Journal for Research in Mathematics Education
2020, Vol. 51, No. 3, 301–333

Effects on Mathematics and Executive Function of a Mathematics and Play Intervention Versus Mathematics Alone

Douglas H. Clements and Julie Sarama
University of Denver

Carolyn Layzer
Abt Associates

Fatih Unlu
RAND Corporation

Lily Fesler
Stanford University

Early education is replete with debates about "academic" versus "play" approaches. We evaluated 2 interventions, the *Building Blocks* (BB) mathematics curriculum and the BB synthesized with scaffolding of play to promote executive function (BBSEF), compared to a business-as-usual (BAU) control using a 3-armed cluster randomized trial with more than 1,000 children in 84 preschool classrooms across three districts (multiracial or multiethnic, low income, 27% English Language Learner). Impact estimates for BBSEF were mixed in sign, small in magnitude, and insignificant. Most impact estimates for BB were positive, but only a few were statistically significant, with more in the kindergarten year (delayed effects), including both mathematics achievement and executive function (EF) competencies. Gains in both mathematics and EF can be mutually supportive and thus resist the fade-out effect.

Keywords: Executive function; Geometry; Integrated curricula

Early childhood education is replete with debates about the role of content-focused or "academic" and "play-based" approaches (Chien et al., 2010). Many hold that these approaches stand in opposition, with a deleterious effect on children's learning of mathematics (Clements, et al., 2017), whereas others believe that they can be synergistically combined. To provide evidence on these

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A080200 and R305A080700. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. Although the research is concerned with theoretical issues, not particular curricula, components of the intervention used in this research have been published by the authors and their collaborators on the project, who thus could have a vested interest in the results. Researchers from an independent institution oversaw the research design, data collection, and analysis and confirmed findings and procedures. The authors wish to express appreciation to the school districts, teachers, and children who participated in this research and Carrie Germeroth, PhD, who contributed to previous versions of the research. Fatih Unlu and Lily Fesler were at Abt Associates when most of the reported research was conducted.

issues, we evaluated two preschool interventions, the *Building Blocks* (BB) mathematics curriculum based on learning trajectories (Clements & Sarama, 2007/2013) and BB synthesized with scaffolding of play to promote executive function (BBSEF). Executive function (EF) has been suggested as foundational to the learning of mathematics (Bull & Lee, 2014; Clements et al., 2016). Scaffolding of play to promote executive function (SEF) is the theoretical and operationalized pedagogical core of the *Tools of the Mind* (TotM; Bodrova & Leong, 2001) curriculum. We analyzed the effects of the two interventions on teachers' practice and on students' math achievement, EF, language, and literacy outcomes immediately at the end of preschool as well as the persistence of those effects at the end of kindergarten.

Background

Play-based preschool programs have a long history; however, recent concerns about children's achievement have set up a perceived conflict in which educators believe that they are being asked to abandon such approaches. Another perspective is that such approaches may be synergistically combined. To evaluate the latter, the authors of the BB and TotM curricula agreed to collaborate to study these conflicting viewpoints directly by producing and then testing a synthesized, theoretically based approach. The authors viewed such a synthesis as valid because of the compatibility and complementarity of their approaches, both theoretically (e.g., focus on children's agency and self-direction) and practically (use of scaffolding more than either laissez-faire or direct-instruction approaches).

BB (Clements & Sarama, 2007/2013) has produced positive effects on mathematics in rigorous evaluations (Clements & Sarama, 2007, 2008), including large-scale implementations across diverse settings (Clements, et al., 2011; Sarama et al., 2012) as well as improvements in oral language (Sarama, Lange, et al., 2012). Another evaluation reported moderate-to-large impacts on children's language, literacy, numeracy, and mathematics skills as well as small impacts on children's executive functioning and a measure of emotion recognition (Weiland & Yoshikawa, 2013); however, BB was one component of the PK program that also included an evaluated literacy component, and therefore any causal link between BB and effects on other domains such as EF is confounded. BB's basic approach entails finding the mathematics in, and developing mathematics from, children's activity following learning trajectories. Children are guided to extend and mathematize their everyday activities, from block building to art to songs to puzzles, through sequenced, intentional activities.

For the play component, the TotM authors chose to implement the theoretical and operationalized core of their approach—scaffolding dramatic and make-believe play—because they believed that these components develop children's EF skills in a way that content-oriented teaching may not (Bodrova et al., 2013; Golinkoff et al., 2006). SEF includes supports for planning, articulation, and especially maintenance of dramatic roles throughout an extended time frame, putting considerable demands on, and thus developing, children's EF processes (Bodrova & Leong, 2006, 2007b). At the time that we planned this study, research indicated that the TotM program supported the development of specific EF skills

(Barnett et al., 2008; A. Diamond et al., 2007). After we began, some research showed few such effects (Farran, Lipsey, & Wilson, 2011; Morris et al., 2014), although results from other evaluations were more promising (Blair & Raver, 2014).

Further supporting the synthesized approach, several studies suggested that SEF can improve mathematics learning (Clements et al., 2016). Indeed, in one sample of classrooms using such strategies, children did better than those in comparison classrooms that did not use SEF strategies on math tests without changes in the content of the curriculum, which was focused on literacy and EF (Barnett et al., 2006). Thus, EF may allow children to use and further develop cognitive processes necessary for academic learning.

The authors of TotM synthesized these two approaches in two ways. First, throughout the day, especially during "free play" (children choose their activities) periods, teachers in the BBSEF condition were taught to use scaffolding strategies that support mature intentional play. They also were encouraged to incorporate mathematical ideas into their play contexts and interactions. Second, large-group, small-group, and transition activities were altered as necessary to avoid situations that might negatively affect EF (e.g., limiting long periods of whole-group activities dominated by teacher's talk) and include scaffolding strategies designed to support the use of private speech or the use of external visual and auditory aids to support children's ability to focus or follow directions. There was some empirical support for such a synthesized approach. For example, some report that curricula designed to both improve EF and enhance early academic abilities are most effective in helping children succeed in school (e.g., Blair & Razza, 2007). Further, young children's EF scores correlate with both concurrent and future mathematics achievement scores even more strongly than other attributes such as IQ (Best et al., 2011; Blair et al., 2011; Blair & Razza, 2007; Clements et al., 2016; Neuenschwander et al., 2012). Some studies show that EF is more highly associated with mathematics than literacy or language (Blair et al., 2011; McClelland et al., 2014). However, there is little research that investigates the foundation of these abilities and analyzes cause-and-effect relationships among specific components of these abilities (Clements et al., 2016; Jacob & Parkinson, 2015).

Our main hypotheses were that the synthesized (BBSEF) approach would increase children's EF and support greater mathematics achievement than either the mathematics-curriculum-only (BB) approach or the business-as-usual (BAU) approach and that the BB approach would support greater mathematics achievement gains than BAU. Moreover, we posited that neither experimental approach would come at the cost of achievement in other areas, such as early language and literacy.

Research Design

To test our hypotheses, we conducted a three-armed cluster randomized control trial in which classrooms in participating schools or early childhood centers were randomly assigned to the study conditions (BB, BBSEF, and BAU). Random assignment was conducted separately for schools or centers with only one participating classroom (Group A) and those with two classrooms (Group B). Classrooms

in Group A were placed into five randomization blocks such that each block consisted of all half-day or full-day PK classrooms in each study district (one district had only full day). Within each block, schools or centers were sorted with respect to categories created based on prior math achievement: percent eligible for free or reduced-price lunch and percent English learners (school level).[1] Schools or centers were randomly assigned to the three conditions using the systematic circular sampling scheme (Lahiri, 1951), ensuring that the groups were balanced in geography, length of the program, and key background characteristics. For Group B schools or centers, random assignment was conducted within each school or center, where the two classrooms were randomly assigned to two conditions that were determined randomly. We assessed the balance by examining the characteristics of schools, classrooms, teachers, and students that were primarily obtained through teacher surveys conducted before random assignment.[2] Table 1 shows the averages of these characteristics in each group and the *p*-value from hypothesis tests (joint *F*-tests that also accounted for the randomization blocks) that assessed the statistical significance of the differences across the three groups. All the differences in Table 1 are small and not statistically significant, suggesting that groups were statistically equivalent at baseline.

Research Questions

This design allowed us to answer three main research questions, each a cluster of several components.

Research question 1: Can the two interventions be implemented with high fidelity and have substantial positive effects on teachers' practice? Prior research suggests that most teachers can implement each of the BB components (Clements & Sarama, 2007) and, separately, the SEF component (Bodrova & Leong, 2005; Morris et al., 2014) with acceptable fidelity and can make significant gains in knowledge and practice if provided adequate professional knowledge and support. Our hypothesis was that all aspects of the synthesized intervention could similarly be adequately implemented.

