students, racial/ethnic composition of students, and percentage of students receiving free or reduced-price lunch. Administrative data, such as student mobility, disciplinary actions, suspensions, and absenteeism were also collected from participating schools.

4.3.2. Teacher assessment of student behavior

Teachers completed online surveys of student behavior via the Datstat Illume system (DatStat Inc., Seattle, WA). The first was the teacher version of the Devereux Student Strengths Assessment—*Second Step*® Edition (DESSA-SSE; Devereux Center for Resilient Children, 2012). The DESSA-SSE is a 36-item, standardized, norm-referenced behavior rating scale that assesses the social–emotional competencies that serve as protective factors for children in kindergarten through the eighth grade and map onto the *Second Step*® program: (a) skills for learning ($n = 9$, $\alpha = .95$), empathy ($n = 9$, $\alpha = .95$), emotion management ($n = 9$, $\alpha = .91$), problem solving ($n = 9$, $\alpha = .94$), and social–emotional composite ($n = 36$, $\alpha = .98$). The DESSA scale from which the DESSA-SSE was derived has been shown to have acceptable reliability and validity evidence (Nickerson & Fishman, 2009).

Teachers also completed the Strengths Difficulties Questionnaire (SDQ; Goodman, 1997). The SDQ is a brief behavior rating scale for 3–16 year olds that assesses functioning in five domains: peer problems ($n = 5$, $\alpha = .63$), hyperactivity ($n = 5$, $\alpha = .90$), conduct problems ($n = 5$, $\alpha = .77$), and prosocial ($n = 5$, $\alpha = .83$) and emotional symptoms ($n = 5$, $\alpha = .80$). Alpha coefficients are calculated based on fall data. Scores for the SDQ range from 0 (*not true*) to 2 (*certainly true*) on a 3-point Likert scale. The SDQ has demonstrated acceptable internal consistency and stability reliability and validity (Goodman, 2001) and has been shown to be at least as good as the Child Behavior Checklist (CBCL; Achenbach, 1991) at detecting conduct and emotional problems (Goodman & Scott, 1999).

4.3.3. Behavioral observation

To record class-wide and individual student behavior, a behavioral observation system was developed based on the Behavioral Observation of Students in Schools (BOSS; Shapiro & Kratochwill, 2000). The three behavioral coding categories consisted of on-task behavior, off-task behavior, and disruptive behavior (DB). The present study focused on one aspect of classroom behavior that is more closely tied to social–emotional competence, DB, defined as behaviors that were disruptive to learning or the classroom environment (e.g., call outs, talking to peers when not permitted, out of seat, behavior that draws peers off-task, playing with an object, etc.). There was some overlap between the behavioral codes, as the coding procedures made it is possible for a student to be coded as off-task/non-disruptive or off-task/disruptive.

Observations were conducted in all classrooms (early and delayed start) across both sites by trained graduate students during core academic instruction time in the fall, winter, and spring. Each student was observed for 2 minutes total, divided into 10-s intervals. To obtain class-wide estimates of DB, observers were instructed to begin with an identified student in the front or back of the classroom and systematically move to the next student to the left after each interval. After the observers made their way through all students in the class, they repeated the same process until the observation time elapsed. A minimum of 12 intervals of data per student and roughly 300 total intervals across all students were obtained. This observation system allowed for the calculation of class-wide and individual student estimates.

Prior to conducting the observations, graduate students were trained on the observation system. Before beginning baseline data collection, each student was required to reach at least 90% agreement during practice trials with an identified observer who served as the anchor measure. Inter-observer agreement (IOA) data consisting of two observers conducting the observation at the same time on the same students were collected on roughly 20% of the observation sessions. IOA was calculated using the point-by-point method, which consists of calculating agreement for each and every interval. This method has been shown to be a more accurate estimate of the agreement between raters for direct observation systems with interval recording formats (Shapiro & Kratochwill, 2000). The results revealed that IOA averaged 88% (minimum = 72% and maximum = 100%), which was associated with a kappa value of .71 and is considered to be an acceptable level of inter-rater reliability (Viera & Garrett, 2005).

4.3.4. Proactive classroom management

Data on proactive classroom management were collected using the Proactive Classroom Management Rating Form (PCM-RF; Cook, 2009). Trained graduate students completed a 20-item Proactive Classroom Management Survey based on observed classroom management and proactive behavior management strategies that were included in the PCM training ($\alpha = .94$). Items were rated based on a 4-point Likert scale rating (very untrue to very true), the degree to which PCM strategies were being implemented. Cook and Browning-Wright (2010) found that the PCM-RF was found to have acceptable inter-rater reliability ($r = .72$), as well as evidence in support of its criterion-related validity, with moderate correlations with measures of student academic engaged time and disruptive class behavior (Cook & Browning-Wright, 2010).

4.4. Statistical analysis

We assessed intervention effects on each of the primary outcomes with a mixed-model time × condition analysis (Murray, 1998) to account for the intraclass correlation associated with students nested within schools, the unit of assignment. The analysis tests net differences between conditions on change in outcomes from the fall (T1) to spring (T2) with gains for individual students clustered within schools. The test of net differences provides an unbiased and straightforward interpretation of the results (Cribbie & Jamieson, 2000; Fitzmaurice, Laird, & Ware, 2004). The basic statistical model includes time, condition, and the time × condition interaction, with time coded 0 at T1 and 1 at T2, and condition coded 0 for control and 1 for intervention. With 61 schools, tests of time × condition used 59 degrees of freedom (*df*).

The basic model was expanded to include covariates and to test for a differential response due to student-level and classroom-level variables (moderation). To test moderation, we expanded the model to test interactions. The statistical model included a predictor and its interaction with condition, time, and the time × condition term, resulting in a three-way interaction, all corresponding two-way interactions, and individual (conditional) effects. The three-way interaction of the predictor, time, and condition provides an estimate of whether the condition effect varied by the predictor. The analysis included dichotomous and continuous predictors, and we used continuous variables whenever possible.

4.4.1. Model estimation

We fit models to our data with SAS PROC MIXED version 9.2 (SAS Institute, 2009) using restricted maximum likelihood and included all available data, whether or not students' scores were present at both time points. Maximum likelihood estimation with all available data produces potentially unbiased results even in the face of substantial attrition, provided the missing data were missing at random (Schafer & Graham, 2002). In the present study, we did not believe that attrition or other missing data represented a meaningful departure from the missing at random assumption, meaning that missing data likely did not depend on unobserved determinants of the outcomes of interest (Little & Rubin, 2002). Most missing data involved students who were absent on the day of assessment or transferred to a new school.

