Contemporary School Psychology (2023) 27:198–213
https://doi.org/10.1007/s40688-022-00441-5

TOOLBOX Effects on K-2 Student Growth in Social-Emotional Competence: Evidence from a Quasi-Experimental Study

Valerie B. Shapiro[1] · Juyeon Lee[2] · Bo-Kyung Elizabeth Kim[3]

Accepted: 24 October 2022 / Published online: 12 November 2022

Abstract

TOOLBOX is a school-based social and emotional learning (SEL) program to teach and reinforce 12 tools (e.g., Breathing Tool, Courage Tool) for healthy youth development. Although TOOLBOX has been broadly adopted, it remains largely untested. This quasi-experimental study aims to examine the relationship between TOOLBOX implementation and the growth of social-emotional competence (SEC) among K-2 students. First, we compare the growth trajectories of SEC between TOOLBOX and non-TOOLBOX conditions during one academic year. Then, within the TOOLBOX condition, we compare two implementation directives—TOOLBOX Standard and TOOLBOX Primer—to assess (a) the extent to which TOOLBOX was implemented and (b) the SEC growth trajectories among youth. Two district schools were given an implementation directive and materials to implement TOOLBOX Standard curriculum (full, structured lesson plans and strategies with higher dosage anticipated), and two district schools to implement TOOLBOX Primer (basic, introductory lesson plans and strategies with lower dosage anticipated). There were no dosage benchmarks available for either condition. Two additional district schools practiced as usual (non-TOOLBOX). The study sample consisted of 1766 K-2 students and their 80 classroom teachers. Student SEC was measured three times using the Devereux Student Strengths Assessment-Mini, a teacher-completed brief behavioral rating scale. Implementation dosage and quality were measured three times using teacher self-reports. Latent growth modeling and *t* tests were conducted. In Fall, SEC did not differ between TOOLBOX and non-TOOLBOX conditions. TOOLBOX students, however, showed a greater increase in SEC over the year. Within the TOOLBOX condition, Standard teachers incorporated the "tools" into academic instruction more frequently in Spring compared to Primer teachers. For most other implementation variables, no statistical difference was found between TOOLBOX implementation directives. Student SEC growth trajectories did not differ between Standard and Primer schools. Quasi-experimental evidence supports the overall benefits of TOOLBOX on K-2 students' social and emotional growth. We discuss possible reasons for non-differential outcomes between Standard and Primer implementation directives within the TOOLBOX condition (e.g., gap between implementation directives and actual implementation behaviors). The gap between implementation directives and behaviors may provide an impetus for practicing school psychologists to contribute to more robust forms of implementation leadership (e.g., remove obstacles for implementation) and offer implementation support (e.g., coaching, feedback loops).

Keywords Social and emotional learning (SEL) · Program evaluation · Social-emotional competence · Growth trajectory · Program implementation · TOOLBOX

Introduction

Social and emotional learning (SEL) seeks to promote the social-emotional competencies (SEC) essential to healthy relationships, education and job success, and engaged citizenry (Greenberg & Weissberg, 2018). SEL is often delivered universally (i.e., to all children) within a school or out-of-school time setting through free-standing skill-focused lessons. In a multi-tiered system of support, SEL is often performed as a Tier I intervention, intended to prevent problem behaviors

✉ Valerie B. Shapiro
vshapiro@berkeley.edu

[1] University of California, Berkeley, 120 Haviland Hall, Berkeley, CA 94720-7400, USA

[2] The University of Hong Kong, Pok Fu Lam, Hong Kong

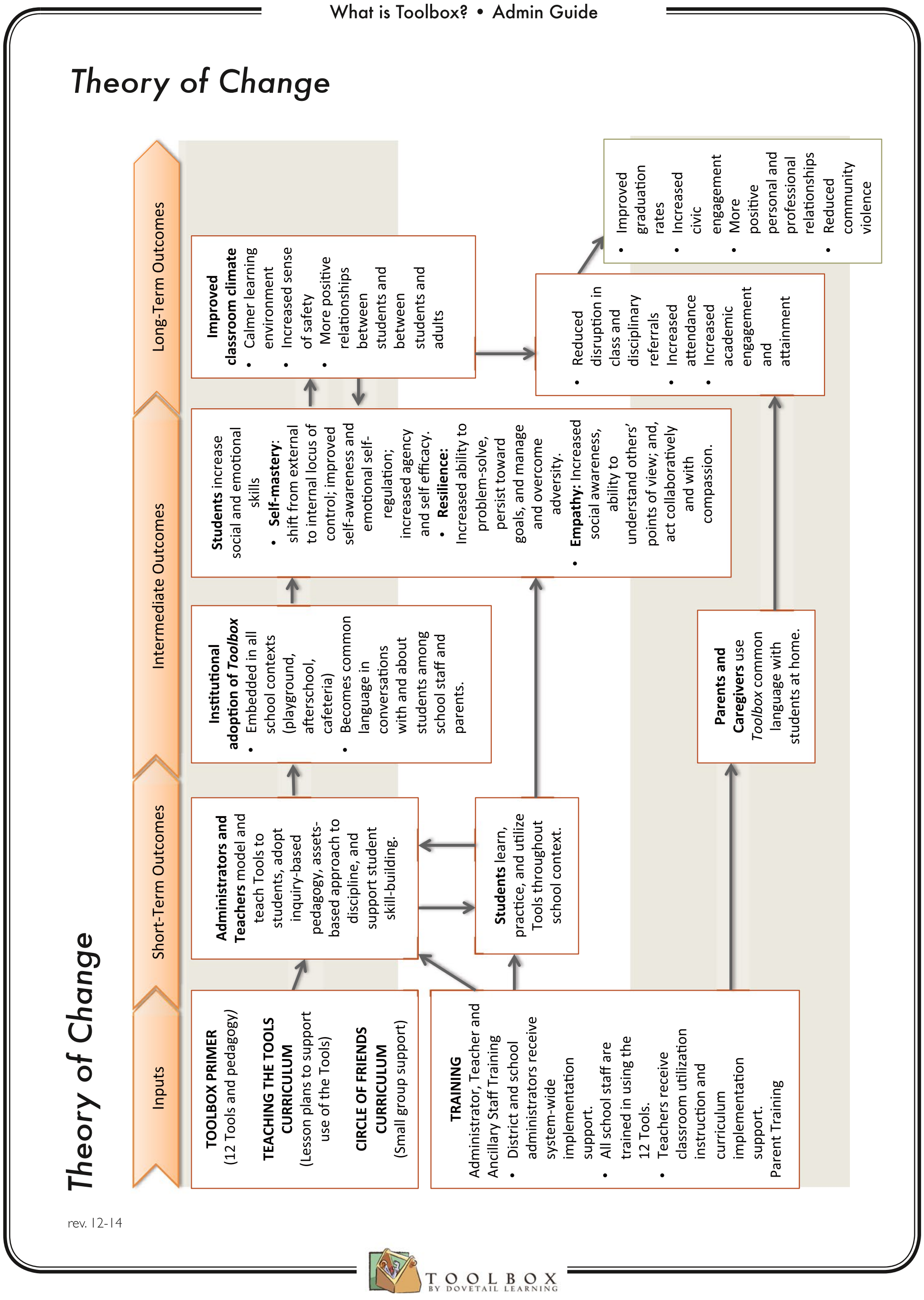
[3] University of Southern California, Los Angeles, USA

and encourage positive behaviors in all students (Collins et al., 2016; Kilgus et al., 2015). This model of SEL program has a considerable evidence base (e.g., Corcoran et al., 2018; Durlak et al., 2011; Sklad et al., 2012; Taylor et al., 2017), which has been leveraged to call for the wide-scale adoption of programs, such as those featured in successive program guides (2005, 2013, 2015) created by the Collaborative for Academic, Social, and Emotional Learning (CASEL). Significant challenges have been noted, however, in the uptake of free-standing, skills-focused SEL programs (Fagan et al., 2015; Jones & Bouffard, 2012). Primary concerns are with their "fit" into the school day as "one more thing" to do (Weston et al., 2018), the diversity of students upon which effective programs have been tested (Rowe & Trickett, 2018), and whether lessons typically generalize out of their specific instructional context, especially in the absence of positive school climates, relationships, and other scaffolds and supports (Darling-Hammond et al., 2020). Recently, SEL approaches have begun evolving to include (a) the integration of SEL into academic curricula, (b) fostering teachers' own wellbeing and capacities to promote SEC as part of their instruction, and (c) whole-school strategies (Domitrovich et al., 2017; Dusenbury et al., 2015). This paper considers the effectiveness of a popular SEL program and explores student growth across two different implementation directives, which differentially emphasize lesson-based delivery and integration strategies.