Research question 2: What are the immediate effects of the two interventions (BB and BBSEF), as implemented under diverse conditions, on children's achievement and EF? Assuming adequate fidelity of implementation, our hypothesis was that children in both intervention groups would outperform children from the BAU control classrooms in mathematics, with no significant differences in measures of language and literacy (thus, the time spent on the

[1] Five categories were created for each of the three baseline characteristics (math achievement, % free or reduced price lunch, and % English learners) and sorting of the schools within each block was conducted for the resulting three categorical variables in the specified order. Using continuous variables in the sorting would have given the most weight to the first variable. This alternative to matching schools or centers prior to randomization was preferred because the use of continuous variables would have decreased the degrees of freedom for the analyses and statistical power.

[2] We intended to collect baseline measures of students' math achievement and teachers' classroom practices through student assessments and classroom observations. Problems with gaining access to classrooms delayed collection of these data, making them potentially contaminated. Therefore, we did not use them to assess the equivalence of the groups at baseline.

Table 1
Baseline Characteristics of Schools, Classrooms, and Teachers.

	BB ($N = 25$)	BBSEF ($N = 30$)	BAU ($N = 29$)	*p*-value for joint test of significance
Locale				
Urban	52%	52%	63%	.56
Suburban	34%	29%	22%	
Rural	7%	6%	6%	
Classroom characteristics				
Class size	24	22.5	24.3	.27
Children with IEPs	1.5	1	1.5	.31
English learners	11.5	14	13.8	.15
Teacher characteristics				
Associate's degree	14%	39%	38%	.10
Bachelor's degree	72%	52%	53%	
Master's degree	14%	6%	9%	
Year of teaching experience	12	11.2	13.2	.70
Years taught in current school	4.6	6.1	6.1	.34
CDA credential	17%	32%	19%	.10
ECE credential	66%	58%	53%	.37
ETC credential	10%	6%	9%	.47

Note. All measures are obtained from teacher surveys conducted prior to random assignment. *P*-values were obtained from an omnibus *F*-test that assesses the significance of the differences in the group means for each characteristic.

interventions would facilitate language growth at least as much as the practices of the BAU group who may devote more time to these topics). In addition, we hypothesized that the children in the BBSEF group would outperform the other two groups on measures of mathematics and EF.

Research question 3: What are the longer term effects of the two interventions? Similar to Research Question 2, we hypothesized that children in both experimental groups would outperform those in the BAU group in math achievement at the end of kindergarten and that the BBSEF group would outperform children in the BB group as well.

Interventions

The two interventions used one or both of the two theoretically and empirically grounded components. The first intervention was the implementation of just the first, the BB curriculum. The second component, SEF, was synthesized with BB to form the second intervention (BBSEF).

BB Component

Math intervention. The BB curriculum (Clements & Sarama, 2007/2013) is an early math curriculum based on a theoretical research and development framework (Clements, 2007). The curriculum is structured around empirically based learning trajectories and includes whole-group (10 minutes per day), small-group, and computer activities (both about 10 minutes per week) as well as learning centers and ideas integrating mathematics throughout the school day. For example, on and off computer, children play board games corresponding to the developmental levels along the BB learning trajectories on subitizing and counting leading to arithmetic. That is, children might use a single cube with only one, two, and three dots, then one with one to six dots, then one with five to 10 dots, and finally two cubes with dots or numerals that they have to add. Teachers model games in the whole group, work with two to three pairs of children in small groups (adjusting the level as necessary for individuals), and encourage children to play the game in learning centers.

BB professional development. Teacher professional development was designed to develop teachers' learning of all three parts of the learning trajectories: goal, developmental progression of levels of thinking, and instructional activities designed to build the mental actions-on-objects that enable thinking at each higher level (Clements & Sarama, 2014; Sarama & Clements, 2009). First, the sessions develop teachers' content knowledge by explicating the mathematical concepts, principles, and processes involved in each level and the relationships across levels and topics. For example, sessions on geometry began by exploring the components of geometric shapes, including a correct definition of *side*. Teachers then learned about relationships between components, such as sides forming a right angle. Finally, they used such attributes to describe shape categories and relationships between categories, such as squares as a subcategory of rectangles (Clements et al., 2011). Second, the sessions increase teachers' knowledge of students' developmental progressions in learning that content (for example, moving from intuitively recognizing shapes as unanalyzed visual wholes, to recognizing components of shapes, to hierarchically classifying shape categories). Learning in these two areas encourages and enables teachers to engage children in more challenging mathematical activities. Third, the sessions enhance teachers' knowledge of the instructional activities designed to teach children the content and processes defining the level of thinking in the developmental progressions and to inform teachers of the rationale for the instructional design of each activity (e.g., why certain length sticks are provided to children with the challenge to build specific shapes). Knowledge of these learning trajectories supports curriculum enactment with fidelity in that the learning trajectories connect the developmental progressions to the instructional tasks, providing multiple guidelines and sources of stability in teachers' implementation of the instructional activities. Finally, BB's learning trajectories are designed to motivate and support the use of formative assessment.

The professional development had two components: (a) training for each of the 2 years of their involvement, with 2 days of training during the first month of

school in Year 1, 2 days during the school day in the fall, and 2 days during the spring (with the project paying for substitutes) and (b) coaching within each teacher's classroom (similar to that used in Clements & Sarama, 2007, 2008 and Clements et al., 2011, except that coaches were from the district and not provided by the project, as described in a later section). Training included the following topics: learning trajectories for each math topic, using learning trajectories for formative assessment, recognizing and supporting math throughout the day, setting up math learning centers, teaching with computers, small-group activities, and supporting mathematical development in the home. A main tool was the Building Blocks Learning Trajectory (BBLT) web application, providing descriptions, videos, and commentaries (Clements & Sarama, 2007, 2008; Clements et al., 2011; Sarama & Clements, 2013).

The first year was a pilot and training year, given that it can require at least 2 years for teachers to begin implementing a program with fidelity (Berends et al., 2001; Campbell & Silver, 1999; Cobb et al., 2003; Heck et al., 2002; Weiss, 2002). In the second year, teachers continued to work on BBLT, and they brought case studies of particular situations that occurred in their classrooms to their training groups (Gallimore & Stigler, 2003).

SEF Component

SEF. The TotM authors focused on the core of the curriculum (Bodrova & Leong, 2001), promoting EF via pedagogical strategies that optimize the benefits of mature, intentional dramatic play (Bodrova & Leong, 2007a). Further, SEF in nonplay activities is accomplished by redesigning the social context for these activities as well as by teaching children to use specific EF "tools" that assist them in managing their own behaviors (Bodrova & Leong, 2007b). Teachers were taught two types of strategies for SEF.

The first and major type included strategies that focus on supporting intentional and mature dramatic play. These include (a) using toys and props in a symbolic way, (b) developing consistent and extended-play scenarios, (c) taking on and staying in a pretend role for an extended-play episode or a series of play episodes, and (d) consistently following the rules determining what each pretend character can or cannot do.

The second type included strategies that support various aspects of EF in a more focused and specific way. The latter strategies were implemented throughout the day both in the context of teaching the BB curriculum and in the context of other activities. The BB curriculum remained intact but was supplemented by these strategies. For example, BB already emphasizes involvement of all children during whole-group lessons, but the reason that having pairs of children talk about their solutions benefits learning of both math and EF was explicitly discussed with teachers, along with more specific strategies for supporting that synergistic learning. One such strategy was the use of visual aids depicting the sequence of steps in a board game and the role each child takes when working on an activity. Such SEF strategies were used by teachers for all large-group and small-group activities, including math and other content areas. Table A1 in Appendix A

provides a small sample of the strategies used and a comparison of the three experimental conditions.

The intervention addressed some EF strategies directly: for example, children's learning and application of skills in an authentic environment by practicing sustaining attention and inhibiting "first impulse" responses, switching attention, focusing attention on specific attributes in authentic contexts such as play and movement games, and building working memory. Specifically, children were encouraged to pay attention to their physical actions (e.g., movement games in which a series of directions are to be carried out only if the phrase "Simon says" precedes them), verbal behaviors (e.g., word play and riddles), and dramatic play (e.g., introducing multifunctional props that change their function repeatedly).