The models assume independent and normally distributed observations. We addressed the first assumption (van Belle, 2008) by explicitly modeling the multilevel nature of the data. Regression methods have been found quite robust to violations of normality and outliers have a limited influence on the results in a variety of multilevel modeling scenarios (Bloom, Bos, & Lee, 1999; Donner & Klar, 1996; Fitzmaurice et al., 2004; Hannan & Murray, 1996; Murray et al., 2006). Murray et al. (2006) showed that violations of normality at either or both the individual and group levels do not bias results as long as the study is balanced at the group level.

4.4.2. Effect sizes

To ease interpretation, we computed an effect size, Hedges' g (Hedges, 1981), for each fixed effect according to the What Works Clearinghouse (WWC, 2014) standards. Hedges' *g* is comparable to Cohen's *d* (Cohen, 1988). Both represent individual-level effect sizes, but we suggest caution during interpretation, as this study is designed for inferences about schools, which do not necessarily apply to individuals (ecological fallacy).

5. Results

Table 1 presents descriptive statistics for each measure. Conduct problems represented the largest departure from the normal distribution among the ten survey scales, with skewness index value of 2.09 and kurtosis index value of 4.59. All other scales had skewness and skewness index values between within −1 and 2 and kurtosis index values between −1 and 3.5. The observations

Table 1
Descriptive statistics for dependent variables at pretest (T1) and posttest (T2).

		Early start			Delayed start			Percentiles							Percent missing
		Mean	*SD*	*N*	Mean	*SD*	*N*	Min	5th	25th	50th	75th	95th	Max	
SDQ conduct problems	T1	1.07	1.68	3359	1.05	1.74	3147	0	0	0	0	1	5	10	12
	T2	1.07	1.72	2895	1.10	1.81	2764	0	0	0	0	2	5	10	24
SDQ emotional problems	T1	1.09	1.61	3349	1.03	1.63	3147	0	0	0	0	2	5	9	12
	T2	1.10	1.62	2894	1.19	1.66	2765	0	0	0	0	2	5	10	24
SDQ hyperactivity	T1	3.53	3.10	3368	3.36	3.12	3147	0	0	1	3	5	10	10	12
	T2	3.11	3.00	2913	3.24	3.13	2764	0	0	0	2	5	9	10	23
SDQ peer problems	T1	1.46	1.66	3340	1.41	1.68	3145	0	1	2	2	4	6	10	13
	T2	1.18	1.57	2885	1.25	1.64	2762	0	1	2	2	3	5	9	24
SDQ Prosocial	T1	7.15	2.48	3328	7.01	2.51	3120	0	3	5	7	9	10	10	13
	T2	7.74	2.34	2879	7.66	2.37	2762	0	3	6	8	10	10	10	24
DESSA emotional management	T1	24.12	6.35	3323	23.47	6.54	3086	0	13	20	24	27	35	36	14
	T2	26.60	6.34	2871	25.25	6.66	2741	0	15	22	26	31	36	36	24
DESSA empathy	T1	23.60	7.00	3341	23.15	7.30	3141	0	11	19	24	27	36	36	13
	T2	26.54	6.76	2893	25.31	6.99	2757	0	14	21	27	31	36	36	24
DESSA problem solving	T1	23.77	6.72	3341	23.32	6.93	3135	0	12	19	24	28	35	36	13
	T2	26.30	6.60	2881	25.18	6.91	2758	0	14	21	26	31	36	36	24
DESSA skills learning	T1	24.58	7.16	3359	24.49	7.33	3144	0	12	20	25	29	36	36	12
	T2	26.95	6.89	2914	26.10	7.31	2756	0	13	22	27	33	36	36	24
DESSA social–emotional	T1	96.10	25.57	3320	94.24	26.14	3064	0	51	78	97	111	139	144	14
	T2	106.48	25.10	2864	101.80	26.07	2737	0	60	88	106	124	144	144	25
Observations of disruptive behavior	T1	9.53	15.50	3331	8.81	14.37	3270	0	0	0	0	17	42	100	11
	T2	8.62	14.65	3080	9.60	16.50	2999	0	0	0	0	17	42	100	18

Note. SDQ = Strengths and Difficulties Questionnaire and DESSA = Devereux Student Strengths Assessment.

of disruptive behavior had a skewness index value of 2.15 and kurtosis index value of 5.16. While these values were high for conduct problems, emotional problems, and observations of disruptive behavior, they would not be considered extreme (Kline, 2005). The following sections address attrition, baseline equivalence, tests of efficacy, differential response to the *Second Step*® intervention, and the associated between engagement or dosage and student outcomes.

5.1. Attrition

Student attrition was defined as students with data at T1 but missing data at T2, and we examined attrition with respect to the sample of 7244 students, 3594 in comparison schools and 3650 in intervention schools. We experienced 11.0% attrition at T2, with 419 students missing T2 data in comparison schools and 381 students missing T2 data in intervention schools, and attrition rates did not differ between conditions ($\chi^2 = 2.74$, $df = 1$, $p = .098$). Although differential rates of attrition are undesirable, differential scores by condition present a greater threat to validity (Barry, 2005). We conducted an analysis to test whether student scores were differentially affected by attrition across conditions. We examined the effects of condition, attrition status, and the interaction between the two on pretest scores within a mixed-model analysis of variance (Murray, 1998), which nests students' T1 scores within schools and condition. We tested scores for all DESSA and SDQ scales and observations of disruptive behavior. We found no evidence of differential attrition effects for any of our dependent variables: $p > .149$ for all tests. We also found no evidence of differential attrition effects for students' gender, ethnicity (e.g., white, Hispanic), or special education status or teachers' proactive classroom management.

5.2. Baseline equivalence

We tested the difference between conditions at baseline within the models that test efficacy. The condition effect (not crossed by time) shows the difference between conditions at pretest (see Table 2). This difference was not statistically significant for any of the social–emotional competence measures.

5.3. Efficacy

Consistent with our hypothesis that *Second Step*® would promote social–emotional competence and reduce disruptive behaviors, we tested whether students in students in intervention schools would perform better on several social–emotional measures than students in comparison schools. Table 2 presents the results of all tests of main-effect tests of treatment efficacy. The time × condition row represents the critical test of condition on gains for each measure from fall to spring, and the bottom two rows in the table shows the effect sizes and *p*-values for that critical test. We initially included three pretest covariates: grade level, percent of students who receive free and reduced-price lunch in each school, and proactive classroom management. While the covariates predicted the student outcomes, they did not influence the time × condition estimates, so we reported the simpler models without covariates in Table 2. We also report the ICCs for gains in each measure as described by Murray (1998, see p. 301).