TOOLBOX (Collin, 2015) is a universal, school-based SEL program that aims to promote elementary school children's SEC through the instruction and reinforcement of 12 tools (e.g., Breathing Tool, Listening Tool, Courage Tool). The developer employed the metaphor of *tools* to provide students with rhetorical devices that call upon their intrinsic capacities to achieve emotional, social, and academic wellbeing and resilience. While TOOLBOX is taught formally as a curriculum, the essence of TOOLBOX is providing a common language and practices used across contexts throughout the school day. In TOOLBOX, adults are asked to "go first" and internalize the 12 tools for their own wellbeing. Then, teachers are trained to use an inquiry-based approach to support student discovery and decision-making. For example, teachers are trained to ask children "what three tools might you try?" during a classroom conversation or "what tools did you try?" when trying to unpack a playground conflict. Each child constructs a manila toolbox to keep at their work station that can be personalized and referenced. Each tool has an icon and hand gesture, enabling teachers to non-verbally suggest a tool to students during community meetings or academic instruction, or to pause and "name" which tools children could use in the moment. Staff hang posters and wear fandecks on lanyards with the tool icons to facilitate quick references. The whole school community uses mantras for the tools (e.g., Patience Tool: I am strong enough to wait!) and invites families to reference the tools at home.

TOOLBOX promotes a flexible approach to program implementation, such that program delivery can occur with an emphasis on longer, free-standing lessons (in the tradition of classic social skills curriculum) or a more integrative "common language" approach of referencing and reinforcing tools after only a brief introductory lesson. The Standard implementation strategy provides structured stand-alone lesson plans and comprehensive resources for delivery while the Primer implementation strategy provides only the brief, introductory "light touch" lessons to the 12 tools and only the most essential resources for delivery. Typically, district or school leaders choose the overall implementation approach that they believe to be best suited for their educational context. The choice between the *Standard* implementation strategy and the *Primer* implementation strategy depends on local resources, readiness, and preferences. Theoretically, higher dosage (i.e., how much of the program components have been delivered; Durlak & DuPre, 2008) is anticipated in the Standard implementation than in the Primer implementation. Within either of these leadership-selected implementation choices, classroom teachers vary in their delivery of program components through lesson-based methods (e.g., explicit instruction about the concepts of each tool) and non-lesson-based strategies (e.g., modeling how to use tools, incorporating tools into academic curriculum, and applying tools to daily classroom interactions) based on their individual strengths, resources, constraints, and preferences.

These flexible implementation features had made TOOLBOX appealing to many educators. More than 40 school districts in Northern California have implemented TOOLBOX. Yet, only two unpublished studies have been conducted to date to examine the program's theory of change (see Fig. 1), thus far exploring (a) the acceptability and utility of training and resource inputs, (b) the presence and strength of instructional output in the classroom, and (c) the proximal student outcomes perceived by teachers, all conducted in contexts of educational leaders selecting the standard implementation protocols (De Long-Cotty, 2010; Dovetail Learning, 2013). These studies found that elementary school teachers positively rated the value of program materials and training resources, implemented lessons and other delivery strategies in their classrooms, and observed students using the "tools" and encouraging others to use the tools. Through pre/post comparison on the Behavioral and Emotional Rating Scale (BERS; Epstein, 2004; Epstein & Sharma, 1998), teacher ratings of students' intrapersonal and affective strengths increased over 3 months. Change was not detected, however, on BERS teacher ratings of interpersonal strengths or school functioning, nor on any of the BERS parent rating scales (De Long-Cotty, 2010). No prior study has compared students experiencing TOOLBOX to a comparison group, comparing student development under TOOLBOX conditions to the typical maturation of social-emotional development.



What is Toolbox? • Admin Guide

Theory of Change

Theory of Change

Inputs

Short-Term Outcomes

Intermediate Outcomes

Long-Term Outcomes

TOOLBOX PRIMER
(12 Tools and pedagogy)

TEACHING THE TOOLS CURRICULUM
(Lesson plans to support use of the Tools)

CIRCLE OF FRIENDS CURRICULUM
(Small group support)

TRAINING
Administrator, Teacher and Ancillary Staff Training
- District and school administrators receive system-wide implementation support.
- All school staff are trained in using the 12 Tools.
- Teachers receive classroom utilization instruction and curriculum implementation support.

Parent Training

Administrators and Teachers model and teach Tools to students, adopt inquiry-based pedagogy, assets-based approach to discipline, and support student skill-building.

Students learn, practice, and utilize Tools throughout school context.

Institutional adoption of *Toolbox*
- Embedded in all school contexts (playground, afterschool, cafeteria)
- Becomes common language in conversations with and about students among school staff and parents.

Parents and Caregivers use *Toolbox* common language with students at home.

Students increase social and emotional skills
- **Self-mastery**: shift from external to internal locus of control; improved self-awareness and emotional self-regulation; increased agency and self efficacy.
- **Resilience:** Increased ability to problem-solve, persist toward goals, and manage and overcome adversity.
- **Empathy:** Increased social awareness, ability to understand others' points of view; and, act collaboratively and with compassion.

Improved classroom climate
- Calmer learning environment
- Increased sense of safety
- More positive relationships between students and between students and adults

- Reduced disruption in class and disciplinary referrals
- Increased attendance
- Increased academic engagement and attainment

- Improved graduation rates
- Increased civic engagement
- More positive personal and professional relationships
- Reduced community violence

rev. 12-14

TOOLBOX BY DOVETAIL LEARNING

Fig. 1 TOOLBOX theory of change as communicated in the 2014 TOOLBOX Project Administrator's Guide (Copyright Mark A. Collin. All rights reserved.)

This quasi-experimental study aims to examine the relationship between TOOLBOX, as implemented in *routine* school settings (i.e., no additional coaching or technical assistance made available by nature of being studied), and the development of K-2 students' SEC. This study first explored the overall effectiveness of TOOLBOX by comparing student SEC growth trajectories between TOOLBOX and non-TOOLBOX conditions over one academic year. Then, within the TOOLBOX condition, this study examined the extent to which TOOLBOX was implemented differently across two implementation directives (i.e., Standard and Primer) as well as the extent to which the two implementation directives had differential effects on student SEC growth trajectories. The research questions and hypotheses are as follows:

- Question 1: To what extent was the TOOLBOX intervention related to growth trajectories of students' SEC? Hypothesis 1: Students in the TOOLBOX conditions will have higher rates of growth in SEC as compared to students in non-TOOLBOX conditions.

- Question 2: To what extent was TOOLBOX implemented across two different implementation directives? Hypothesis 2: Although there are no benchmarks to guide specific hypotheses about dosage, we hypothesize that standard implementation teachers will report higher levels of implementation dosage as compared to Primer implementation teachers. Indicators of implementation quality will not be different across TOOLBOX implementation directives.
- Question 3: To what extent were the TOOLBOX implementation directives related to growth trajectories of students' SEC? Hypothesis 3: Students in schools with the Standard implementation directive will have higher rates of growth in SEC compared to students in schools with the Primer implementation directive.

Method

Design and Sample

The TOOLBOX Implementation Research Project (TIRP) aimed to understand variation in the routine implementation of TOOLBOX, as distributed, at the time of study, by Dovetail Learning, and to explore the relationship between one academic year of TOOLBOX implementation and student outcomes. The TIRP is a quasi-experimental study situated within a single California school district (who initiated practice-driven research). According to the publicly available district statistics, 59.1% of elementary school students in this district were identified as Hispanic/Latinx, 39.8% were English language learners, 71.3% were eligible for free and reduced priced lunch, and students meeting or exceeding the state educational standards in were 23.2% English language arts and 24.0% in mathematics (District, 2016).

Funding enabled four elementary schools to initiate TOOLBOX, and two comparison schools to participate in a measurement-only, practice-as-usual (non-TOOLBOX) condition, during the 2015–2016 academic year. Schools were assigned to conditions in a way that intentionally distributed student demographic characteristics as evenly as possible across TOOLBOX and non-TOOLBOX conditions. Of the four TOOLBOX schools, two were given resources to implement the TOOLBOX Standard package, and the other two were given resources to implement the TOOLBOX Primer package.

In August, prior to the beginning of the fall semester, a 6-hour training was provided to teachers and staff from the four TOOLBOX schools, and 94% of classroom teachers attended the training. More detailed descriptions of the training and findings from a post-training teacher survey (including teacher attitudes, capacities, expectations for implementation and impact, etc.) are presented in a paper by Shapiro et al. (2020). To monitor implementation, the SEL Implementation Survey (SEL-IS), a self-report survey on program implementation behaviors, was administered to teachers and staff in these four schools at three time points throughout the year (October, December, and May). Also, classroom teachers were asked to complete the Devereux Student Strengths Assessment-Mini (DESSA-Mini), a brief 8-item behavioral rating scale assessing student SEC, at three time points throughout the year (October, December/January, and April/May). All research protocols were approved by the institutional review board at the University of California, Berkeley.