SEF professional development. The content of professional development for the BBSEF group included the same training that was in BB as described previously plus additional professional development on the SEF component. The SEF training, delivered by authors of TotM (Bodrova & Leong, 2007b) and their colleagues, included an additional 6 days of training in each of the 2 years of the teachers' involvement. Trainers met with the classroom teachers following the same schedule as the BB group. The SEF training followed the same general organizational structure as the BB training and included the topics of development of EF in early childhood, how dramatic play supports EF, and how teachers can scaffold mature and intentional dramatic play. Like BB professional development, SEF professional development combined building teachers' knowledge of young children's learning and development with helping teachers master effective instructional strategies designed to support this learning and development. Focus was on mature dramatic play as a critical component in promoting EF and on the need to scaffold such play; videotapes illustrating various stages in play development as well as best practices of scaffolding play were shown and discussed. In addition to SEF-specific training, the BBSEF teachers received training on modified BB instructional strategies redesigned to maximally promote EF. Further, the EF strategies added to the math curriculum combined focal activities implemented both inside and outside of existing math activities.

Coaching. The coaches were those already working in each district, and they participated in the same professional development as the teachers for both the BB curriculum and the SEF. They also participated in a half day of professional development on coaching by project staff and participated in on-site coaching support for hour-long, biweekly sessions. Coaches provided teachers in the treatment groups with feedback using a structured observation form (Germeroth & Sarama, 2017). Coaches also provided off-site coaching support, being available to teachers and research coordinators via email, phone, or fax. Because coaches provided support to teachers across the conditions, one focus of their training was maintaining fidelity to the condition in all support provided. This was monitored through joint classroom visits by a coach and the lead coach, an experienced local coach who coordinated and supervised the coaches.

BAU Control

The BAU control used standard district practices and curricula. Standard practices in all three participating districts included implementation of well-regarded published curricula: The two districts contributing the most classrooms used *Developing Math Concepts in Pre-Kindergarten* by Kathy Richardson (2008; http://mathperspectives.com) and the third used *Everyday Mathematics* (University of Chicago School Mathematics Project, 1995/1997). All teachers received extensive training on these curricula prior to the first year of this study—for example, by viewing videos of children, talking about their mathematical thinking and sense making, and implementing the curriculum (Richardson, 2008). In summary, BAU in these districts was substantially different from that in the control classrooms of previous evaluations of BB (Clements & Sarama, 2007, 2008) and almost all other evaluations of early curricula (e.g., Lewis et al., 2015; Preschool Curriculum Evaluation Research Consortium, 2008) because they included mathematics curricula that shared many characteristics of the BB curriculum (research based; commitment to meaningful, conceptual learning; extensive, sequential activities), and the teachers had received substantial professional development in implementing these curricula that also shared characteristics of the professional development provided in support of BB and BBSEF (including creating a positive environment, formative assessment, and promoting ways to help children develop understanding and skills).

The evaluation thus had three conditions: BB as it has been implemented in previous research, BB enhanced with SEF (BBSEF), and a BAU control using standard district practices and curricula. All three districts employed instructional coaches for mathematics, and the same coaches within each district worked with teachers in up to three of the study conditions. The time during which coaches engaged with teachers was unchanged, with the coaching for BB and BBSEF replacing some BAU coaching foci. In summary, comparisons among the conditions, including the unusually strong counterfactual, constituted a rigorous and conservative evaluation of the introduction of a math curriculum and professional development based on learning trajectories (BB) and this in combination with an approach to scaffolding mature play designed to support children's development of EF.

Sample

Our analytical sample varies across time points and outcome measures. Table 2 depicts the number of classrooms and students with valid data for at least one outcome measure. A large proportion of the sample are multiracial or multiethnic Hispanic children—the majority minority at 39%, Asian Pacific Islander at 18%, African American at 11%, and non-Hispanic White at 31%. On average, 27% of the students are English Language Learners (roughly 20% of the total group have Spanish as the primary language).

Measures

Child Outcomes

Early mathematics. We used two instruments to measure early mathematics achievement, the Tools for Early Assessment of Mathematics (TEAM; Clements

Table 2
Size of the Analytical Sample in Fall 2010, Spring 2011, and Spring 2012

	BB		BBSEF		BAU	
	Number of classrooms	Number of students	Number of classrooms	Number of students	Number of classrooms	Number of students
Fall 2010	25	329	30	391	29	365
Spring 2011	24	264	28	298	28	275
Spring 2012	25	301	30	354	29	298

Note. The number of classrooms presented in this table is based on the initial PK classrooms in which students were enrolled at the time of random assignment.

et al., 2008; Clements & Sarama 2011) and a Spanish-language version administered to those children identified by their teachers as English Language Learners (<5%). TEAM is a measure of preschool children's mathematical knowledge and skills that features two individual interviews of each child, with explicit protocol, coding, and scoring procedures. Both videotapes and the TEAM record forms were evaluated by trained coders who were naïve to the group assignment of the child. Assessments were evaluated for item accuracy as well as item solution strategies and error type. Concurrent validity was initially established with a .86 correlation with a separate research-based instrument, and there was a .89 correlation with the Woodcock Johnson III in pilot testing (Woodcock et al., 2001). The assessment was refined in three pilot tests and a Rasch model analysis was computed, yielding a reliability of .94 for a similar population of children (Clements et al., 2008). The use of the Rasch model provides strong inference that the measured behaviors are expressions of the underlying construct of mathematics ability, supporting the instrument's construct validity (Clements et al., 2008).

The second measure of early mathematics competencies is the mathematics section of the direct cognitive child assessment used in the Early Childhood Longitudinal Study-Birth (ECLS-B; http://nces.ed.gov/ecls/birth.asp). This study tracks physical, cognitive, language, social, and emotional development of a representative sample of 14,000 children in the United States from 9 months through kindergarten. The mathematics test from the ECLS-B measures proficiencies relevant to preschool-age children, including number sense, counting, operations, geometry, and patterns, and comprises items from the direct cognitive assessment from other psychometrically validated assessments.

EF. We tested inhibitory control (Head-Toes-Knees-Shoulders, Peg Tapping), working memory (Backward Digit Span), and phonological processing (Forward Digit Span). We considered aggregating these measures into one or more constructs but decided to analyze them separately because they assessed different skills and the correlations between the individual measures were low (ranging between .11 and.39).

Head-Toes-Knees-Shoulders (HTKS) task. In this task, which requires inhibitory control, attention, and working memory (though inhibitory control is the main focus), children are asked to play a game in which they must do the

"opposite" of what the experimenter says (McClelland et al., 2014). For example, the experimenter could instruct children to touch their head, but instead of following the command, the children are supposed to do the "opposite" and touch their toes. In subsequent trials, commands to touch shoulders and knees are added, following the same rule that when the experimenter instructs the child to touch his shoulders, he should touch his knees, and when told to touch his knees, he should touch his shoulders. The analysis of 12 EF measures showed the HTKS to significantly predict achievement gain from the beginning of PK to the end of kindergarten and identified it as one of the top performing measures, which supports its validity (Lipsey et al., 2017). Coefficient alpha in this evaluation ranged from .85 to .96, and test-retest reliability was .78 for PK and .93 for kindergarten.

Forward digit and backward digit span. The children are told that they will hear some numbers and they will first repeat the numbers back to the examiner in the same order in which they were presented, and then later they are asked to repeat a different sequence of numbers in the reverse order to that in which they were presented. Difficulty increases by increasing the span of the pattern. These are considered measures of updating working memory (Lipsey et al., 2017). The forward digit span is considered to measure short-term auditory memory, the phonological processing component of memory. Backward digit span measures the ability to manipulate verbal information while in working memory. Coefficient alphas for this study's sample were .74 for PK and .78 for kindergarten for forward and .73 and .77 for backward.

Peg Tapping. The Peg Tapping task has been normed and widely used to measure EF and, more specifically, inhibitory control. Students are asked to tap a peg on a desk either once or twice after watching the assessor tap. The student must tap once if the assessor taps twice and tap twice if the assessor taps once. A student must attend to the instructions and his or her response while inhibiting the desire to tap the same number of times as the assessor. The test is individually administered and takes approximately 5–8 minutes, depending on the ability of the child. Peg Tapping has been shown to significantly predict achievement gain on every outcome (Lipsey et al., 2017). Coefficient alphas in this evaluation ranged from .87 to .88, and test–retest reliability was .80 for PK and .81 for kindergarten.

Language and literacy: Oral language. We used one measure of oral language at baseline and two at the two posttest time points. At baseline, we used the Peabody Picture Vocabulary Test, 3rd Edition (PPVT-III; Dunn, 1997), which is a widely used test of receptive vocabulary. Although there are two parallel forms, only one form (A) was used as we planned to administer it as a baseline measure only. Internal consistency (Cronbach's alpha) ranges from .92 to .98 (median: .95) and split-half reliability ranges from .86 to .97 (median: .94). At follow-up, we used two expressive oral language measures—the Expressive Vocabulary Test (EVT-2; Williams, 2007) and the Renfrew Bus Story (RBS; Glasgow & Cowley, 1994)—which are described below.