Table 2
Results from mixed-model time × condition analysis of condition effects on fall-to-spring gains in social–emotional behavior.~, *, **, ***

Effect or statistic		SDQ emotional problems	SDQ conduct problems	SDQ hyperactivity	SDQ peer problems	SDQ prosocial	DESSA social–emotional	DESSA skills learning	DESSA empathy	DESSA emotional management	DESSA problem solving	Observations of disruptive behavior
Fixed effects	Intercept	1.03***	1.06***	3.37***	1.42***	7.00***	94.31***	24.48***	23.11***	23.50***	23.25***	8.88***
		(.07)	(.06)	(.09)	(.07)	(.12)	(1.34)	(.30)	(.41)	(.35)	(.35)	(.79)
	Time	.17***	.09 ~	−.05	−.15*	.60***	6.71***	1.41***	2.04***	1.58***	1.69***	.76
		(.05)	(.05)	(.07)	(.06)	(.11)	(1.18)	(.24)	(.38)	(.32)	(.28)	(.97)
	Condition	.07	.01	.17	.04	.13	1.54	.05	.45	.56	.44	.71
		(.09)	(.09)	(.13)	(.10)	(.17)	(1.88)	(.42)	(.57)	(.49)	(.49)	(1.11)
	Time × condition	−.17*	−.07	−.33**	−.11	−.04	3.19 ~	.81*	.81	.82 ~	.72 ~	−1.68
		(.07)	(.07)	(.10)	(.09)	(.15)	(1.66)	(.34)	(.54)	(.45)	(.39)	(1.36)
Variances	School intercept	.09***	.06**	.12**	.09***	.22***	29.07***	1.54***	2.50***	1.89***	2.24***	4.37 ~
		(.02)	(.02)	(.04)	(.02)	(.06)	(8.25)	(.43)	(.74)	(.56)	(.58)	(2.52)
	School gains	.02**	.02***	.05***	.04***	.15***	18.74***	.73***	2.03***	1.42***	1.01***	12.25***
		(.01)	(.01)	(.01)	(.01)	(.03)	(3.89)	(.17)	(.41)	(.29)	(.22)	(2.61)
	Student	1.29***	1.98***	7.13***	1.37***	3.47***	443.31***	35.98***	29.05***	26.39***	31.01***	21.55***
		(.04)	(.05)	(.15)	(.04)	(.09)	(9.81)	(.78)	(.69)	(.61)	(.68)	(2.99)
	Residual	1.27***	.99***	2.30***	1.22***	2.14***	179.14***	13.91***	16.22***	12.60***	12.54***	195.46***
		(.02)	(.02)	(.04)	(.02)	(.04)	(3.49)	(.27)	(.31)	(.24)	(.24)	(3.67)
ICC		.015	.024	.021	.035	.064	.095	.050	.111	.101	.074	.059
Hedges' *g*	Time × condition	−.104	−.040	−.109	−.067	−.016	.125	.114	.118	.126	.107	−.108
p-value	Time × condition	.0120	.3023	.0014	.2229	.8000	.0587	.0221	.1365	.0755	.0689	.2238

Note. Table entries show parameter estimates with standard errors in parentheses except for intraclass correlations (ICCs), Hedges' *g* values, and *p*-values. Tests of fixed effects (first four rows) used 59 *df* to account for the school as the unit of analysis. DESSA = Devereux Student Strengths Assessment; SDQ = Strengths and Difficulties Questionnaire; ICC = intraclass correlation coefficient. ICCs calculated as per Murray (1998, p. 301).

~ $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Students in schools that implemented *Second Step*® showed greater improvements in DESSA-SSE skills learning ($p = .022$; $g = .11$), as compared to students in control schools, and greater reductions in SDQ emotional problems ($p = .012$; $g = -.10$) and SDQ hyperactivity ($p = .001$; $g = -.11$) over a 1-year period. The effect sizes for the DESSA social–emotional composite and emotional management and problem-solving sub-scales approached or exceeded these in magnitude, with $g = .13$, .13, and .11, respectively, but the condition effects were only marginally significant ($p = .059$, $p = .079$, and $p = .069$). When we applied the Benjamini–Hochberg procedure to control false discovery rate (Benjamini & Hochberg, 1995), only SDQ hyperactivity was statistically significant, with an adjusted *p*-value of .015. SDQ emotional problems and DESSA-SSE skills learning had adjusted *p*-values of .066 and .081, respectively.

5.4. Differential effects

For many students, improvement in social–emotional measures was not likely because students had scored in an acceptable range at pretest. Because we expected little gains from well-adjusted students, tests of efficacy across all students can suppress information about the treatment effects. We therefore tested whether students' response to the *Second Step*® intervention depended on students' pretest scores with a moderation analysis. Table 3 presents the results of the moderation analyses. The critical test of moderation in this table is the time × condition × pretest row.

The time × condition effect was moderated by pretest scores for SDQ conduct problems ($p = .004$), hyperactivity ($p < .001$), peer problems ($p < .001$), and prosocial behaviors ($p = .003$). Moderation effects were also found for the DESSA social-emotional skills ($p < .001$), skills for learning ($p < .001$), emotional management ($p < .001$), and problem solving ($p < .001$) scales. Pretest marginally (but not significantly) moderated the DESSA empathy scale ($p = .051$) and the observations of disruptive behavior ($p = .061$).

To help interpret the results, Fig. 3 provides graphs of condition effects for four scales from the SDQ across the range of pretest scores. Fig. 4 depicts condition effects by pretest scores for four DESSA scales: social–emotional, skills learning, emotion management, and problem solving. Fig. 3A, for example, shows the moderation results for conduct problems. Students with fewer conduct problems (left side of Fig. 3A) did not differ between conditions, which is apparent from the mean difference (center, heavier line) of about zero and 95% confidence bounds (lighter outer lines) that include zero. Moving from left to right, the 95% confidence bounds begin to exclude zero at a pretest score of about 2.3, implying a statistically significant difference between conditions for students who scored 2.3 or higher on conduct problems.

It is important to interpret the moderation effects, however, in the context of the distribution of scores, summarized in Table 1. At least 50% of students scored a 0 and 75% scored a 1 or less on the conduct problems scale. Interpolating from Table 1, we estimate that the differences between conditions are statistically significant for about the highest-scoring 15–20% of the sample.