The current study uses a sample of 1766 K-2 students. The mean age of students in the sample at the beginning of the year was 6.05 years (SD = 0.89), and 48.6% of students in the sample were female. Administrative records provided by the district indicated that more than half of the students (55.3%) in the sample were identified as Hispanic/Latinx, followed by Asian/Asian American (13.6%), Black/African American (10.9%), White (7.5%), and Others (7.5%, including Filipinx, Pacific Islander, and American Indian/Native American); about half of the students in the sample (48.2%) were identified as English language learners (ELL; primary language used at home includes Spanish [70.9%], Cantonese [12.8%], Tagalog [4.3%], Vietnamese [4.3%], and Arabic [3.1%]); 8% of children in the sample were receiving special education (SPED) services; and 67.8% were eligible for free and reduced price lunch (FRL) based on their household economic status.

There were 562 students in schools under the TOOLBOX Standard implementation directive, 608 in schools under the Primer implementation directive, and 596 in the non-TOOLBOX condition. No difference was observed in the distribution of gender, age, SPED, and FRL across three study conditions. Race/ethnicity distribution differed between TOOLBOX and non-TOOLBOX conditions ($\chi^2(5) = 20.72$, $p < 0.01$, Cramer's $V = 0.11$, indicating a small-to-medium difference): in the TOOLBOX condition, there were fewer Hispanic/Latinx (51.9% versus 62.1%) and more Black/African American (12.6% versus 7.6%) students. Within the TOOLBOX condition, race/ethnicity distribution ($\chi^2(5) = 36.60$, $p < 0.001$, Cramer's $V = 0.18$, indicating a medium-to-large difference) as well as ELL status ($\chi^2(1) = 10.91$, $p < 0.01$, $\phi = 0.10$, indicating a small difference) were different between Standard and Primer implementation directives: in TOOLBOX Standard, there were more Black/African American students (15.8% versus 9.5%), fewer Asian/Asian American students (10.5% versus 18.4%), fewer Others (5.2% versus 10.5%), and fewer ELL students (41.3% versus 51.6%). Table 1 describes student demographic characteristics for the entire student sample and disaggregated by study condition.

Students were nested within 85 classrooms (26 Standard, 30 Primer, 29 non-TOOLBOX). In the current study, a

Table 1 Student demographic characteristics by study condition

	Total (N = 1766)		TOOLBOX (N = 1170)				Non-TOOLBOX (N = 596)	
			Standard (N = 562)		Primer (N = 608)			
	N	%	N	%	N	%	N	%
Age	M = 6.05	SD = 0.89	M = 6.01	SD = 0.95	M = 6.08	SD = 0.84	M = 6.05	SD = 0.87
Boy	907	51.4	284	50.5	321	52.8	302	50.7
Girl	859	48.6	278	49.5	287	47.2	294	49.3
Hispanic/Latinx	977	55.3	300	53.4	307	50.5	370	62.1
Asian/Asian American	240	13.6	59	10.5	112	18.4	69	11.6
Black/African American	192	10.9	89	15.8	58	9.5	45	7.6
White	133	7.5	49	8.7	39	6.4	45	7.6
Others	133	7.5	29	5.2	64	10.5	40	6.7
Missing race/ethnicity	91	5.2	36	6.4	28	4.6	27	4.5
Non-ELL	829	46.9	296	52.7	269	44.2	264	44.3
ELL	851	48.2	232	41.3	314	51.6	305	51.2
Missing ELL	86	4.9	34	6.0	25	4.1	27	4.5
Non-SPED	1538	87.1	486	86.5	535	88.0	517	86.7
SPED	142	8.0	42	7.5	48	7.9	52	8.7
Missing SPED	86	4.9	34	6.0	25	4.1	27	4.5
Non-FRL	482	27.3	155	27.6	177	29.1	150	25.2
FRL	1198	67.8	373	66.4	406	66.8	419	70.3
Missing FRL	86	4.9	34	6.0	25	4.1	27	4.5

ELL English language learner, *SPED* special education, *FRL* free and reduced priced lunch eligibility

sample of 80 K-2 classroom teachers is used (25 Standard, 28 Primer, 27 non-TOOLBOX), excluding five non-responding/research-consenting teachers. The majority of teachers (93.8%) in the sample identified themselves as female. Sixty-five percent of teachers in the sample identified as White, 17.5% as Hispanic/Latinx, and 8.8% as Asian/Asian American. No difference was found in the distribution of teachers' gender and race/ethnicity across study conditions.

Measurement

Social-Emotional Competence

Teachers assessed students' SEC using the DESSA-Mini, a strength-based behavioral rating scale (Naglieri et al., 2011a), at three time points throughout the year of 2015–2016: October (Fall), December/January (Winter), and April/May (Spring). Following the recommended use of the DESSA-Mini, teachers rated the frequency (*never* = 0, *rarely* = 1, *occasionally* = 2, *frequently* = 3, *very frequently* = 4) of students' positive behaviors (e.g., do something nice for somebody) over the past 4 weeks (Simmons et al., 2016). The sum of eight items—transformed into a T score based on national norms (an expected sample mean of 50 and standard deviation of 10)—yields a social-emotional total (SET) score for each student at each time (Naglieri et al., 2013). The adequacy of the DESSA norms has been independently reviewed (e.g., Atlas, 2010; Malcomb, 2010) and determined to be sufficiently large and diverse (Merrell & Gueldner, 2010). The DESSA-Mini is a brief version of the DESSA, which assesses 8 social and emotional competencies: self-awareness (e.g., describe how they were feeling; 7 items), social-awareness (e.g., get along with different types of people; 9 items), self-management (e.g., stay calm when faced with a challenge; 11 items), goal-directed behavior (e.g., keeping trying when unsuccessful; 10 items), relationship skills (e.g., express concern for another person; 10 items), personal responsibility (e.g., remember important information; 10 items), decision-making (e.g., learn from experience; 8 items), and optimistic thinking (e.g., look forward to classes or activities at school; 7 items). The DESSA-Mini has four alternative forms, each comprised of eight different indicators of social-emotional competence from the full DESSA. The alternative forms can be used in rotation to limit practice effects (LeBuffe et al., 2018; Lee et al., 2022a). Prior studies have shown that alternative form reliability meets or exceeds 0.90 across all forms. DESSA-Mini forms 1 (Fall), 2 (Winter), and 3 (Spring) were used in this study.

The DESSA-Mini has been evaluated against commonly accepted criteria for brief behavior rating scales measuring social, emotional, and behavioral risks (e.g., Glover & Albers, 2007; Jenkins et al., 2014). DESSA-Mini scores have been shown to be reliable (e.g., Shapiro et al., 2017a); sensitive and specific (e.g., Naglieri et al., 2011b); discerning between children with and without mental problems, emotional problems, behavioral problems, impairments, and adaptive skills (e.g., Goldstein & Naglieri, 2016; Nickerson & Fishman, 2009; Shapiro & Lebuffe, 2006); and predictive of serious disciplinary infractions (Shapiro et al., 2017b) and academic achievement (Balfanz & Byrnes, 2020; Chain et al., 2017). The DESSA and DESSA-Mini were designed with strategies to avoid rating bias (Mahoney et al., 2022), have been used with diverse populations across diverse settings (Hwang et al., 2022), and have been empirically tested for measurement invariance across subgroups of students as characterized by gender, race and ethnicity, special education, English language learning, and socioeconomic status (Lee et al., 2022b). The vast majority of variance in DESSA-Mini scores is attributable to differences between students, relative to differences attributable to differences between teachers (Shapiro et al., 2016; Tanner et al., 2018).

Implementation Variables

Implementation dosage (e.g., count, frequency, duration) and quality were measured through the SEL-IS, a self-report survey on program implementation behaviors, at three time points throughout the year of 2015–2016: October (Fall), December (Winter), and May (Spring). We used both composite indicators (an average of multiple related items) and individual items in our comparisons to balance the discordant desires of limiting the likelihood of a type 2 error through multiple comparisons, reducing measurement error, and promoting interpretability and clear implications for practice. Where composite scores were generated, we report an assessment of their internal reliability (i.e., the extent to which the items on the same scale measure the same underlying construct) using Cronbach's alpha (Cronbach, 1951). Noting that Cronbach's alpha is highly influenced by the number of items on a given scale, and the need for practical scales for applied uses to be brief, a 0.60 criterion was used to indicate acceptable internal reliability (Schmitt, 1996).