The second edition of the EVT-2, Pearson Education, often used in conjunction with the PPVT-III, offers a test of expressive vocabulary. Split-half reliabilities range from .83 to .97 with a median of .91. Alphas range from .90 to .98 with a median of .95. Test–retest studies with four separate age samples resulted in reliability coefficients ranging from .77 to .90, indicating a strong degree of test score stability. Children's oral expressiveness was measured to determine whether the intervention had any effect on these skills of young children. The current study hypothesized that there would be no negative effects on these skills, but another possibility was that the BB curriculum alone, or the combined condition, might have a positive impact on these abilities because of the interactive and metacognitive nature of both curricula.

The RBS (Glasgow & Cowley, 1994), a standardized measure of oral language using narrative retell, was used to evaluate children's oral language. The assessment involves telling a child a story and then asking the child to retell the story using the pictures in the wordless storybook as prompts. At the end of the story, assessors asked children an inferential question. To score well on this measure, children must remember key concepts (memory), know the meaning of the words representing the concepts well enough to use them appropriately in their retell (vocabulary), and have a sufficient understanding of story structure to use the words or concepts in the right sequence (book or story knowledge). Previous research has demonstrated strong predictive relationships with literacy and language skills 3 years after initial assessment (Pankratz et al., 2007). The total raw score, with a maximum possible score of 52, is then converted to a standardized score.

Language and literacy: Emergent literacy. We included an emergent literacy measure to test whether attention devoted to math or SEF would affect literacy outcomes. The Alphabet Knowledge subtest from the Phonological Awareness Literacy Screening (PALS; Invernizzi et al., 2004) assesses children's knowledge of the alphabetic code (alphabet recognition and phonemic awareness) and consists of three parts: Upper-Case Alphabet Recognition, Lower-Case Alphabet Recognition, and Letter Sounds. Lowercase alphabet recognition is only assessed if a threshold number of uppercase letters is identified (16 or more uppercase letters), and letter sounds are only assessed if a threshold is met in lowercase letter identification (nine or more lowercase letters). Coefficient alphas for this study's population were .75 for PK and .79 for kindergarten.

After training, all assessors were required to pass certifications prior to being allowed to conduct the assessments. Each assessor's first assessment in the field comprised the final certification and was observed by a trainer, and feedback provided included correctness of administration, accuracy of recordkeeping, and rapport and appropriateness of behavior with child and with school and center staff. Only assessors with no errors in administration that would have compromised the quality of the data were certified.

Classroom Observations and Surveys

Classroom observations were conducted to rate general quality and to assess fidelity to the two components of the interventions. Teachers were surveyed to

assess baseline equivalence across the three groups and to understand intervention impacts on teachers' knowledge, attitudes, beliefs, and self-reported practices.

Classroom quality. The Classroom Assessment Scoring System (CLASS) Pre-K is an observational assessment of classroom quality in preschool. The 10 dimensions measured by the CLASS Pre-K focus on the quality of teachers' emotional, organizational, and instructional interactions with students in the classroom. This measure allows comparisons across treatment groups of quality of the classroom on dimensions such as positive emotional climate, behavior management, and the degree to which teachers promote higher order thinking. These dimensions represent some of the primary mediators that the EF components of the proposed curriculum aimed to improve. The CLASS Pre-K has been validated in several large studies, one of which found the emotional and instructional support factor scores to be correlated with another classroom observation scale, the Early Childhood Environment Rating Scale (ECERS; Harms et al., 1998), total score, .52, $p < .0001$, and .40, $p < .0001$ (La Paro et al., 2004). Studies have demonstrated that children make more academic progress in classrooms characterized by positive and sensitive interactions among peers and teachers, effective organization of time and behavior, and consistent instructional feedback and support of higher level cognition (La Paro et al., 2004; Pianta et al., 2008). A minimum of 2 hours is required for administration and coding per class observed. Observation is required for at least four periods of 20 minutes throughout the school day, observing in each activity except outside recess, with 10 minutes of coding after each 20-minute period. Interrater reliability for the CLASS, computed via simultaneous classroom visits by pairs of observers (10% of all observations, with pair memberships rotated), was 90%.[3]

Fidelity to the SEF approach. The Mature Play Observation Tool (MPOT; Germeroth et al., 2019) was developed to measure the fidelity of implementation of the SEF intervention but was also used across conditions for measuring the extent to which the setting and caregivers support make-believe play and the extent to which mature play is taking place in the setting. The measure includes a Play Routine Checklist consisting of eight items: Dimension 1: Child Actions scale (five items) and Dimension 2: Teacher Actions (three items). The observation is conducted over 3 or 4 hours of class time during which observers note evidence of each item. At the end of the observation period, observers assign a rating (1 to 4) on each item using the provided rubric and basing ratings on the evidence noted. The MPOT yields three scores (one for each of the scales above). Interrater reliability for the MPOT, computed via simultaneous classroom visits by pairs of observers (10% of all observations, with pair memberships rotated), was 97%.

Classroom observation of math. The Classroom Observation of Early Mathematics—Environment and Teaching (COEMET) was used to measure the

[3] The procedure for determining observer reliability followed the CLASS published guidelines, with the percentage agreement exceeding 80% across all items and dimension-level agreement on at least three of five segments coded. The 90% cited is agreement across all items.

quantity and quality of mathematics in the classrooms. Based on a body of research on the characteristics and teaching strategies of effective teachers of early childhood mathematics (Sarama & Clements, 2019), the COEMET measures the quality of the mathematics environment and activities with an observation of 3 or more hours, similar to the MPOT. It allows for intervention-control condition contrasts. However, because the research bases for the COEMET and BB were similar, it can also be used to indicate the degree of fidelity to the BB curriculum, similar to the MPOT. There are 31 items, all but four of which are 5-point Likert scales. An example of one of the three items in the section "Personal Attributes of the Teacher" is: "The teacher appeared to be knowledgeable and confident about mathematics (i.e., demonstrated accurate knowledge of mathematical ideas and procedures, demonstrated knowledge of connections between, or sequences of, mathematical ideas)." Observers spend no less than a half day in the classroom—for example, from before the children arrive until the end of the half day (e.g., until lunch). All mathematics activities are observed and evaluated without reference to any printed curriculum. The COEMET has three main sections: classroom elements, classroom culture, and specific mathematics activities (SMAs). Assessors complete the first two sections once to reflect their entire observation. They complete an SMA form for each observed mathematics activity. A mathematics activity is defined as one conducted intentionally by the teacher involving several interactions with one or more children or one set up to develop mathematics knowledge (this would not include, for instance, a single, informal comment). Interrater reliability for the COEMET, computed via simultaneous classroom visits by pairs of observers (10% of all observations, with pair memberships rotated), was 86%; 99% of the disagreements were the same polarity (i.e., if one was agree, the other was strongly agree). Coefficient alpha (inter-item correlations) for the COEMET ranges from .95 to .97 (Clements & Sarama, 2008; Clements et al., 2011). The maximum possible scores for each Likert-based subtest are as follows: classroom culture mean score, 5; SMA mean score, 5.

One set of observers was trained and certified in the CLASS in a 3-day intensive session, with the first 2 days following the official CLASS training package and the third day consisting of a field practice in a local preschool classroom followed by interrater reliability tests and debriefing. Another set of observers was trained in the COEMET in a 2-day training session that included in-depth explanation of each item, field practice in a local preschool classroom, and debriefing. Interrater reliability was calculated during the first observation using the trainers as the standard and at two additional points in each round of data collection using more experienced "gold standard" observers.[4] Training on the MPOT was conducted by the developer of the measure with the assistance of the co-PI. In addition to explanation of each item's definitions and rubric, videotaped examples were provided to assist observers in understanding the key concepts. A field practice followed by debriefing was the final part of this training as well. The field practice was conducted in nonstudy classrooms, and debriefing included a focus on

[4] "Gold standard" experienced observers were those who had achieved the highest level of interrater reliability in the observation measures.

identifying behaviors that were "treatment-like"—in other words, behaviors such as planning for play (which is a common practice in Head Start classrooms), use of props in play, role-playing behaviors, and child communication during play—and how to rate what the observers witnessed. The trainer emphasized that it would be possible to observe treatment-like behaviors in any classroom.