Table 3
Results from mixed-model time × condition × pretest moderation analysis on fall-to-spring gains in social–emotional behavior.~, *, **, ***, ****

Effect or statistic		SDQ emotional problems	SDQ conduct problems	SDQ hyperactivity	SDQ peer problems	SDQ prosocial	DESSA social–emotional	DESSA skills learning	DESSA empathy	DESSA emotional management	DESSA problem solving	Observations of disruptive behavior
Fixed effects	Intercept	1.06***	1.06***	3.45***	1.44***	7.08***	95.21***	24.54***	23.38***	23.81***	23.55***	9.17***
		(.04)	(.03)	(.05)	(.04)	(.07)	(.79)	(.17)	(.24)	(.22)	(.19)	(.48)
	Time	.16**	.09*	−.06	−.15*	.57***	6.28***	1.39***	1.92***	1.43***	1.61***	.26
		(.05)	(.05)	(.07)	(.06)	(.10)	(1.13)	(.24)	(.34)	(.31)	(.27)	(.68)
	Condition	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		(.05)	(.05)	(.07)	(.06)	(.10)	(1.11)	(.23)	(.34)	(.31)	(.27)	(.67)
	Time × condition	−.14 ~	−.06	−.30**	−.09	.03	3.98*	.86*	1.05*	1.10*	.90*	−.79
		(.08)	(.06)	(.10)	(.08)	(.14)	(1.58)	(.33)	(.48)	(.44)	(.38)	(.96)
	Pretest	1.00***	1.00***	1.00***	1.00***	1.00***	1.00***	1.00***	1.00***	1.00***	1.00***	1.00***
		(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
	Condition × pretest	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)
	Time × pretest	−.48***	−.27***	−.21***	−.43***	−.39***	−.25***	−.25***	−.37***	−.26***	−.26***	−.92***
		(.02)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)
	Time × condition × pretest	−.02	−.07***	−.09***	−.13***	−.06**	−.09***	−.10***	−.04 ~	−.13***	−.09***	.05 ~
		(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.03)
Variances	School intercept	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		(.01)	(.00)	(.01)	(.01)	(.02)	(2.47)	(.11)	(.23)	(.19)	(.14)	(.90)
	School gains	.03***	.02***	.06***	.05***	.12***	17.48***	.72***	1.64***	1.36***	.98***	5.95***
		(.01)	(.01)	(.01)	(.01)	(.03)	(3.55)	(.16)	(.33)	(.27)	(.20)	(1.29)
	Student	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		(.01)	(.01)	(.02)	(.01)	(.02)	(1.72)	(.13)	(.14)	(.12)	(.12)	(1.28)
	Residual	.88***	.78***	1.80***	.81***	1.47***	136.55***	10.60***	11.58***	9.54***	9.60***	103.22***
		(.02)	(.01)	(.03)	(.01)	(.03)	(2.48)	(.19)	(.21)	(.17)	(.17)	(1.84)

Note. Table entries show parameter estimates with standard errors in parentheses except for intraclass correlations (ICCs), Hedges' *g* values, and *p*-values. Tests of fixed effects (first four rows) used 57 *df* to account for the school as the unit of analysis. DESSA = Devereux Student Strengths Assessment; SDQ = Strengths and Difficulties Questionnaire; ICC = intraclass correlation coefficient. ICC calculated as per Murray (1998, p. 301).

~ $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.
**** $p < .0001$.

The confidence bounds widen toward the right side of the chart because fewer students contributed to the condition effect estimate at higher pretest scores. Most other scales were less skewed and produced differences between conditions for a larger portion of the sample. For social–emotional problems, students with scores below about 100, approximately 60% of the sample, benefitted from *Second Step*®. Similarly, nearly 60% of the students, those with lower scores, improved their problem-solving skills when compared to students in control schools. In every test of moderation, the students with more problems or fewer social skills appeared to benefit from the *Second Step*® intervention, while we found no differences between conditions for those students who scored more positively at pretest.

Proactive classroom management conditioned teacher reports of students' peer problems ($t = -3.45$, $df = 57$, $p = .001$) and empathy ($t = 2.02$, $df = 57$, $p = .049$). Teachers with about the highest 60% of proactive classroom management scores reported fewer peer problems in *Second Step*® schools compared to controls. For teachers with roughly the highest 75% of proactive classroom management scores, students exposed to *Second Step*® improved in empathy over controls. We found nonsignificant moderation effects by proactive classroom management for SDQ emotional symptoms ($t = -1.01$, $df = 57$, $p = .316$), conduct problems ($t = 0.40$, $df = 57$, $p = .692$), hyperactivity ($t = -0.52$, $df = 57$, $p = .607$), and prosocial behavior ($t = 0.10$, $df = 57$, $p = .923$) and DESSA social–emotional ($t = 1.39$, $df = 57$, $p = .170$), skills learning ($t = 1.21$, $df = 57$, $p = .230$), emotional management ($t = 0.84$, $df = 57$, $p = .404$), and problem solving ($t = 0.32$, $df = 57$, $p = .753$), and observations of disruptive behavior ($t = -0.86$, $df = 57$, $p = .392$).

We found one statistically significant moderation effect for grade level ($t = 2.18$, $df = 58$, $p = .034$), with students in grade 2 benefitting more on the emotion management scale of the DESSA than students in kindergarten or grade 1. We found nonsignificant moderation effects by grade level for SDQ emotional symptoms ($t = -1.22$, $df = 57$, $p = .227$), conduct problems ($t = -0.57$, $df = 57$, $p = .571$), hyperactivity ($t = 0.06$, $df = 57$, $p = .953$), peer problems ($t = 0.61$, $df = 57$, $p = .545$), and prosocial behavior ($t = -0.73$, $df = 57$, $p = .467$) and DESSA social–emotional ($t = 1.59$, $df = 57$, $p = .18$), skills learning ($t = 0.63$, $df = 57$, $p = .534$), empathy ($t = 1.91$, $df = 57$, $p = .061$), and problem solving ($t = 0.52$, $df = 57$, $p = .604$), and observations of disruptive behavior ($t = -0.01$, $df = 57$, $p = .992$).

We tested for moderation by state and found nonsignificant moderation effects for SDQ emotional symptoms ($t = -1.08$, $df = 57$, $p = .286$), conduct problems ($t = -0.98$, $df = 57$, $p = .331$), hyperactivity ($t = -1.87$, $df = 57$, $p = .067$), peer problems ($t = -1.51$, $df = 57$, $p = .137$), and prosocial behavior ($t = 1.04$, $df = 57$, $p = .301$) and DESSA social–emotional ($t = 1.52$, $df = 57$, $p = .133$), skills learning ($t = 1.14$, $df = 57$, $p = .259$), empathy ($t = 1.38$, $df = 57$, $p = .174$), emotional