Implementation dosage was measured using four distinct dose forms (i.e., ways of delivering program components): lesson delivery, modeling, incorporation, and application. Lesson delivery dosage was measured by summing up the number of lessons teachers reported teaching until the time of survey administration (by checking off the ones they instructed from the list of 17 total possible lessons). Teachers were then asked to select their "most favorite" and "least favorite" lesson from the ones they reported teaching as a referent for some of the subsequent questions. Modeling dosage was measured by two items: the frequency of teachers (a) using tools themselves in the classroom and (b) telling students the tools teachers need in the moment. Incorporation dosage was measured by three items: the frequency of incorporating tools into (a) writing, (b) literature, and (c) arts. Application dosage was measured by three items: the frequency of (a) discussing and (b) asking how students can use tools in their daily lives and (c) naming tools in the moment that students are using. Modeling, incorporation, and application dosage items were scored using a 5-point Likert scale (0 = never, 1 = rarely, 2 = occasionally, 3 = often, and 4 = very frequently). Cronbach's alpha coefficients for these eight non-lesson-based (i.e., modeling, incorporation, and application) dosage items were 0.76 in Fall, 0.65 in Winter, and 0.86 in Spring. In addition to these count and frequency indicators of dosage, duration was also measured by three items: time spent (a) teaching most favorite lesson, (b) teaching least favorite lesson, and (c) using other strategies outside of lesson structure in a typical week (0 = less than 10 min, 1 = 20 min, 2 = 30 min, 3 = 40 min, 4 = 50 min or more). Cronbach's alpha coefficients for the two lesson-based duration items were 0.55 in Fall, 0.74 in Winter, and 0.66 in Spring.

Implementation quality was measured by two items asking teaching quality when teaching their most favorite lesson and their least favorite lesson, in addition to one item asking the quality when using strategies outside of lesson structure (1 = F to 13 = A +). Cronbach's alpha coefficients for the two lesson-based quality items were 0.74 in Fall, 0.63 in Winter, and 0.59 in Spring.

Student Demographic Characteristics

Student demographic characteristics were included as covariates when comparing student SEC growth trajectories by condition. Student age and gender were reported by teachers when they completed the DESSA-Mini. Other student characteristics including race/ethnicity, ELL status, SPED status, and FRL eligibility status were collected from the 2015–2016 district administrative records. Student age in years in Fall was included as a continuous variable. Variables measured dichotomously were dummy coded, including gender (0 = male, 1 = female), ELL status (0 = non-ELL, 1 = ELL), SPED status (0 = no SPED services, 1 = SPED services), and FRL eligibility status (0 = not eligible for FRL, 1 = eligible for FRL). The race/ethnicity variable was transformed using the effect coding method, such that the mean of each subgroup can be compared to the grand mean across all subgroups. This method is especially useful for examining variables like race/ethnicity without assuming any specific group is normative, against which all other groups are compared (Mayhew & Simonoff, 2015).

Analytic Procedures

To investigate question 1, latent growth modeling (LGM) was conducted to assess the extent to which the TOOLBOX intervention was related to student SEC growth trajectory, including the initial level (i.e., intercept) and the rate of change (i.e., slope). The unconditional LGM was first performed to test whether a linear growth trajectory model fit our sample data. Then, the conditional LGM was performed to compare student SEC growth trajectories by intervention condition, while accounting for any variations in outcomes associated with student demographic characteristics. To investigate question 2, independent samples *t* tests were conducted to compare the means of implementation variables at the teacher level between Standard and Primer implementation directives. To investigate question 3, the conditional LGM was performed to compare student SEC growth trajectories between Standard and Primer directives, while accounting for any variations related to student demographic characteristics.

For LGM (questions 1 and 3), a goodness-of-fit was assessed following Hair and colleagues' (2009) fit assessment guidelines for a moderately complex model (i.e., having 12 to 30 observed variables) with a large sample of $n > 250$). They suggest that a significant chi-square statistic ($p < 0.05$) is expected, while a comparative fit index (CFI) or Tucker-Lewis index (TLI) higher than 0.92, a standardized root mean squared residual (SRMR) lower than 0.08, and a root mean square error of approximation (RMSEA) lower than 0.07 demonstrate goodness-of-fit. LGM was conducted using the full information maximum likelihood method with a sandwich estimator in Mplus version 8 in order to address combinations of different types of predictor variables, compute standard errors that are robust to non-normality, and handle missing data efficiently (Muthén & Muthén, 1998). For independent samples *t* tests (question 2), Levene's tests for homogeneity of variances were first conducted. If there was a violation of the homogeneity assumption, the degrees of freedom were adjusted using the Welch-Satterthwaite method. An alpha level of 0.05 was applied to assess a statistically significant difference between group means. *t* tests were conducted in SPSS version 25.

Results

Table 2 presents descriptive statistics and missing rates of DESSA-Mini scores at each time point by condition. In our sample, the total average SEC score in Fall was $M = 50.37$ (SD = 11.08), which is close to the national norm of $M = 50$ (SD = 10), with no differences in Fall SEC across study conditions ($F(2, 1622) = 0.33$, $p = 0.72$). DESSA-Mini scores were 8% missing in Fall, 15.6% in Winter, and 15.1% in Spring. The missing rates did not differ by condition in Spring, but differed at the first two waves. In Fall, there were more missing data in the TOOLBOX condition versus non-TOOLBOX (10.3% versus 3.5%; $\chi^2(1) = 24.36$, $p < 0.001$) and in Primer versus Standard (18.4% versus 1.4%; $\chi^2(1) = 91.67$, $p < 0.001$). In Winter, more data were missing in the non-TOOLBOX sample versus TOOLBOX sample (30.2% versus 8.1%, $\chi^2(1) = 146.45$, $p < 0.001$) and in Primer versus Standard (12.7% versus 3.2%; $\chi^2(1) = 35.05$, $p < 0.001$). We present this information transparently, since the design is quasi-experimental.

Table 2 Student SEC descriptive statistics and missing rates by study condition

	M	SD	Skewness		Kurtosis		Missing *N*	
			Est	SE	Est	SE	*N*	%
Total (*N* = 1766)								
Fall	50.37	11.08	0.24	0.06	−0.66	0.12	141	8.0
Winter	50.97	11.26	0.17	0.06	−0.68	0.13	275	15.6
Spring	53.14	11.66	0.13	0.06	−0.91	0.13	267	15.1
TOOLBOX Standard (*N* = 562)								
Fall	50.65	11.21	0.28	0.10	−0.76	0.21	8	1.4
Winter	51.44	11.20	0.16	0.10	−0.71	0.21	18	3.2
Spring	53.79	11.90	0.10	0.11	−1.05	0.22	65	11.6
TOOLBOX Primer (*N* = 608)								
Fall	50.34	11.20	0.14	0.11	−0.58	0.22	112	18.4
Winter	51.47	11.28	0.15	0.11	−0.62	0.21	77	12.7
Spring	53.97	11.73	−0.01	0.11	−0.86	0.22	106	17.4
Non-TOOLBOX (*N* = 596)								
Fall	50.12	10.84	0.31	0.10	−0.63	0.2	21	3.5
Winter	49.73	11.23	0.23	0.12	−0.70	0.24	180	30.2
Spring	51.68	11.22	0.31	0.11	−0.72	0.22	96	16.1

The unconditional linear growth trajectory model, with actual measurement time span reflected in slope loadings (i.e., 0, 0.28, 1) and equal residual variance assumed across three time points, showed an acceptable fit ($\chi^2(3) = 9.66$, $p = 0.02$; CFI = 0.996; TLI = 0.996; SRMR = 0.032, RMSEA = 0.035, 90% CI = [0.012, 0.062]). In this unconditional model, the mean intercept was estimated to be 50.30, and the average rate of growth over the academic year was 2.60 T score points. The correlation between intercept and slope was not statistically significant ($r = -0.07$, $p = 0.27$), suggesting that across the full sample, a student's SEC level at the start of the year did not predict a student's growth in SEC throughout the year.

Question 1: to What Extent Was the TOOLBOX Intervention Related to K-2 Students' Growth Trajectories of Social-Emotional Competence?

Table 3 presents the estimation results of these two conditional growth trajectory modeling approaches. The model without any covariates adjusted showed an acceptable fit to our data ($\chi^2(4) = 11.39$, $p = 0.02$; CFI = 0.996; TLI = 0.994; SRMR = 0.029; RMSEA = 0.032, 90% CI = [0.01, 0.06]). The intercept was not different by the TOOLBOX intervention condition ($b = 0.89$, $p = 0.10$), but the slope differed by 1.60 point ($p < 0.01$) by Spring. The mean slope for non-TOOLBOX students was estimated to be 1.53 T score point over the year, and the mean slope for TOOLBOX students was estimated to be 3.13 T score points. After including demographic covariates, the model still showed an acceptable fit ($\chi^2(13) = 21.22$, $p = 0.07$; CFI = 0.996; TLI = 0.990; SRMR = 0.014; RMSEA = 0.020, 90% CI = [0.00, 0.04]). Holding all the other covariates constant, the intercept did not differ by the TOOLBOX intervention condition ($b = 0.68$, $p = 0.19$), but the slope differed by 1.74 point ($p < 0.01$) by Spring. These findings suggest that although student SEC started at a similar level in Fall, TOOLBOX students, on average, demonstrated a higher rate of growth in SEC than non-TOOLBOX students over the year.