Teacher attitudes, beliefs, and self-reported practices. Teachers were surveyed twice, early in their first implementation year (Fall 2009) prior to the training sessions and at the end of the second year of implementation (Spring 2011). The first survey was about perceptions of teaching mathematics and the environment. The second survey was distributed to teachers by mail and by hand as well as with the assistance of the district's instructional coaches; they were returned by mail and by a private carrier. The first survey included 53 questions, 12 of which used a Likert-scale response style and four of which allowed for open responses. The second survey included the same questions as the first (aside from teacher demographic questions), and a series of 10 items was added to the version of the second survey that was only administered to the BBSEF group to gather information about their perceptions of implementation of that intervention and perceived outcomes for their students.

Data Collection

Teachers' classroom instructional behaviors were measured through surveys and classroom observations, respectively, at two points in time—after random assignment and again after the end of 2 years of implementation (the first year of implementation was a pilot or training year—child data were collected starting in the second year). Observations of mathematics teaching in the classroom were also conducted at the end of the first school year of implementation. The baseline observations were conducted later than intended and thus reveal what could be early treatment effects in either intervention condition. Child assessments were conducted at two points in the second-implementation PK year—one intended to be at baseline and the second at the end of the PK year—and again at the end of kindergarten, using measures of early mathematics and EF. Problems with gaining access to classrooms delayed child assessments at the baseline time point, making assessing the similarity or differences between the three groups at baseline challenging.

Schedule and procedures for the collection of achievement measures. Children were individually assessed by trained and certified assessors in two 45-minute sessions after roughly one month at the start of the preschool year (to capture children's abilities at baseline), near the end of the preschool year (to capture children's abilities immediately following treatment and at a comparable time for nontreatment), and at the end of the kindergarten year (to capture possible longer term differences in children's abilities). Table 3 illustrates the data collection schedule for specific child-assessment measures.

Schedule and procedures for the collection of classroom observation and survey data. Trained observers conducted observations in all participating

Table 3
Data Collection Schedule: Child Assessments

	Year 1		Year 2		Year 3	
Child assessment measures	Fall 2009	Spring 2009	Fall 2010	Spring 2011	Fall 2011	Spring 2012
Mathematics						
TEAM			X	X		X
ECLS-B Pre-K Math				X		
ECLS-B K Math						X
Self-regulation/executive function						
HTKS			X	X		X
Peg tapping			X	X		X
Forward digit span			X	X		X
Backward digit span			X	X		X
Language and literacy						
PPVT-III			X			X
EVT-2				X		X
RBS				X		X
PALS—alphabet knowledge			X	X		

preschool classrooms during the instructional time[5] at three time points in the first and second years of training and implementation. The first observation point was intended to be prior to the training of the treatment group teachers, during the fall of the first implementation year[6]; the second observation point was in spring of the same year; and the third observation point was in the spring of the second implementation year (Year 3 involved no implementation but only child data collection at the end of their kindergarten year). Table 4 illustrates the data collection schedule for teacher and classroom measures.

Analytical Strategies and Results

This section describes the analyses conducted to address the research questions. Specifically, we first examine the analytical implications of the delay in the collection of baseline classroom and student measures. Next, we assess the fidelity of implementation and impacts of the two interventions on the teacher practices. We then examine the impact of the two interventions on student outcome measures collected at the end of PK (immediate posttest) and kindergarten (follow-up posttest).

[5] The longer of either 3 hours or the entire instructional time, excluding lunch, outdoor or gross motor play, and nap times.

[6] In fact, challenges in gaining permission for data collectors to enter classrooms delayed the baseline data collection until nearly one month after the initial training in some cases.

Table 4
Data Collection Schedule: Teacher and Classroom Data.

Teacher and classroom measures	Year 1		Year 2		Year 3	
	Fall 2009	Spring 2009	Fall 2010	Spring 2011	Fall 2011	Spring 2012
Teacher survey	X			X		
COEMET—mathematics teaching observation	X	X		X		
MPOT—self-regulation observation				X		

Assessing Intervention Fidelity and Impacts on Teacher Practices

The first research question pertains to how well the two interventions were implemented by teachers and the impacts of the two interventions on teachers' practice. In previous experiments, the COEMET items that were mediators of the effects of the BB intervention fell into five categories: classroom culture score,[7] number of SMAs, average SMA score,[8] percentage of SMAs conducted as small-group activities, and number of computers turned on and working for students. We looked at the distributions of each of these COEMET constructs and found that there was not enough variation in the number of classroom computers between classrooms (70% of treatment classrooms had two computers, and 20% had no computers). We *z*-scored[9] the remaining four chosen COEMET constructs and calculated the average of the four *z*-scores to construct the overall COEMET intervention fidelity (IF) measure.

For the MPOT IF measure, we included the play routine checklist score, child actions score, and adult actions score. As with the COEMET IF measure, we *z*-scored each dimension score and calculated the average of the three *z*-scores to construct the overall MPOT IF measure. Figure 1 shows the distribution of the COEMET IF measure within the BAU, BB, and BBSEF groups, respectively. This figure suggests that BB and BBSEF classrooms had similar COEMET IF score distributions, and both groups' scores exceeded those of the BAU classrooms. Figure 2 displays similar histograms for the MPOT IF measure. This figure shows that BBSEF classrooms scored much higher on this measure than the BB and BAU classrooms, and the distribution of the MPOT IF index was similar in the latter two groups. Altogether, these graphs provide some empirical evidence for potential heterogeneity in the fidelity of intervention within the BB and BBSEF groups, with some classrooms exceeding the fidelity scores of other classrooms in their respective groups or the BAU group.

[7] The classroom culture score includes the classroom environment and interaction, as well as personal attributes of the teacher.

[8] The average specific math activity score includes scores on the activity's mathematical focus; organization, teaching approaches, and interactions; expectations; eliciting children's solution methods; supporting children's conceptual understanding; extending children's mathematical thinking; and assessment and instructional adjustment.

[9] *Z*-scoring each measure involved subtracting the measure's mean from each observation and dividing the result by the measure's standard deviation. Means and standard deviations are obtained from the control classrooms.

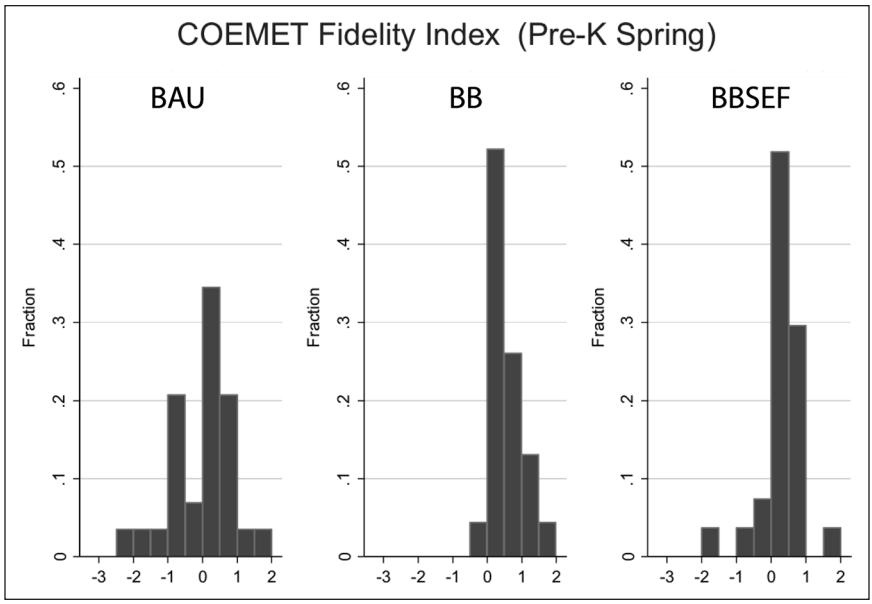

Figure 1. Distribution of COEMET Intervention Fidelity (IF) measure across the three groups.