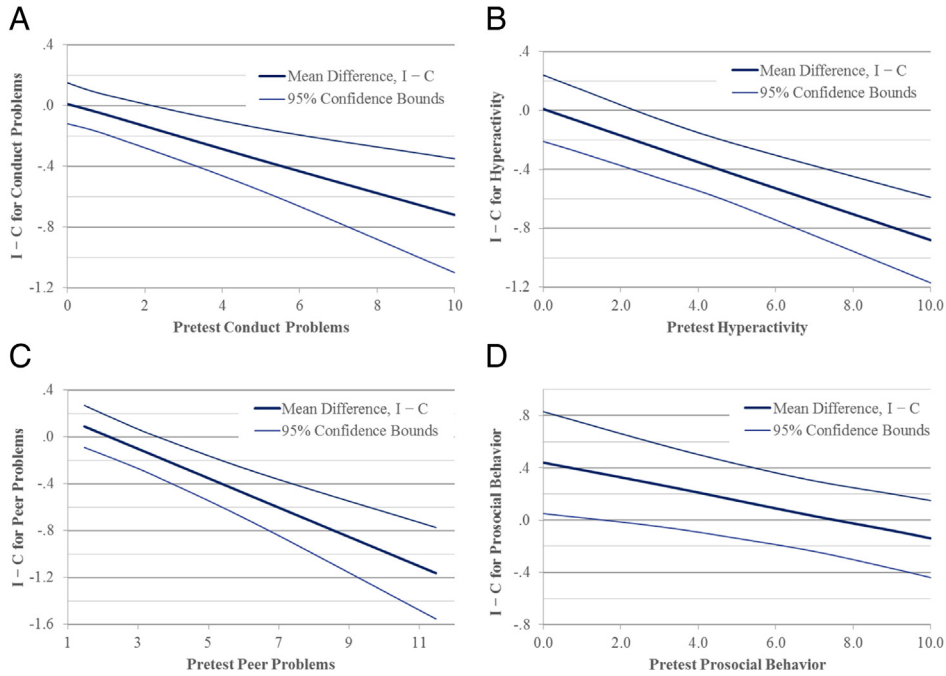

Fig. 3. Differences between conditions for four scales from the Strengths and Difficulties Questionnaire plotted by pretest scores. The vertical axis shows the difference between conditions (I–C) on gains in each scale across the range of the pretest scores; a gain of zero represents no difference between conditions. The heavy line depicts the mean difference estimate. The two thin, outer lines show the 95% confidence interval around the mean estimate. The confidence intervals exclude zero, indicating a statistically significant difference, at values above 2.3 (81st percentile) for pretest conduct problems, 2.4 (43rd percentile) for pretest hyperactivity, and 3.6 (76th percentile) for pretest peer problems. For prosocial behavior, conditions differ significantly for pretest values of 1.5 (3rd percentile) or less.

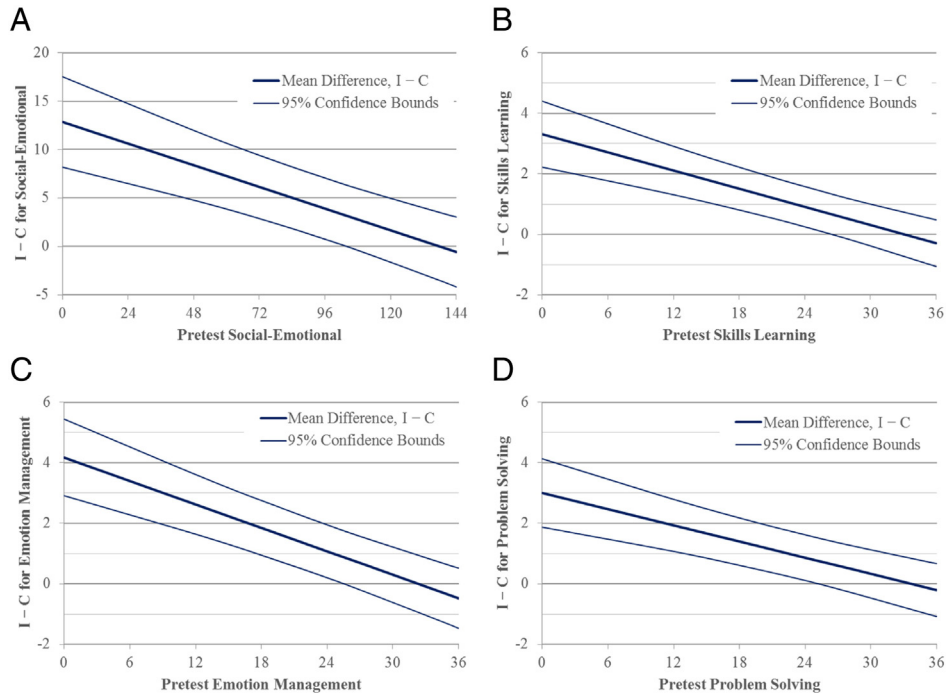

Fig. 4. Differences between conditions for four scales from the Devereux Student Strengths Assessment plotted by pretest scores. The vertical axis shows the difference between conditions (I–C) on gains in each scale across the range of the pretest scores; a gain of zero represents no difference between conditions. The heavy line depicts the mean difference estimate. The two thin, outer lines show the 95% confidence interval around the mean estimate. The confidence intervals exclude zero, indicating a statistically significant difference between conditions, for values below 103 (62nd percentile) for the pretest social–emotional scale, 26 (59th percentile) for pretest skills learning, 26 (62nd percentile) for pretest emotion management, and 25 (57th percentile) for pretest problem solving.

management ($t = 1.62$, $df = 57$, $p = .110$), and problem solving ($t = 1.50$, $df = 57$, $p = .139$), and observations of disruptive behavior ($t = 1.40$, $df = 57$, $p = .166$).

Finally, we tested moderation by the proportion of white students and found nonsignificant moderation effects for SDQ emotional symptoms ($t = 1.17$, $df = 57$, $p = .248$), conduct problems ($t = 1.18$, $df = 57$, $p = .243$), hyperactivity ($t = 1.42$, $df = 57$, $p = .163$), peer problems ($t = 0.01$, $df = 57$, $p = .993$), and prosocial behavior ($t = -0.12$, $df = 57$, $p = .908$) and DESSA social–emotional ($t = 0.60$, $df = 57$, $p = .554$), skills learning ($t = 1.10$, $df = 57$, $p = .278$), empathy ($t = 0.30$, $df = 57$, $p = .763$), emotional management ($t = 0.65$, $df = 57$, $p = .518$), and problem solving ($t = 0.55$, $df = 57$, $p = .584$), and observations of disruptive behavior ($t = 0.88$, $df = 57$, $p = .384$).

5.5. Engagement and dosage

We next predicted each of the student posttest measures, controlling for their pretest values, with two measures of implementation: the number of lessons completed and engagement in *Second Step*®, given that adherence had little variability. The number of completed lessons and engagement were available only within the intervention schools, which prohibited a formal mediation analysis. The statistically significant associations between these two implementation measures, however, offer some evidence that *Second Step*® activities likely produced the differences observed between conditions. We fit these data to multilevel models and report standardized regression estimates (β).