Table 3 Difference in student SEC growth trajectory by intervention condition

	Model 1: no covariates			Model 2: covariates adjusted		
	Est	SE	*p*	Est	SE	*p*
Intercept regressed on						
TOOLBOX	0.89	0.54	0.10	0.68	0.52	0.19
Girl				4.45	0.51	0.00
Age				0.91	0.27	0.00
Hispanic				0.58	0.43	0.18
Asian				2.65	0.64	0.00
Black				−3.47	0.69	0.00
White				−0.13	0.72	0.86
ELL				−0.95	0.56	0.09
SPED				−8.37	0.87	0.00
FRL				−0.65	0.57	0.25
Slope regressed on						
TOOLBOX	1.60	0.50	0.00	1.74	0.51	0.00
Girl				0.99	0.50	0.05
Age				−0.81	0.29	0.01
Hispanic				0.45	0.44	0.31
Asian				2.01	0.63	0.00
Black				−1.63	0.69	0.02
White				−1.25	0.70	0.07
ELL				0.95	0.56	0.09
SPED				0.41	0.79	0.61
FRL				−0.64	0.58	0.27
Correlation						
Intercept with slope	−0.07	0.06	0.25	−0.12	0.07	0.07
Mean						
Intercept	49.70	0.43	0.00	43.50	1.76	0.00
Slope	1.53	0.39	0.00	5.53	1.91	0.00

ELL English language learner, *SPED* special education, *FRL* free and reduced priced lunch eligibility

Question 2: to What Extent Was TOOLBOX Implemented Across Two Different Implementation Directives?

Table 4 presents descriptive statistics of implementation variables (both composite scales and individual items) by implementation directive and the corresponding *t* test results. On average, teachers taught about 5 lessons in Fall, 8 lessons by Winter, and 11 lessons by Spring. Besides the cumulative number of lessons taught, no clear descriptive pattern was observed for other implementation variables. *t* test results showed no scale-level differences in teacher reports of their implementation behaviors between Standard and Primer implementation directives. Item-level analysis revealed that Standard teachers tended to use the three incorporation strategies (i.e., incorporating tools within academic curricula) more frequently in Spring than Primer teachers: incorporation into writing (Standard $M = 1.35$, Primer $M = 0.64$, $t(43) = 2.67$, $p < 0.05$), into literature (Standard $M = 2.17$, Primer $M = 1.41$, $t(43) = 2.23$, $p < 0.05$), and into arts and crafts (Standard $M = 1.43$, Primer $M = 0.73$, $t(43) = 2.59$, $p < 0.05$). In addition, Standard teachers reported a higher level of quality of teaching their least favorite lesson in Spring (Standard $M = 9.09$, Primer $M = 8.00$, $t(43) = 2.27$, $p < 0.05$) relative to Primer teachers. Primer teachers reported *naming tools that students are using in the moment* more frequently in Winter than Standard teachers (Standard $M = 2.33$, Primer $M = 2.76$, $t(43) = -2.05$, $p < 0.05$). For all the other implementation variables, no statistical difference was found between the two implementation directives at the $p < 0.05$ level.

Question 3: to What Extent Was the TOOLBOX Implementation Directives Related to K-2 Students' Growth Trajectories of Social-Emotional Competence?

Table 5 presents the estimation results of these two conditional growth trajectory modeling approaches within the TOOLBOX sample (Standard versus Primer implementation directives). The model without any covariates adjusted showed an acceptable fit to our data ($\chi^2(4) = 1.69$, $p = 0.79$; CFI = 1.00; TLI = 1.00; SRMR = 0.012; RMSEA = 0.000, 90% CI = [0.00, 0.03]). Neither intercept ($b = -0.10$, $p = 0.88$) nor slope ($b = 0.01$, $p = 0.99$) differed by implementation directive. After including demographic covariates, the model still showed an acceptable fit ($\chi^2(13) = 19.37$, $p = 0.11$; CFI = 0.996; TLI = 0.989; SRMR = 0.010, RMSEA = 0.022 [0.00, 0.04]), and neither intercept ($b = 0.92$, $p = 0.14$) nor slope ($b = 0.06$, $p = 0.93$) differed by implementation directive. These findings suggest that the two different TOOLBOX implementation directives, which largely did not change implementation behavior, had no differential effects on student growth in SEC.

Discussion

This quasi-experimental study provides promising evidence to support TOOLBOX effects on K-2 students' social and emotional growth in a routine practice setting. On average, students in TOOLBOX schools gained 3.13 T score points across the school year, 1.60 more *T* score points than non-TOOLBOX students. To interpret the magnitude of these gains, one might compare results from this study of TOOLBOX to a study in which the DESSA-Mini was used to measure the SEC of students receiving the Promoting Alternative Thinking Strategies (PATHS; Kusché & Greenberg, 1994) curriculum (Shapiro et al., 2018). The Blueprints for Healthy Youth Development (Mihalic & Elliott, 2015)—a clearinghouse created to help consumers determine "what works"—lists PATHS as a "model" program. Shapiro et al. (2018) observed that K-2 students exposed to PATHS, with robust technical assistance, gained an average of 3.66 DESSA-Mini *T* score points across the school year (no comparison group available). Therefore, available evidence suggests that TOOLBOX may also be a promising approach for augmenting student SEC.

In response to a secondary aim, this study fails to provide robust evidence that the school-level decision to purchase either the TOOLBOX Standard or the TOOLBOX Primer package (i.e., the school-level implementation directive) differentially shaped teacher implementation behavior or student outcomes. The extra resources provided with the Standard curriculum were not associated with reports of higher-quality instruction. Teachers in the Standard TOOLBOX directive did not report a higher level of quality, with one exception—in their least favorite lesson in the springtime. Also counter to expectation, teachers did not report many differences in dosage across diverse dose forms. Item-level analysis indicated that teachers in the Standard TOOLBOX directive were *more* likely than teachers in the Primer directive to incorporate the program into academic lessons, despite the Standard directive emphasizing stand-alone lessons and the Primer directive emphasizing integration. These findings could imply that more highly scripted lessons have an under-acknowledged benefit as a proxy for professional development, enabling the teachers to learn the material themselves before teaching it, and then to flexibly integrate SEL content into other curricular areas and to teach less-resonate lessons well. This is concordant with prior research indicating a plurality of teachers hold a preference for initial structure with increasing flexibility when implementing a new initiative (Shapiro et al., 2016).

The challenge of taking high-quality SEL to scale is formidable, and innovative approaches to adoption

Table 4 Teacher implementation comparisons between TOOLBOX Standard and Primer directives

Items (range)	Standard		Primer		t statistics	
	M	SD	*M*	SD	*t*	*df*
Dosage indicators						
Number of lessons taught (0–17)						
Fall	5.24	4.15	5.91	3.15	−0.62	45
Winter	8.04	3.69	8.46	4.05	−0.39	49
Spring	11.64	4.26	11.04	4.90	0.48	51
Frequency of using non-lesson-based strategies (past week; 0=never, 1=less than daily or 4 times, 2=at least daily, 3=at least twice daily, 4=more than three times a day or 15 times)						
Fall	1.85	0.65	1.73	0.57	0.65	40
Winter	1.96	0.47	1.99	0.55	−0.24	43
Spring	2.05	0.62	1.74	0.71	1.53	43
Modeling: Using the tools themselves (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	2.61	0.89	2.53	1.02	0.28	40
Winter	2.42	0.88	2.71	1.06	−1.03	43
Spring	2.35	0.71	2.18	1.01	0.64	43
Modeling: Letting students know which tools teachers need (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	2.41	0.80	2.42	1.02	−0.04	39
Winter	2.33	0.87	2.67	1.15	−1.10	43
Spring	2.26	0.75	2.18	1.10	0.28	43
Incorporating into writing lessons (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	0.87	0.87	0.68	0.89	0.68	40
Winter	1.38	1.06	1.00	0.95	1.25	43
Spring	1.35	0.98	0.64	0.79	2.67*	43
Incorporating into literature discussions (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	1.43	1.04	1.11	1.20	0.96	40
Winter	1.54	0.98	1.52	1.12	0.06	43
Spring	2.17	1.11	1.41	1.18	2.23*	43
Incorporating/Expressing through arts and crafts (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	1.04	0.98	0.63	0.76	1.50	40
Winter	1.25	0.90	0.90	1.09	1.16	43
Spring	1.43	0.84	0.73	0.98	2.59*	43
Application: Discussing how to use tools in their daily lives (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	1.83	1.19	1.74	1.15	0.25	40
Winter	2.13	0.74	1.81	0.81	1.36	43
Spring	1.91	0.79	1.73	0.88	0.74	43
Application: Asking how to use tools in their daily lives (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	2.09	1.12	2.32	1.16	−0.65	40
Winter	2.29	0.75	2.57	0.98	−1.08	43
Spring	2.67	0.73	2.59	1.01	0.28	41
Application: Naming tools that students are using (0=never, 1=rarely, 2=occasionally, 3=often, 4=very frequently)						
Fall	2.57	0.95	2.42	1.07	0.46	40
Winter	2.33	0.64	2.76	0.77	−2.05*	43
Spring	2.35	0.83	2.50	1.10	−0.52	43
Time spent teaching lessons (average; 0=less than 10m, 1=20m, 2=30m, 3=40m, 4=50m or more)						
Fall†	2.53	1.01	2.63	1.27	−0.27	37
Winter	2.98	1.30	2.55	0.99	1.24	42
Spring	3.61	1.20	3.24	1.13	1.08	43