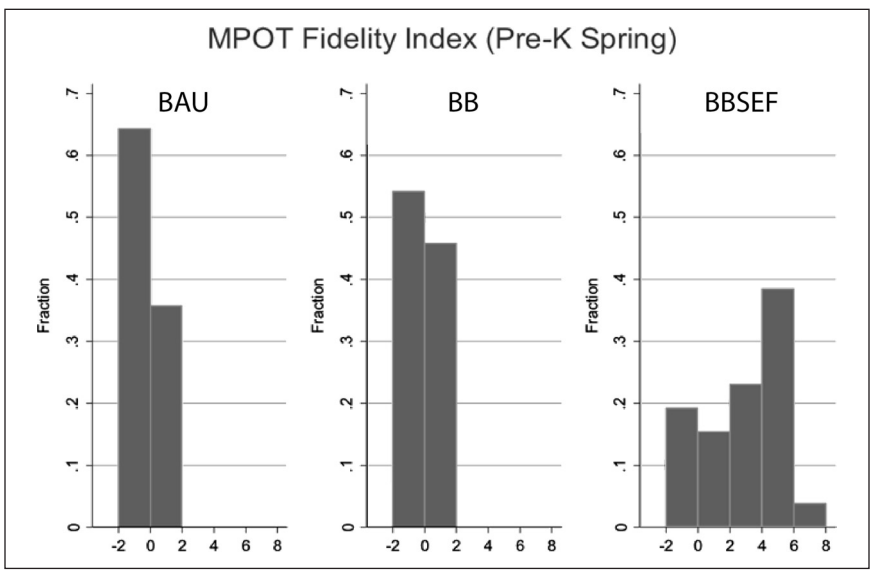

Figure 2. Distribution of MPOT Intervention Fidelity (IF) measure across the three groups.

To address the second part of the first research question regarding the impact of the two interventions on teachers' practice, we compared selected measures from COEMET, MPOT, and Teacher Surveys administered during Spring 2011 across the three study conditions. We selected a priori items that we hypothesized

would be most strongly related to the two interventions' logic models. We then compared outcomes using simple regression models that include indicators for BB and BBSEF conditions and the randomization blocks. The corresponding results are shown in Table 5.

Table 5 shows that, compared to the BAU condition, BB had sizeable and statistically significant positive impacts on three of the four COEMET measures tested (Culture, SMA Number, and Average SMA Score), whereas BBSEF had large positive effects on two measures (Culture and Average SMA Score). As for the MPOT measures, BB and BAU classrooms had similar scores, whereas BBSEF classrooms' scores greatly exceeded those of the other two groups. Impacts on the Teacher Survey items were mixed: BB had significant and positive impacts at the $p < .05$ level on 10 of the 16 measures tested, whereas BBSEF had significant impacts on five measures at the $p < .05$ level. The teachers' report on the surveys for the BBSEF classrooms echoed the observed practices. Responses on the teacher survey for BBSEF teachers showed that about half of the responding BBSEF teachers reported that they had difficulty integrating the critical components of the synthesized approach, 83% reported that time was a significant barrier to effective implementation, and 58% reported that support was not sufficient to support implementation. However, about 71% indicated that they were aware of their children's EF needs, and 75% reported that their children had adequate opportunities to practice EF skills. About 78% reported that they integrated BB into themed play scenarios and EF into activities very frequently at several times per week. Taken together, these results support the argument that the two interventions had considerable impacts on teachers' practice.

Analysis of Student Measures Collected in Fall 2010

Although Fall 2010 student assessments were intended to serve as baseline measures for the cohort of students who started PK during that school year (note that random assignment of teachers had occurred in Fall 2009, but the student sample for the outcomes analysis was composed of students who entered PK in the fall of the 2010–2011 school year), data collection faced obstacles in some districts because of delays in obtaining official permission to begin assessments, completing certification of data collectors on the TEAM assessment, obtaining official district clearance of data collectors in each district, and scheduling directly with teachers. Figure 3 displays the distribution of the time (in days) between the start of the school year and the administration of baseline assessments across the study classrooms separately for the three study groups. As this figure suggests, in almost half of the classrooms, the data collection was completed as many as 2 months after the school's start date. We tried to minimize these differences as much as possible, but some were inevitable. Because the teachers (presumably) implemented the BB and BBSEF interventions during this time, we were concerned that the collected measures might have been contaminated by late pretests (i.e., reflect early treatment effects).

We addressed the late pretest issue via two sets of analyses, which are described in more detail in Appendix B. First, we compared the average scores of the BB and BBSEF classrooms on the pretest measures to those of the BAU classrooms. Table A1 shows the results of these comparisons, which show that (a) BB students

Table 5
Impacts on COEMET, MPOT, and the Teacher Survey in 2011.

Measure	Number of observations	Standardized differences (effect sizes)			Unadjusted	
		BB versus BAU	BBSEF versus BAU	BB versus BBSEF	BAU mean	BAU *SD*
COEMET 2011						
Culture	79	.728*	.601*	.127	3.466	.587
SMA number	79	.649*	.257	.392	2.207	1.719
Average SMA score	79	.604*	.591*	.013	2.780	1.336
Percent small-group SMAs	79	.168	-.311	.479	.063	.163
MPOT 2011						
Play routine checklist score	78	-.006	3.929*	-3.935*	1.321	1.156
Child activity	78	.163	2.385*	-2.221*	6.786	1.707
Adult activity	78	-.194	2.029*	-2.222*	5.179	1.422
Teacher survey 2011						
Teacher perception: adequacy of time for math content	65	.447*	.110	.336	3.136	.941
Frequency: embed assessment in regular class activities	66	.274	.333	-.059	3.435	.788
Frequency (students): hands-on math	66	.372*	.127	.246	3.826	.491
Frequency (students): use computers or calculators–learn/ practice skills	63	1.032*	1.124*	-.092	2.636	.953
Frequency of: large group instruction	66	.221	.008	.213	3.957	.209
Frequency of: center-based activities/choice time	66	.162	-.554	.716	3.913	.288
Teacher perception: adequacy of time for math content	64	.070	.237	-.167	3.227	.752
Frequency: provide opportunities for students to discuss math with one another	66	.412*	.234	.178	3.261	.864
Importance to you of [a range of instructional strategies]	66	.583*	.698*	-.115	3.571	.342

Table 5 (continued)
Impacts on COEMET, MPOT, and the Teacher Survey in 2011.

Measure	Number of observations	Standardized differences (effect sizes)			Unadjusted	
		BB versus BAU	BBSEF versus BAU	BB versus BBSEF	BAU mean	BAU *SD*
Frequency (students): discussions with teacher to further math understanding	64	.584*	.640*	-.056	3.478	.665
Frequency (students): memorize math facts, rules, and formulas	66	.039	-.314	.353	2.783	.998
Frequency (students): student-led discussions	64	.391*	.420*	-.030	2.957	1.065
Adjust task to child's developmental level	66	.399*	.642*	-.243	2.674	.792
Math professional development at own school	66	.405*	.587*	-.182	2.754	.645
Support from administration	66	.158	.274	-.116	3.118	.619

*$p < .05$.

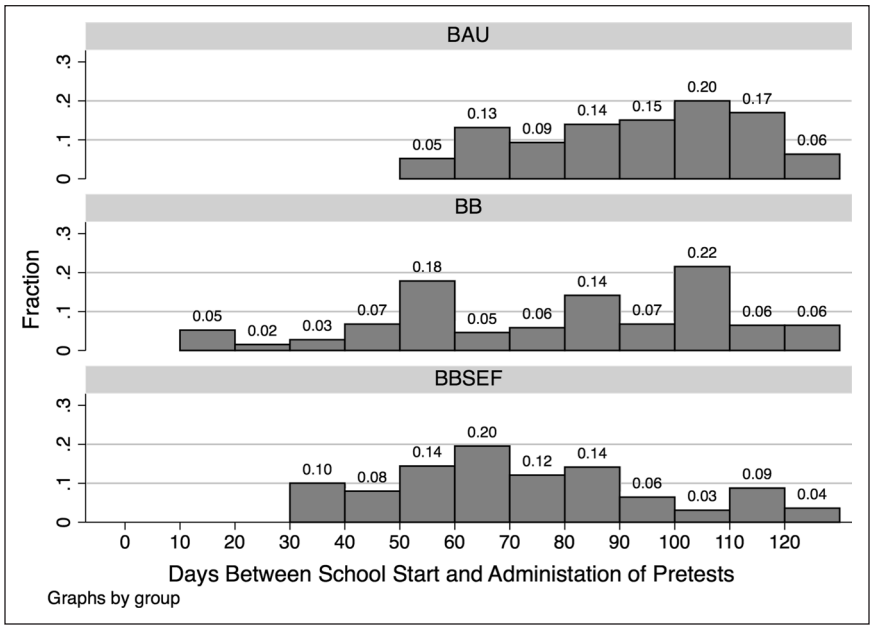

Figure 3. Number of days between the start of the school year and the administration of baseline assessments.

scored between 0.1 and 0.2 standard deviations higher than their BAU peers on most of the mathematics and EF measures and (b) estimated differences between the BBSEF and BAU students were smaller and in favor of the BAU students for most outcomes. The second set of analyses was motivated by this suggestive evidence for the early treatment effects on the pretest measures (at least for the BB group). These analyses followed Schochet (2008) and examined the analytical implications of using late pretests in the estimation of impacts on posttest measures. As discussed in Appendix B, these analyses led us to not use the pretest measures in the estimation of impacts on posttest measures.