The number of lessons completed and engagement, respectively, predicted SDQ emotional problems ($\beta = -.02$, $p = .265$; $\beta = -.13$, $p < .000$), conduct problems ($\beta = -.06$, $p = .0025$; $\beta = -.12$, $p < .000$), hyperactivity ($\beta = -.03$, $p = .089$; $\beta = -.03$, $p = .046$), peer problems ($\beta = -.02$, $p = .332$; $\beta = -.15$, $p < .000$), prosocial ($\beta = .06$, $p = .005$; $\beta = .05$, $p = .003$). Lessons completed and engagement predicted DESSA social–emotional ($\beta = .06$, $p = .001$; $\beta = .12$, $p < .000$), skills learning ($\beta = .05$, $p = .002$; $\beta = .11$, $p < .000$), empathy ($\beta = .06$, $p = .000$; $\beta = .12$, $p < .000$), emotional management ($\beta = 06$, $p = .001$; $\beta = .13$, $p < .000$), problem solving ($\beta = .06$, $p = .000$; $\beta = .10$, $p < .000$). Finally, lessons completed and engagement predicted the observations of disruptive behavior ($\beta = -.09$, $p = .000$; $\beta = -.08$, $p < .000$). In general, engagement was a more important predictor (mean absolute $\beta = .10$) than the number of lessons completed (mean absolute $\beta = .05$).

6. Discussion

Too many children start off their school careers without the necessary social–emotional skills to be optimally engaged in the classroom, which can dampen the cumulative profit from their learning experiences. Several organizations have highlighted and validated the central role of social and executive function skills in learning outcomes, spawning a number of SEL programs, which are now in widespread use. For both policy and practice, it remains important to identify promising programs through rigorous evaluations and expand our understanding of who is benefiting from such programming. *Second Step*® is one of the most widely used SEL programs in the country, warranting further evaluation of its efficacy. The present study aimed to (1) examine main effects of the revised *Second Step*® program on a range of social–behavioral indices, (2) examine moderating effects of baseline skill levels (i.e., differential response) and classroom management, and (3) examine the influence of implementation on SEL outcomes among an early elementary population.

Overall, the data support the internal validity of the program but suggest that the benefits of *Second Step*® are most pronounced for children with lower baseline competencies. Indeed, the hypothesis that all children would benefit similarly (i.e., main effect) from *Second Step*® did not receive strong support—with statistically significant condition differences found for only two out of the eleven outcomes tested. Rather, tests of moderation indicate that for the most part, *Second Step*® produces larger differences between conditions among students with initially higher levels of problem behavior versus lower levels of problem behaviors. Specifically, positive effects were found for conduct problems, hyperactivity, peer problems, prosocial skills, SEL skills, skills for learning, emotion management and problem solving—as reported by teachers. However, these effects were specific to children who were generally in the lower half of their peers (50th percentile). For those dependent variables where pretest moderated the condition effect, between 15% and 60% of students benefitted from *Second Step*®. Thus, although the program is universal in delivery and format (i.e., non-targeted), teachers are primarily noticing improvements in youth with skills and behaviors that are mostly below the average peer. These patterns are not surprising, given that youth have variable levels of skills and will be expected to respond differently to intervention (Walker et al., 1996). Notably, these effects were not moderated by ethnicity or grade level.

No effects were found with the behavioral observations of disruptive behavior. Change in observed behaviors is often difficult to demonstrate due to the limited time observers are able to remain in each classroom, in contrast to teachers who have daily observations and interactions with students over many months. However, it is encouraging that the baseline moderation with observed disruptive behavior was marginally significant.

As a covariate, baseline classroom management skills were related to all study outcomes but did not account for the relation between the intervention status and outcomes, and thus was removed (from further analyses) for parsimony. However, the results suggested that *Second Step*® improved empathy and peer problems in classrooms with more proactive, positive classroom management. This is important to highlight because while *Second Step*® can yield benefits to students regardless of classroom management skill, the magnitude of benefit on certain key competencies can be greater when in the context of proactive, positive classroom climate and supports.

Although tests of mediation by implementation measures were not possible, the correlations between student SEL outcomes and implementation indices support the conclusion that *Second Step*® produced the differences between treatment conditions. The associations between implementation indices and outcomes within *Second Step*® schools were also similar in magnitude to the differences between conditions.

This study contributes to the prevention science and SEL literature in a few important ways. First, this study was designed and conducted with several features of an effectiveness trial, but overall maintained little researcher control. Although the literature is replete with efficacy trials, few studies have focused on evaluating the effectiveness of SEL programs under real-world educational conditions, so as to enhance their translational value (Merrell & Gueldner, 2010). Second, this study included a large, heterogeneous sample of geographically, ethnically, and socio-economically diverse students. Last, we utilized both survey and observational data, and rigorous data analytic methods.

Moderation analyses from this study warrant more elaboration and offer insights into the individuals who benefit most from universal programs such as *Second Step*®. Findings from this study reflect differential effectiveness, which has received little attention in the context of universal SEL supports. In the present case, children who were most responsive to the program were those who struggle to consistently deploy positive social–emotional skills and positive behaviors (under demands of a classroom environment). That is, children who began with a higher than average number of problematic behaviors or skill deficits showed greater improvement, driving program effects. Although it is important to be attentive to children who are at risk for behavior problems, it is important to note that children with adequate levels of these skills were able to maintain those skills. It would be misleading, and a consequence of the ecological fallacy, to conclude that only certain children benefited from *Second Step*®. Indeed, if one considers the benefits of *Second Step*® at the classroom level, versus individual level, socially sophisticated children clearly benefit indirectly by reaching the students who might impede classroom instruction through disruptive behaviors. It may also be that students who began with the most challenges benefited from the positive examples of their peers. Thus, we caution against making inferences about individual students from a universal intervention delivered at the classroom level.

Universal programming has several benefits. *Second Step*® can reach a large number of youth and can be taught in a regular classroom. Universal programs do not demand additional resources to screen targeted youth and do not have the stigma associated with identifying at-risk kids (see Horowitz, Garber, Ciesla, Young, & Mufson, 2007); and yet, as evidenced by the current data, universal programs can still reach at-risk children at an opportune developmental period. It also serves as a reminder that universal programming does not mean all children will show similar rates of improvement. Some students who receive universal programming have minimal room for improvement in social–emotional competence and adaptive behaviors (e.g., reductions in problem behavior), but they may need such supports in order to maintain their level of functioning over time.

6.1. Limitations

This study, like most, has limitations that readers should be aware of and that pinpoint directions for future work. First, we focused on child status when conducting the moderation analyses to avoid capitalization on chance. Second, this study focused on year one findings, limiting our ability to formally examine and interpret mediation, and mechanisms of change. Third, the conclusions reached in this study cannot necessarily extend beyond early elementary populations; further longitudinal work is needed to determine sustainability of these findings over time. Fourth, given the inclusion of the PCM training with the *Second Step*®, it is difficult to disentangle what was responsible for the effects. That being said, data suggest that PCM may serve as moderator rather than a direct intervention that promotes social–emotional competence. Last, the only outcome measures that were significant were teacher ratings, limiting the robustness of the findings. However, this is not altogether surprising, given that observational data provide a less reliable sampling of behavior. Future research and replication are nonetheless recommended in order to validate the observed benefits of *Second Step*® (and related programs).