Table 4 (continued)

Items (range)	Standard		Primer		t statistics	
	M	SD	*M*	SD	*t*	*df*
Time spent teaching most favorite lesson (0=less than 10m, 1=20m, 2=30m, 3=40m, 4=50m or more)						
Fall	2.84	1.34	2.47	1.26	0.87	36
Winter	3.14	1.39	2.59	1.18	1.40	42
Spring	3.77	1.45	3.26	1.29	1.26	43
Time spent teaching least favorite lesson (0=less than 10m, 1=20m, 2=30m, 3=40m, 4=50m or more)						
Fall	2.21	1.40	2.80	1.44	− 1.30	37
Winter	2.86	1.39	2.50	1.22	0.90	41
Spring	3.45	1.30	3.23	1.41	0.56	42
Time spent referencing toolbox outside of lessons in a typical week (0=less than 10m, 1=20m, 2=30m, 3=40m, 4=50m or more)						
Fall	1.96	0.93	2.10	1.07	− 0.47	41
Winter	1.96	0.91	1.65	0.81	1.18	42
Spring	2.71	1.20	2.45	1.44	0.65	44
Quality indicators						
Quality of teaching lessons (Average; 1=F to 13=A+; e.g., 9=B)						
Fall	9.42	1.13	8.83	1.40	1.46	37
Winter	9.14	1.66	8.82	1.27	0.72	41
Spring†	9.50	1.50	8.78	1.37	1.68	43
Quality of teaching most favorite lesson (1=F to 13=A+; e.g., 10=B+)						
Fall	9.74	1.33	9.20	1.28	1.29	37
Winter	9.62	2.13	9.27	1.52	0.62	41
Spring	9.91	1.74	9.57	1.85	0.64	43
Quality of teaching least favorite lesson (1=F to 13=A+; e.g., 8=B-)						
Fall	9.11	1.20	8.45	1.85	1.31	37
Winter	8.45	1.28	8.36	1.43	0.21	40
Spring	9.09	1.72	8.00	1.51	2.27*	43
Quality of teaching while referencing tools outside of lessons (1=F to 13=A+)						
Fall	8.61	2.48	8.11	1.94	0.72	40
Winter	9.13	1.77	8.24	1.95	1.59	42
Spring	8.71	1.08	7.82	2.26	1.68	29.57[a]

M mean; *SD* standard deviation; *$p < 0.05$ (statistical significant difference between Standard and Primer conditions)

†Scale alpha coefficient < 0.60 (interpret with caution)

[a] Degrees of freedom (df) was adjusted using the Welch-Satterthwaite method to correct for the violation of homogeneity of variance assumption

abound. Many are guided by the *foot-in-the-door* compliance tactic (Freedman & Fraser, 1966), and the *diffusion theory* premise that less disruption is better for adoption (Rogers, 1995). Some program developers and administrators have, in turn, conceived that a lower-burden, flexible approach (e.g., a "primer") may prepare the implementation environment for the subsequent adoption of a more comprehensive ("standard") curriculum. Yet, a "flexible" approach may not necessarily be a lower burden approach. Given the complexity of integrating SEL into academic instruction, often in the absence of ongoing training and technical assistance, and with an array of competing mandates, additional research should consider whether diffuse directives and flexibility are the best way to promote wide-scale implementation in schools. This research should be longitudinal, as it may also be the case that different implementation directives will lead to more disparate implementation behaviors or student outcomes over time. It might also be helpful for programs to be analyzed in micro-randomized trials to understand the direct and interactive effects of their component parts to optimize for implementation and effectiveness (Collins et al., 2014).

Although it is now well established that *how* a practice is put into place shapes SEL program outcomes (Durlak et al., 2011; Rojas-Andrade & Bahamondes, 2019), we did not find evidence that the TOOLBOX *implementation*

Table 5 Difference in student SEC growth trajectory by implementation directive

	Model 1: no covariates			Model 2: covariates adjusted		
	Est	SE	*p*	Est	SE	*p*
Intercept regressed on						
Standard	−0.10	0.64	0.88	0.92	0.63	0.14
Girl				5.02	0.63	0.00
Age				0.68	0.33	0.04
Hispanic				0.05	0.53	0.93
Asian				3.20	0.78	0.00
Black				−3.99	0.85	0.00
White				0.62	0.90	0.49
ELL				−1.36	0.69	0.05
SPED				−8.92	1.07	0.00
FRL				−0.86	0.71	0.23
Slope regressed on						
Standard	0.01	0.61	0.99	0.06	0.64	0.93
Girl				1.16	0.63	0.07
Age				−0.64	0.35	0.07
Hispanic				0.84	0.53	0.11
Asian				2.02	0.78	0.01
Black				−0.83	0.82	0.31
White				−1.51	0.85	0.08
ELL				1.26	0.71	0.08
SPED				1.42	0.96	0.14
FRL				−0.82	0.70	0.24
Correlation						
Intercept with slope	−0.10	0.07	0.18	−0.10	0.08	0.21
Mean						
Intercept	50.64	0.46	0.00	45.47	2.11	0.00
Slope	3.12	0.43	0.00	5.82	2.36	0.01

ELL English language learner, *SPED* special education, *FRL* free and reduced priced lunch eligibility

directive ultimately shaped the growth of student SEC within the first year of implementation. This does not imply that actual implementation behavior does not shape the growth of student SEC, but rather, that the instructions and materials provided at the start of the academic year were not necessarily the most meaningful source of variance determining individual teacher implementation behaviors. Now that we have observed the relationship between the directives and student outcomes, future research should examine the direct relationship between teacher implementation behaviors and child outcomes through multi-level analysis. This analysis is beyond the scope of this paper, which sought to explore TOOLBOX effectiveness under different conditions created by school- or district-level adoption decisions.

The completion of this research project renders it no longer the case that students are receiving an SEL program (i.e., TOOLBOX) without the demonstration of growth relative to a comparison group, yet several limitations should be considered. In order for teacher ratings of student SEC to be completed, teachers needed to be familiar with students. This means that our initial assessment was in October, following an August TOOLBOX training, and approximately 1 month of instruction. We realize we may have missed some initial growth in SEC by nature of this limitation, but differences by condition were not detected at the time of our initial assessment. The quasi-experimental design (i.e., lack of random assignment to condition), detection bias (i.e., informants were likely aware of their assignment to condition and therefore potentially biased evaluators of their own work), and slight differences in missing rates by condition limit the potential for strong causal claims, but the "routine" practice conditions enhance the study's utility for informing practice-as-usual decisions. Although there were many students and teachers within each condition, the scant number of schools assigned to each condition, and their origins within a single district, limits the generalizability of our findings. On the other hand, the diverse student body, thoroughly described in this study, is a strength relative to much of the SEL literature (Rowe & Trickett, 2018).

Finally, it is important to recognize the limitations of self-reported implementation variables. Third-party observations of implementation behaviors have typically shown a stronger relationship to outcomes than self-report data (Lillehoj et al., 2004). Third-party observations, however, are more typical of free-standing, lesson-based SEL programs than SEL programs intended to be integrated, modeled, and applied throughout the school day (Shapiro et al., 2018). The direct observation of behavior is best suited for observing the frequency of high-prevalence, discrete behaviors, against their own baseline, for a minimum of five 30-min sessions to achieve reliable estimates (Doll & Elliott, 1994). The SEL-IS is designed to be a pragmatic alternative for monitoring implementation in routine practice, primarily for continuous quality improvement purposes. Its use for research is largely exploratory. For example, single-item indicators, used to enhance efficiency and problem-solving in practice settings, may contain more measurement error than is desired for research purposes, and some tactics used to normalize distributions of SEL-IS self-report data (e.g., how fully teachers are implementing TOOLBOX *compared to other colleagues*) could also obscure between-school comparisons. On the other hand, although many self-report efforts find teachers uniformly rating themselves favorably, the data from the SEL-IS were not particularly skewed.