Excluding pretest measures from the estimation on impacts on posttest measures caused a substantial loss in statistical power.[10] To compensate for this unanticipated power loss and given that the analyses of the late pretest measures provided suggestive evidence for the positive impact of the BB condition, we used one-sided hypothesis tests conducted at the $p < .05$ level when comparing the posttest measures of the BB group to those of the BAU group. When comparing the outcomes of the BBSEF group to the other two groups, we used two-sided significance tests at the $p < .05$ level given our initial hypothesis that this condition would have better outcomes than the other conditions at posttest, but the analyses of the late pretest measures did not support this hypothesis.

Analysis of Student Measures Collected in Spring 2011 (Immediate Posttest) With Those of the Full Study Sample

To address the second research question pertaining to the immediate impacts of the two intervention conditions, we calculated the BB and BBSEF impacts on the achievement measures collected in Spring 2011. These analyses were conducted within the intent-to-treat (ITT) framework in which classrooms and schools were analyzed according to the group to which they had been initially randomized. We did not conduct additional treatment-to-treated analyses, given that noncompliance with random assignment was virtually zero.

The impact estimates were calculated using two-level hierarchical linear models (HLMs) (students nested within classrooms[11]) that controlled for student gender and age but not the potentially contaminated pretest measures. The model also included indicators (i.e., fixed effects) for randomization blocks described in the Research Design section. Specifically, we used the following combined two-level model specification to estimate the impacts:

$$Y_{ij} = \beta_{00} + \beta_{10} BB_j + \beta_{20} BBSEF_j + \sum_{n=1}^{N} \beta_{(2+n)0} X_{nij} + v_j + \varepsilon_{ij} \qquad (1)$$

[10] The ex-ante power analyses that we conducted at the design stage of the project assumed that pretest measures would explain 75% of the cluster-level variance and 50% of the student-level variance of the posttest measures. Under this assumption, we determined the target sample size of the study for a minimum detectable effect size (MDES) of .2. The actual outcome-covariate correlations without the pretest measures were much smaller and increased the ex-post MDES to .3.

[11] We did not include an explicit school level in this model because most of the schools had only one classroom. Sensitivity analyses conducted with the alternative model specification that included a separate school level (as Level 3 in three-level HLMs) yielded very similar estimates to this model, and those results are available upon request.

In this model, Y_{ij} denotes the outcome measure (e.g., TEAM score in Spring 2011) for student *i* taught by teacher *j;* BB_j is the indicator variable for the BB group for teacher *j* (set to 1 if teacher *j* is assigned to the BB group and to 0 if otherwise); $BBSEF_j$ is the indicator variable for the BBSEF group for teacher *j* (set to 1 if teacher *j* is assigned to the BBSEF group and to 0 if otherwise); X_{nij} is the *n*th covariate including student age, gender, and the indicator variables for randomization blocks (which is set to 1 if teacher *j* is in the randomization block indicated by that variable and 0 if otherwise[12]); v_j is the random effect for teacher *j*; and ε_{ij} is the usual error term for student *i* taught by teacher *j*, which is assumed to be independent of the classroom (teacher) error term.[13]

We estimated the model via the *xtmixed* procedure in Stata using restricted maximum likelihood. In the model output, we interpreted the estimate of β_{10} as the adjusted difference between the BB and BAU groups (i.e., impact of BB vs. BAU) and the estimate of β_{20} as the adjusted difference between the BBSEF and BAU groups (i.e., impact of BBSEF). The difference between the two coefficients yields the adjusted difference between the BB and BBSEF groups. We converted the impact estimates to standardized effect size units using the BAU group standard deviations.[14]

Table 6 presents the corresponding impact estimates (or, more accurately, pairwise comparisons of the outcomes between any two of the three study groups) that are expressed in effect sizes using the standard deviation of the BAU group (BAU group mean and standard deviations are also shown in this table in the last two columns). Table 6 shows that although most of the mathematics and EF impacts for the BB students (BB vs. BAU contrast) are positive and larger than 0.1 standard deviations, only one impact attains statistical significance at the $p < .05$ level (Backward Digit Span, effect size = 0.19). The pattern in the impacts for BBSEF students is somewhat mixed, with some estimates being positive and some negative but none statistically significant. Overall, these results do not make a strong case for students being positively affected by either intervention condition compared to the BAU curriculum through the end of PK, the only year in which the interventions were implemented.

Given that the ECLS-B measure was used on a national sample (*T*-scores, $M = 50$, $SD = 10$), this study's scores can be compared to those of the U.S. population: *T*-scores for 2011 and 2012 were 45.5 and 50.0, respectively, for BB; 44.0 and 48.5, respectively, for BBSEF; and 44.1 and 48.6, respectively, for BAU. Most were slightly below the national norms, but the Spring 2012 score for the BAU group

[12] As explained previously, for teachers in Group A, randomized blocks were based on the district and whether the program was full day or half day. For teachers in Group B, schools served as randomization blocks. One of the blocks is considered as the reference block and its indicator is excluded from the model specification.

[13] We also estimated an unconditional version of this model (i.e., model with no covariates) for each outcome to calculate unconditional intra-class correlations, which varied between .02 and .10. The results of analyses presented here arise from model testing; therefore, we do not conduct any adjustments for multiple comparisons.

[14] Resulting effect sizes are sometimes referred to as Glass's delta and are preferred to other effect-size metrics because they do not reflect any effects that the treatment conditions may have had on the standard deviation of the outcome measures.

Table 6
Impacts of BB and BBSEF on Spring 2011 Assessment Measures (Immediate Posttest, PK).

Measure	Number of observations	Standardized differences (effect sizes)			Unadjusted	
		BB versus BAU	BBSEF versus BAU	BB versus BBSEF	BAU mean	BAU *SD*
Mathematics outcomes						
TEAM—scaled score (total)	819	.116	.092	.025	356.3	91.6
ECLS-B Math (Pre-K)	781	.151	-.017	.168	18.7	6.5
Executive function outcomes						
HTKS score	779	.165	.051	.113	14.8	14.3
Forward digit span	789	.111	-.012	.124	3.7	1.3
Backward digit span	785	.187*	-.001	.188*	.5	.9
Peg Tapping	783	.095	.024	.071	10.0	6.0
Language and literacy outcomes						
EVT-2 score	797	.074	.032	.042	54.1	20.2
RBS—composite	680	-.034	-.128	.095	2.62	.98
PALS—alphabet knowledge (A + B + C)	791	-.015	-.079	.064	39.3	28.0

Note. As explained in the main text, the significance of the BB versus BAU differences was tested using one-sided tests and the contrasts involving BBSEF (BBSEF vs. BAU and BB vs. BBSEF) were tested using two-sided tests.
*$p < .05$.

indicates that the districts' emphasis on mathematics professional development may have been effective. This increased emphasis could also explain some of the changes observed in the BB and BBSEF classrooms.

Analysis of Student Measures Collected in Spring 2012 (Follow-Up Posttest) With the Full Study Sample

This section presents the results of the student achievement measures collected at the end of the kindergarten year (Spring 2012), during which students were exposed to the BAU instructional practices in their respective schools and districts. These analyses address the third research question, regarding the impacts of the two interventions on follow-up posttest measures, and were conducted similarly to the immediate posttest measures using multivariate two-level HLMs that cluster students in the schools and centers at PK (as they were exposed to the interventions in those settings) but do not control for the late pretest measures.

Table 7 presents the corresponding results. Comparing the first column in this table with the first column in Table 5 suggests that the size of the analytical

Table 7
Impacts of BB and BBSEF on Spring 2012 Student Assessment Measures (Follow-Up Posttest).

Measure	Number of observations	Standardized differences (effect sizes)			Unadjusted	
		BB versus BAU	BBSEF versus BAU	BB versus BBSEF	BAU mean	BAU *SD*
Mathematics outcomes						
TEAM—scaled score (total)	755	.191*	.105	.086	448.4	60.6
ECLS-B Math (K)	754	.086	-.008	.094	42.6	14.8
Executive function outcomes						
HTKS score	746	.114	.005	.109	25.5	12.9
Forward digit span	751	.196*	.059	.136	4.3	1.2
Backward digit span	740	.127	.045	.082	1.4	1.4
Peg Tapping	754	.156*	.058	.098	13.4	4.0
Language and literacy outcomes						
EVT-2 score	750	.122	.033	.089	68.6	18.3
RBS—composite	668	.129	-.008	.138	3.01	1.02
PALS—letter Identification	755	-.201	-.142	-.059	23.6	4.3

Note. As explained in the main text, the significance of the BB versus BAU differences was tested using one-sided tests and the contrasts involving BBSEF (BBSEF vs. BAU and BB vs. BBSEF) were tested using two-sided tests.
*$p < .05$.

samples for these analyses is roughly 10% smaller than those of the Spring 2011 measures, where the reduction in the sample size is mostly because of the children's mobility between the two time points.