6.2. Conclusions

Despite these limitations, findings translate into several implications for school psychologists or personnel in related positions. First, it is important for school psychologists to understand that *Second Step*® can be implemented with overall integrity in general education classrooms, with limited training and research involvement, and that better implementation corresponds with program impacts. Specifically, psychologists would be advised by these findings to ensure that teachers focus on both engagement, as well as breadth of coverage across lesson areas. Further, given the significant variability in implementation across teachers, school psychologists and other providers could consult with teachers by providing performance-based feedback to facilitate better implementation of SEL programming. Second, many models of delivery rely on school psychologists, counselors, or school social workers to implement the SEL curriculum. This study, however, demonstrates that most teachers are also capable of teaching the program with adequate levels of implementation quality. Indeed, for teachers with strong classroom management skills, impacts of Second Step implementation may be more pronounced.

The data suggest that the expectation of main effects may be overreaching and masking important subgroup differences. Presumably, a certain subset of children (in all schools) will not demonstrate added benefit of social skills programming because they already possess sufficiently strong skills. Conversely, it is important to highlight that *Second Step*® did not result in detrimental effects on children's social–emotional functioning. Further, universally delivered programs such as *Second Step*® may be one way to provide some of the support needed by higher-risk children without the required additional resources to screen children and the stigma associated with identifying students as at risk (see Horowitz et al., 2007). However, while *Second Step*® appears to have

more potency on for children most likely to get referred for social–behavioral concerns, it is unlikely that *Second Step*® alone is sufficient enough to address the needs of children with severe and complex social–emotional or cognitive difficulties. In such cases, there is evidence that those students would also benefit from more intensive services above and beyond *Second Step*® (Cook, Burns, Browning-Wright, & Gresham, 2010).

Acknowledgements

We would like to generously thank the teachers, staff, and administrators of participating districts in Puget Sound Area, WA, and Mesa, AZ. In addition, we would like to acknowledge research coordinators, Jodie Buntain-Ricklefs at the University of Washington and Alexa Ogburn at Arizona State University. Committee for Children funded the study and paid for Dr. Smolkowski's time as an analyst on the project, but all study-related activities, including data collection and analyses, were done without the participation or influence of staff so as to maintain complete objectivity.