Beyond these strengths and limitations, this study inspires an additional research direction. The finding that TOOLBOX was beneficial for student SEC growth, on average, and under various implementation directives, does not necessarily imply that it benefits all students equally. SEC growth trajectories may differ across diverse subgroups of students, and a universal SEL program like TOOLBOX may work differently for different students. A few prior studies have examined how student characteristics such as gender, race/ethnicity, and socioeconomic status relate to their SEC growth in general as well as in response to a universal SEL program (e.g., Frey et al., 2005; Holsen et al., 2009; Jones and Bouffard, 2012; Low et al., 2019; Malti et al., 2011). Nevertheless, current evidence is inadequate for understanding the extent to which socio-demographic disparities in SEC exist in student populations, and whether the benefits of a universal SEL program are distributed equally or differentially across diverse subgroups of students. Our analysis of covariates suggests that student SEC growth trajectories may vary by student characteristics such as age, gender, race/ethnicity, and special education services status, independent of the observed program effects. In order to understand the effectiveness of TOOLBOX, and in addition to randomized trials observing sustained effects, it is an important next step to determine whether the program contributes to educational equity or education disparities. Further research is needed to examine program effects on student social-emotional growth across diverse subgroups of students.

Practicing school psychologists may consider lessons that can be derived from this analysis for delivering social and emotional learning in the context of a Multi-Tiered System of Support (MTSS). In the California MTSS framework, for example, evidence-based practices are provided to all students to support whole child development, a feature of which is inclusive, transformative social and emotional instruction (Orange County Department of Education, 2021). It is therefore important for school psychologists, who may be engaged in decisions to select and adopt programs, to understand whether SEL programs like TOOLBOX have an evidence base and transformative potential. The California MTSS framework further asserts that in order to promote whole child development, there should be strong leadership, educator support, and organizational structures for integration, in the context of a positive school climate, and trusting partnerships with the district, families, and other community institutions. We discovered that implementation behaviors and student growth did not systemically vary based on the school-wide directive to adopt the standard or primer version of TOOLBOX. Instead, it is likely that a school psychologist who contributes to strong implementation leadership (e.g., serves on an inclusive SEL leadership team to develop a written implementation plan and remove implementation obstacles; Lee et al., 2018), and provides implementation support (through training, coaching, tools, and feedback loops; Wandersman et al., 2008), can more effectively shape the implementation behaviors associated with the optimal growth of student social and emotional competence.

Acknowledgements This project, and the associated data collection and analyses, was made possible by the generous support of the Stuart Foundation. The authors wish to acknowledge the teachers, staff, and administrators who participated in this project. The authors also wish to thank colleagues at the Center for Prevention Research in Social Welfare who played essential roles in the completion of this work. The authors appreciate consultation from Mark Collin and Chuck Fisher on the TOOLBOX program, and Paul LeBuffe, Jennifer Fleming Robitaille, Jessica Adamson, and Marc Kirsch on the DESSA-Mini. The opinions expressed in this article do not necessarily reflect the views of any funding agency, resource developer, or community partner.

Funding This work was supported by the Stuart Foundation. The Stuart Foundation did not determine the study design; collect, analyze, or interpret the data; participate in the writing of the report; or make the decision to submit the article for publication.

Declarations

Competing Interests Valerie B. Shapiro has been, but is not currently, a paid research consultant for Aperture Education, the current publisher of a tool used to measure outcomes in this manuscript. No author of this paper receives any direct financial remuneration from the sale of any of the assessments or resources mentioned within this manuscript.

References

Atlas, J. A. (2010). Test review of the Devereux Student Strengths Assessment. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook* (pp. 178–180). Buros Center for Testing.

Balfanz, R., & Byrnes, V. (2020). *Connecting social-emotional development, academic achievement, and on-track outcomes*. The Everyone Graduates Center at the Johns Hopkins University School of Education.

Chain, J., Shapiro, V. B., LeBuffe, P. A., & Bryson, A. M. (2017). Academic achievement of American Indian and Alaska native students: Does social-emotional competence reduce the impact of poverty? *American Indian and Alaska Native Mental Health Research, 24*(1), 1–29.

Collaborative for Academic, Social, and Emotional Learning. (2013). *CASEL guide: Effective social and emotional learning programs – Preschool and elementary* (school). Author.

Collaborative for Academic, Social, and Emotional Learning. (2015). *CASEL guide: Effective social and emotional learning programs – Middle and high* (school). Author.

Collaborative for the Advancement of Social and Emotional Learning. (2005). *Safe and sound: An educational leader's guide to evidence-based social and emotional learning (SEL) programs – Illinois edition*. ERIC Clearinghouse.

Collins, L. M., Dziak, J. J., Kugler, K. C., & Trail, J. B. (2014). Factorial experiments: Efficient tools for evaluation of intervention components. *American Journal of Preventive Medicine, 47*(4), 498–504.

Collins, T. A., Hawkins, R. O., & Nabors, L. A. (2016). Introduction to the special issue: Interventions to improve children's social and emotional functioning at school. *Behavior Modification, 40*(4), 487–492.

Collin, M. A. (2015). *TOOLBOX^{TM} primer*. Dovetail Learning, Inc.

Collin, M.A. (2014). TOOLBOX Project Administrator's Guide. Dovetail Learning, Inc.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science, 24*(2), 97–140.

De Long-Cotty, B. (2010). *Report on the TOOLBOX/Sonoma County Collaboration for Resilient Children.* Unpublished report.

District. (2016). *Data dashboard.* [Masked URL for District Privacy].

Doll, E., & Elliott, S. N. (1994). Consistency of observations of preschoolers' social behavior. *Journal of Early Intervention, 18*(2), 227–238.

Domitrovich, C. E., Syvertsen, A. K., & Calin, S. S. (2017). *Promoting social and emotional learning in the middle and high school years*. Pennsylvania State University.

Dovetail Learning. (2013). *West Contra Costa Unified School District (WCCUSD) TOOLBOX evaluation.* Unpublished report.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

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*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Dusenbury, L., Calin, S., Domitrovich, C., & Weissberg, R. P. (2015). What does evidence-based instruction in social and emotional learning actually look like in practice? A brief on findings from CASEL's program reviews. *Collaborative for Academic, Social, and Emotional Learning*. Retrieved from: https://eric.ed.gov/?id=ED574862

Epstein, M. H., & Sharma, J. (1998). *The Behavior and Emotional Rating Scale: A strength-based approach to assessment*. PRO-ED.

Epstein, M. H. (2004). *Behavioral and Emotional Rating Scale-2nd edition: A strengths-based approach to assessment*. PRO-ED.

Fagan, A. A., Hawkins, J. D., & Shapiro, V. B. (2015). Taking SEL to scale in schools: The role of community coalitions. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice*.

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology, 4*(2), 195. https://doi.org/10.1037/h0023552

Frey, K. S., Hirschstein, M. K., Snell, J. L., Edstrom, L. V. S., MacKenzie, E. P., & Broderick, C. J. (2005). Reducing playground bullying and supporting beliefs: An experimental trial of the steps to respect program. *Developmental Psychology, 41*(3), 479–490.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117–135. https://doi.org/10.1016/j.jsp.2006.05.005

Goldstein, S., & Naglieri, J. A. (2016). *Assessing impairment*. Springer.

Greenberg, M., & Weissberg, R. (2018). *Social and emotional development matters: Taking action now for future generations*. Pennsylvania State University.

Hair, J. K., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed.). Pearson.

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*(2), 51–60. https://doi.org/10.1080/1754730X.2009.9715704

Hwang, S.H.J., Shapiro, V.B., Eldeeb, N., Lee, J., Robitaille, J.L., & Naglieri, J.A. (2022). Assessing social and emotional competencies in educational settings: Supporting resilience in young people. In S. Goldstein & R. Brooks (Eds.), *Handbook of resilience in children* (2nd edition). New York, NY: Kluwer/Academic Press.

Jenkins, L. N., Demaray, M. K., Wren, N. S., Secord, S. M., Lyell, K. M., Magers, A. M., Setmeyer, A. J., Rodelo, C., Newcomb-McNeal, E., & Tennant, J. (2014). A critical review of five commonly used social-emotional and behavioral screeners for elementary or secondary schools. *Contemporary School Psychology, 18*(4), 241–254.

Jones, S. M., & Bouffard, S. M. (2012). Social and emotional learning in schools: From programs to strategies. Social Policy Report. Volume 26, Number 4. *Society for Research in Child Development*. http://eric.ed.gov/?id=ED540203

Kilgus, S. P., Reinke, W. M., & Jimerson, S. R. (2015). Understanding mental health intervention and assessment within a multi-tiered framework: Contemporary science, practice, and policy. *School Psychology Quarterly, 30*(2), 159–165.

Kusché, C. A., & Greenberg, M. T. (1994). *The PATHS curriculum: Promoting alternative thinking strategies*. Developmental Research & Programs.

LeBuffe, P. A., Shapiro, V. B., & Robitaille, J. L. (2018). The Devereux Student Strengths Assessment (DESSA) comprehensive system: Screening, assessing, planning, and monitoring. *Journal of Applied Developmental Psychology, 55*, 62–70.

Lee, J., Shapiro, V. B., Kim, B. K. E., & Yoo, J. P. (2018). Multilevel structural equation modeling (MSEM) for social work researchers: An introduction and an application to healthy youth development. *Journal of the Society for Social Work and Research, 9*(4), 689–719.

Lee, J., Shapiro, V. B., Robitaille, J. L., & LeBuffe, P. A. (2022a). *Measuring the development of social-emotional competence in elementary school years using behavioral rating scales*. Manuscript submitted for publication.

Lee, J., Shapiro, V. B., Robitaille, J. L., & LeBuffe, P. A. (2022b). *Gender, racial-ethnic, and socioeconomic disparities in the development of social-emotional competence among elementary school students*. Manuscript submitted for publication.

Lillehoj, C. J., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education & Behavior, 31*(2), 242–257. https://doi.org/10.1177/1090198103260514

Low, S., Smolkowski, K., Cook, C., & Desfosses, D. (2019). Two-year impact of a universal social-emotional learning curriculum: Group differences from developmentally sensitive trends over time. *Developmental Psychology, 55*(2), 415–433.

Mahoney, J. L., LeBuffe, P. A., Shapiro, V. B., Robitaille, J. L., Johnson, E. S., and Adamson, J. L. (2022). What is assessment bias and how is Aperture Education working to reduce it? (ed.1). Fort Mill, SC: Aperture Education. Retrieved from: www.ApertureEd.com.

Malcomb, K. K. (2010). Test review of the Devereux Student Strengths Assessment. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The eighteenth mental measurements yearbook* (pp. 180–182). Buros Center for Testing.

Malti, T., Ribeaud, D., & Eisner, M. P. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: A cluster randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology, 40*(5), 677–692.

Mayhew, M. J., & Simonoff, J. S. (2015). Non-White, no more: Effect coding as an alternative to dummy coding with implications for higher education researchers. *Journal of College Student Development, 56*(2), 170–175.

Merrell, K. W., & Gueldner, B. A. (2010). *Social and emotional learning in the classroom: Promoting mental health and academic success*. Guilford Press.

Mihalic, S. F., & Elliott, D. S. (2015). Evidence-based programs registry: Blueprints for healthy youth development. *Evaluation and Program Planning, 48*, 124–131.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide*. (Eighth edition). Muthén & Muthén.

Naglieri, J. A., LeBuffe, P., & Shapiro, V. B. (2011b). Universal screening for social–emotional competencies: A study of the reliability and validity of the DESSA-Mini. *Psychology in the Schools, 48*(7), 660–671.

Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. B. (2011a). *The Devereux Student Strengths Assessment—Mini (DESSA-Mini): Assessment, technical manual, and user's guide*. Apperson, Inc.

Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. B. (2013). Assessment of social-emotional competencies related to resilience. In S. Goldstein & R. B. Brooks (Eds.) *Handbook of resilience in children* (2nd edn., pp. 261–272). Springer. https://doi.org/10.1007/978-1-4614-3661-4_15

Nickerson, A. B., & Fishman, C. (2009). Convergent and divergent validity of the Devereux Student Strengths Assessment. *School Psychology Quarterly, 24*(1), 48–59.

Orange County Department of Education (2021). Revised California MTSS Framework. Retrieved from: https://www.link.ocde.us/camtss.

Rogers, E. M. (1995). *The diffusion of innovations*. Free Press.

Rowe, H. L., & Trickett, E. J. (2018). Student diversity representation and reporting in universal school-based social and emotional learning programs: Implications for generalizability. *Educational Psychology Review, 30*(2), 559–583.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.

Shapiro, V. B., & Lebuffe, P. A. (2006). Using protective factors in practice: Lessons learned about resilience from a study of children aged five to thirteen. *Annals of the New York Academy of Sciences, 1094*(1), 350–353. https://doi.org/10.1196/annals.1376.048

Shapiro, V. B., Kim, B. K. E., Accomazzo, S., & Roscoe, J. N. (2016). Predictors of rater bias in the assessment of social-emotional competence. *International Journal of Emotional Education, 8*(2), 25.

Shapiro, V. B., Accomazzo, S., & Robitaille, J. L. (2017a). In the same ballpark or a whole new ball game? Staff as raters of youth behavior. *Journal of Child and Family Studies, 26*(4), 1051–1055. https://doi.org/10.1007/s10826-016-0632-1

Shapiro, V. B., Kim, B. K., Robitaille, J. L., & LeBuffe, P. A. (2017b). Protective factor screening for prevention practice: Sensitivity and specificity of the DESSA-Mini. *School Psychology Quarterly, 32*(4), 449–464. https://doi.org/10.1037/spq0000181

Shapiro, V. B., Kim, B. E., Robitaille, J. L., LeBuffe, P. A., & Ziemer, K. L. (2018). Efficient implementation monitoring in routine prevention practice: A grand challenge for schools. *Journal of the Society for Social Work and Research, 9*(3), 377–394.

Shapiro, V. B., Ziemer, K. L., Accomazzo, S., & Kim, B. K. (2020). Teachers' assessment of "implementation leadership" during a new Social Emotional Learning Initiative. *Contemporary School Psychology, 24*(2), 174–180.

Simmons, C. A., Shapiro, V. B., Accomazzo, S., & Manthey, T. J. (2016). Strengths-based social work: A meta-theory to guide social work research and practice. In N. Coady & P. Lehmann (Eds.) *Theoretical perspectives for direct social work practice* (3rd edn., pp. 131–154). Springer.

Sklad, M., Diekstra, R., Ritter, M. D., 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*(9), 892–909.

Tanner, N., Eklund, K., Kilgus, S. P., & Johnson, A. H. (2018). Generalizability of universal screening measures for behavioral and emotional risk. *School Psychology Review, 47*(1), 3–17.

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development, 88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Wandersman, A., Duffy, J., Flaspohler, P., Noonan, R., Lubell, K., Stillman, S., et al. (2008). Bridging the gap between prevention research and practice: The interactive systems framework for dissemination and implementation. *American Journal of Community Psychology, 41*(3–4), 171–181.

Weston, K., Ott, M., & Rodger, S. (2018). Yet one more expectation for teachers. In Leschied, A., Saklofske, D., Flett, G. (eds), *Handbook of school-based mental health promotion* (pp. 105–126). Springer, Cham. https://doi.org/10.1007/978-3-319-89842-1_7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Valerie B. Shapiro is an Associate Professor jointly appointed in Social Welfare and Public Health at the University of California, Berkeley. Dr. Shapiro studies the promotion of well-being of children and youth through the implementation and continuous improvement of effective prevention practices, such as systematic and transformative Social and Emotional Learning in schools. Dr. Shapiro currently serves as Faculty Co-Director of Innovations for Youth (i4Y), the Chair of Coalition for the Promotion of Behavioral Health, and was recently selected as a William T. Grant Foundation Scholar. Dr. Shapiro is a member of the California Department of Education State SEL Advisory Group, the CalHOPE Student Support statewide leadership team, and is a certified school social worker.

Dr. Juyeon Lee is an Assistant Professor of the Department of Social Work and Social Administration at the University of Hong Kong. Her research concerns promoting social and emotional well-being of children and adolescents. Specifically, her studies have focused on examining (a) sociocultural factors associated with unequal social-emotional developmental trajectories and (b) the questions of whether and how school-based social and emotional learning interventions are both effective and equity-enhancing, when implemented in routine school settings. She has methodological expertise in various structural equation modeling approaches.

Bo-Kyung Elizabeth Kim PhD, MSW is an Assistant Professor at the University of Southern California Suzanne Dworak-Peck School of Social Work. She received a PhD in Social Welfare at the University of Washington and completed her postdoctoral fellowship at the University of Califorina, Berkeley. Her research focuses on informing service systems and evidence-based practice strategies as alternatives to youth incarceration to reduce mental, emotional, and behavioral health inequity experienced by youth of color.