References

Achenbach, T. M. (1991). *Manual for the child behavior checklist/4–18 and 1991 profile*. Burlington, VT: Department of Psychiatry, University of Vermont, 288.
Barry, A. E. (2005). How attrition impacts the internal and external validity of longitudinal research. *Journal of School Health*, *75*, 267–270. http://dx.doi.org/10.1111/j.1746-1561.2005.tb06687.x.
van Belle, G. (2008). *Statistical rules of thumb* (2nd ed.). New York: John Wiley & Sons.
Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological Bulletin*, *121*(1), 65.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.
Bernard, M. E. (2006). It's time we teach social-emotional competence as well as we teach academic competence. *Reading & Writing Quarterly*, *22*, 103–119.
Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist*, *57*(2), 111.
Bloom, H. S., Bos, J. M., & Lee, S. -W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of educational programs. *Evaluation Review*, *23*, 445–469. http://dx.doi.org/10.1177/0193841X9902300405.
Brooks-Gunn, J., Rouse, C. E., & McLanahan, S. (2007). *School readiness and the transition to kindergarten in the era of accountability*.
Cambourne, B. (2002). Holistic, integrated approaches to reading and language arts instruction: The constructivist framework of an instructional theory. *What research has to say about reading instruction*, *3*, 25–47.
Campbell, F. A., & Ramey, C. T. (1995). Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal*, *32*(4), 743–772.
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. http://dx.doi.org/10.1111/1467-9280.00260.
Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*, 489. http://dx.doi.org/10.1037/0022-0663.98.3.489.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children*, *75*, 365–383.
Cook, C. R. (2009). *The Proactive Classroom Management-Rating Form*. Baton Rouge, LA: Louisiana State University.
Cook, C. R., & Browning-Wright, D. (2010). Initial validation of the Proactive Classroom Management Rating Form. *Presentation delivered at the annual National Association of School Psychologists Convention in Chicago, IL*.
Cook, C. R., Burns, M., Browning-Wright, D., & Gresham, F. M. (2010). *Transforming school psychology in the RTI era: A guide for administrators and school psychologists*. Palm Beach: LRP Publications.
Cribbie, R. A., & Jamieson, J. (2000). Structural equation models and the regression bias for measuring correlates of change. *Educational and Psychological Measurement*, *60*, 893–907.
Demetriou, A., Christou, C., Spanoudis, G., & Platsidou, M. (2002). I. Introduction. *Monographs of the Society for Research in Child Development*, *67*(1), 1–38.
Denham, S. A. (2006). Social-emotional competence as support for school readiness: What is it and how do we assess it? *Early Education and Development*, *17*, 57–89.
Denham, S. A., Bassett, H. H., & Zinsser, K. (2012). Early childhood teachers as socializers of young children's emotional competence. *Early Childhood Education Journal*, *40*, 137–143.
Donner, A., & Klar, N. (1996). Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, *49*, 435–439. http://dx.doi.org/10.1016/0895-4356(95)00511-0.
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*, 405–432. http://dx.doi.org/10.1111/j.1467-8624.2010.01564.x.
Elias, M. J., & Haynes, N. M. (2008). Social competence, social support, and academic achievement in minority, low-income, urban elementary school children. *School Psychology Quarterly*, *23*, 474. http://dx.doi.org/10.1037/1045-3830.23.4.474.
Fantuzzo, J. F., Bulotsky-Shearer, R., McDermott, P. A., McWayne, C., Frye, D., & Perlman, S. (2007). Investigation of dimensions of social–emotional classroom behavior and school readiness for low-income urban preschool children. *School Psychology Review*, *36*, 44–62.
Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.
Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., ... Jil, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175. http://dx.doi.org/10.1007/s11121-005-5553-y.
Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of American Child and Adolescent Psychiatry*, *40*, 1337–1345.
Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology*, *27*, 17–24. http://dx.doi.org/10.1023/A:1022658222914.
Gottfredson, G. D., Pas, E. T., Nebbergall, A. J., Nese, J. F., Strein, W., & Shaw, F. (2010). *An experimental outcome evaluation of the Second Step elementary universal prevention program: Technical report*.
Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586.
Gresham, F. (1986). Conceptual issues in the assessment of social competence in children. In P. S. Strain, M. J. Guralnick, & H. M. Walker (Eds.), *Children s social behavior: Development, assessment, and modification* (pp. 143–179). New York: Academic Press.
Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Liu, P. Y., Asher, K. N., Beland, K., ... Rivara, F. P. (1997). Effectiveness of a violence prevention curriculum among children in elementary school: A randomized controlled trial. *JAMA*, *277*(20), 1605–1611.
Hannan, P. J., & Murray, D. M. (1996). Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed model and the logistic mixed model analyses in simulated community trials with a dichotomous outcome variable at the individual level. *Evaluation Review*, *20*, 338–352. http://dx.doi.org/10.1177/0193841X9602000305.
Hanson, T., Dietsch, B., & Zheng, H. (2012). *Lessons in character impact evaluation (No. 2012-4004)*. Research Report.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107–128. http://dx.doi.org/10.3102/10769986006002107.
Holsen, I., Iversen, A. C., & Smith, B. H. (2009). Universal social competence promotion programme in school: Does it work for children with low socio-economic background? *Advances in School Mental Health Promotion*, *2*, 51–60. http://dx.doi.org/10.1080/1754730X.2009.9715704.
Holsen, I., Smith, B. H., & Frey, K. S. (2008). Outcomes of the social competence program Second Step in Norwegian elementary schools. *School Psychology International*, *29*, 71–88. http://dx.doi.org/10.1177/0143034307088504.
Horowitz, J. L., Garber, J., Ciesla, J. A., Young, J. F., & Mufson, L. (2007). Prevention of depressive symptoms in adolescents: A randomized trial of cognitive-behavioral and interpersonal prevention programs. *Journal of Consulting and Clinical Psychology*, *75*, 693. http://dx.doi.org/10.1037/0022-006X.75.5.693.
Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling. 2005.* New York, NY: Guilford.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley & Sons.
Low, S., Van Ryzin, M. J., Brown, E. C., Smith, B. H., & Haggerty, K. P. (2014). Engagement matters: Lessons from assessing classroom implementation of Steps to Respect: A bullying prevention program over a one-year period. *Prevention Science*, *15*, 165–176. http://dx.doi.org/10.1007/s11121-012-0359-1.
Malecki, C. K., & Elliot, S. N. (2002). Children's social behaviors as predictors of academic achievement: A longitudinal analysis. *School Psychology Quarterly*, *17*, 1–23. http://dx.doi.org/10.1521/scpq.17.1.1.19902.
Merrell, K. W., & Gueldner, B. A. (2012). *Social and emotional learning in the classroom: Promoting mental health and academic success.* Guilford Press.
Murray, D. M. (1998). *Design and analysis of group-randomized trials.* New York: Oxford University Press.
Murray, D. M., Hannan, P. J., Pals, S. P., McCowen, R. G., Baker, W. L., & Blitstein, J. L. (2006). A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Statistics in Medicine*, *25*, 378–388. http://dx.doi.org/10.1002/sim.2233.
Nickerson, A., & Fishman, C. (2009). Convergent and divergent validity of the Devereux Student Strengths Assessment. *School Psychology Quarterly*, *24*, 48–59.
Phillips, D. A., & Shonkoff, J. P. (2000). *From neurons to neighborhoods: The science of early childhood development.* National Academies Press.
Proctor, I. E. K., & Brownson, R. C. (2012). Measurement issues in dissemination and implementation research. In R. C. Brownson, G. A. Colditz, & E. K. Proctor (Eds.), *Dissemination and implementation research in health: Translating science to practice* (pp. 261–280). New York: Oxford University Press.
Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist*, *53*(2), 109.
Rennie, D. A., Bull, R., & Diamond, A. (2004). Executive functioning in preschoolers: Reducing the inhibitory demands of the dimensional change card sort task. *Developmental Neuropsychology*, *26*(1), 423–443.
Rimm-Kaufman, S. E., Pianta, R. C., & Cox, M. J. (2000). Teachers' judgments of problems in the transition to kindergarten. *Early Childhood Research Quarterly*, *15*, 147–166. http://dx.doi.org/10.1016/S0885-2006(00)00049-1.
Rueda, M. R., Rothbart, M. K., McCandliss, B. D., Saccomanno, L., & Posner, M. I. (2005). Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(41), 14931–14936.
Sanetti, L. M. H., & Kratochwill, T. R. (2011). An evaluation of the Treatment Integrity Planning Protocol and two schedules of treatment integrity self-report: Impact on implementation and accuracy. *Journal of Educational and Psychological Consultation*, *21*, 284–308.
SAS Institute (2009). *SAS/STAT® 9.2 user's guide* (2nd ed.). Cary, NC: SAS Institute, Inc (Retrieved March 10, 2010, from the SAS Product Documentation web site: http://support.sas.com/documentation/index.html).
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. http://dx.doi.org/10.1037/1082-989X.7.2.147.
Shapiro, E. S., & Kratochwill, T. R. (2000). *Behavioral assessment in schools: Theory, research, and clinical foundations.* Guilford Press.
Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). A review of evidence based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children*, *31*, 351–380. http://dx.doi.org/10.1353/etc.0.0007.
Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*, 15–21. http://dx.doi.org/10.3102/0013189X031007015.
Smolkowski, K., Strycker, L., & Seeley, J. R. (2013). The role of research evaluation for interventions on school-related behavior disorders. In H. M. Walker, & F. M. Gresham (Eds.), *Handbook of evidence-based practices for students having emotional and behavioral disorders* (pp. 552–566). New York: Guilford.
Sutherland, K. S., & Wehby, J. H. (2001). Exploring the relationship between increased opportunities to respond to academic requests and the academic and behavioral outcomes of students with EBD. *Remedial and Special Education*, *22*, 113–121.
Viera, A. J., & Garrett, G. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*, 360–363.
Walker, H. M., Horner, R. H., Sugai, G., Bullis, M., Sprague, J. R., Bricker, D., & Kaufman, M. (1996). Integrated approaches to preventing antisocial behavior patterns among school-age children and youth. *Journal of Emotional and Behavioral Disorders*, *4*, 194–209. http://dx.doi.org/10.1177/106342669600400401.
What Works Clearinghouse (2014). *Procedures and standards handbook (Version 3.0).* Washington DC: U.S. Department of Education, Institute of Education Sciences (Retrieved from the Institute of Education Sciences, National Center for Education Evaluation, WWC: http://ies.ed.gov/ncee/wwc/).
www.centerforresilientchildren.org
www.census.gov
Zins, J. E., Bloodworth, M. R., Weissberg, R. P., & Walberg, H. J. (2007). The scientific base linking social and emotional learning to school success. *Journal of Educational and Psychological Consultation*, *17*, 191–210. http://dx.doi.org/10.1080/10474410701413145.