Almost all the BB versus BAU differences at this time point were larger than in Spring 2011. Impacts on TEAM measures are around 0.2 standard deviations; the effect size for the scaled score is 0.19 and statistically significant at the $p < .05$ level. Other statistically significant impact estimates are found for Forward Digit Span (effect size = 0.2, $p < .05$) and Peg Tapping (effect size = 0.16, $p < .05$). Impact estimates for other measures are generally positive but not statistically significant. Compared to the Spring 2011 results, estimates for the BBSEF versus BAU contrasts are directionally more positive, but none of them reach statistical significance.

Discussion and Implications

Heated debates continue in early childhood education centered on the issue of the proper role for content-oriented and play-based approaches and curricula. We designed this study to collect evidence to ascertain whether these two approaches stand in opposition or could be synergistically combined. To do so, we evaluated

two preschool interventions for fidelity of implementation and effects on child outcomes, the BB mathematics curriculum (Clements & Sarama, 2007/2013) and this curriculum synthesized with TotM-based (Bodrova & Leong, 2001) teaching of EF, including scaffolding of play and similar scaffolding integrated into the mathematics activities (BBSEF). Although the results do not confirm all our hypotheses, they support some, and they suggest an unexpected but promising causal path, raising more questions for future research.

Research Question 1: Can the Two Interventions Be Implemented with High Fidelity and Have Substantial Positive Effects on Teachers' Practice?

Previous work indicated that most teachers can implement each of the BB and SEF components with acceptable fidelity and can make significant gains in research-based practice with adequate professional knowledge and support (e.g., Bodrova & Leong, 2005; Clements & Sarama, 2007). We hypothesized that both components of the synthesized intervention could be adequately implemented. A strong majority of the BBSEF teachers indicated that they were using the intervention and that their children were practicing EF skills, although about one half reported implementation challenges. Further, substantial high-quality professional development and support were provided in the current study and, most importantly, adequate levels of fidelity were achieved on both classroom observation measures in line with the assigned interventions. The only data indicating lower fidelity for the BBSEF intervention was the pattern of higher fidelity for the BB component in the BB compared to the BBSEF groups. Nevertheless, given that measures indicate that implementation was satisfactory, the lack of results for the BBSEF group was mainly a theory failure rather than an implementation failure, especially for the SEF component. We return to these results throughout this discussion.

Research Question 2: What Are the Immediate Effects of the Two Interventions (BB and BBSEF), as Implemented Under Diverse Conditions, on Children's Achievement and EF?

Our results did not support most of our hypotheses. At the end of preschool, outcomes in the synthesized curriculum (BBSEF) were not statistically distinguishable from the BAU control group, with some effect sizes being slightly positive and others slightly negative. Most of the hypothesized impacts for the mathematics curriculum (BB vs. BAU contrast) were positive and larger than 0.1 standard deviations, but only one impact attained statistical significance at the $p < .05$ level (Backward Digit Span, effect size = 0.19; recall that the lack of valid pretest measures attenuated all measured impacts). Surprisingly, the BB versus BBSEF contrasts were also positive, with, again, only the Backward Digit Span reaching statistical significance. Because both treatment groups covered the BB curriculum (and the BAU did substantial instruction in number using Richardson's [2008] materials), simply being exposed to numbers would not explain this difference. Rather, more intensive focus on the BB learning trajectories is the most cogent explanation, given the strong relationships between mathematics competencies and working memory (Clements et al., 2016), which may be bidirectional (van der Ven et al., 2012). As an example, consider the use of counting back to

solve a subtraction problem; one must keep the goal (find the difference) in mind as well as the part-whole relationship, then keep track of how many "counts" one makes (e.g., 8–2, counting back to 7 is 1 count, counting back again to 6 is 2 counts—stop and report "6"). Working through the levels of a learning trajectory that lead up to this competence may exercise and therefore build working memory incrementally. Of course, BB includes many components; however, the two mathematics curricula in the BAU classrooms also include many of these, such as attention to children's thinking, discourse, and active small-group and individual experiences (along with extensive professional development).

Thus, the solution of arithmetical problems may require children to expand their application of working memory but also provide scaffolding for such extension. As one example, the contexts of story problems related to real-world experiences with which children are familiar may provide such scaffolding through the incorporation of the "narrative mode" (Bruner, 1986) of thinking, which provides sequential and interpretable situations that guide children's translation of the situation into the logical and systematic structures of mathematics. However, such results require replication.

Overall, these results do not make a strong case for students being positively affected by either intervention condition compared to the BAU curriculum through the end of PK, the only year during which the interventions were implemented. The results were more positive for the BB intervention, even though the BBSEF teachers received twice as much professional development. Challenges implementing the synthesized intervention might have led to lower performance of the BBSEF group than of the BB group and negligible effects of BBSEF compared with the BAU group. Although both treatment groups scored higher than the BAU group on some measures of the quantity and quality of classroom teaching, only the BB group scored significantly higher on the number of SMAs, a variable that significantly mediated gains in previous research (Clements et al., 2011; the BB group scored higher than the BBSEF group on all measures, although these differences were not statistically significant).

Research Question 3: What Are the Longer Term Effects of the Two Interventions?

The results at the end of kindergarten, with no intervention follow-up after the end of PK, were mixed. Consistent with our initial hypothesis, children in the BB group outperformed those in the BAU group in math achievement (positive on both measures; statistically significant on one). However, the BBSEF group showed no statistically significant differences compared to either of the other two groups, with effect sizes favoring the BAU group on some measures and effect sizes favoring the BB group on all measures.

In contrast to a common concern about "fade out" of effects (Watts et al., 2017; Watts et al., 2018), intervention impacts did not diminish even after a year of kindergarten in which children formerly in the intervention groups were combined with children who were not in the interventions and were taught by teachers who did not receive any intervention training.

Further, the results supported an alternative causal path to gains in EF that we had not hypothesized: The BB group outperformed the BAU group and obtained

directionally more positive outcomes that did not attain statistical significance compared with the group whose intervention was designed to develop EF (BBSEF) on two measures of EF (Forward Digit Span, a measure of phonological processing, and Peg Tapping, a measure of inhibitory control). Both phonological processing and inhibitory control arguably contribute strongly to children's preparedness to learn more advanced content.

These somewhat surprising results may provide some support for three hypotheses. First, as mentioned above, combining two interventions implemented by the BBSEF group may have created challenges for affecting the very EF outcomes that one of the interventions specifically targets. Second, early gains in both mathematics and EF competence can be mutually supportive and thus resist the fade-out effect. Third, the pattern of results suggests that gains may have stemmed from a focus on mathematics learning implemented by teachers focusing on those activities alone (not in combination with another approach). The learning trajectories at the core of the BB curriculum, which motivate and support progressive movement through increasingly challenging activities and levels of thinking, may have helped children develop new EF processes simultaneously with particular mathematical competencies. The greater number of SMAs may have provided additional opportunities both for learning mathematics and for increasing EF competencies. The pattern of results may indicate that the BB teachers engaged children not only more extensively but also more intensely in mathematical thinking, which may have placed more demands on children's use of EF processes (note the BB group did a higher percentage of mathematical activities in small groups, as the curriculum suggests, compared with the BBSEF group).

These hypotheses are also consistent with other recent research. One of the evaluations that found little effect of the TotM program also produced evidence regarding the potential of mathematics curricula alone (Farran et al., 2011). In this large-scale evaluation, the more focus the classroom and teacher had on mathematics, the greater the children's gains in both mathematics and EF (Farran et al., 2011). If replicated, these results have important and wide-ranging implications for curriculum design and pedagogical practices.

In drawing additional implications for both educational research and practice concerning the disappointing results of the SEF component, we consider the results of other recent studies. When we planned this study, extant studies at least tentatively indicated that the TotM-based scaffolding could increase children's EF competencies (Barnett et al., 2008; A. Diamond et al., 2007). However, more recent randomized cluster trial evaluations of the TotM program or the part of TotM targeting EF showed no effects on EF, even with implementations of adequate fidelity (Farran et al., 2011; Lonigan & Phillips, 2012; Morris et al., 2014), and reanalyses of earlier studies reached the same conclusion (Jacob & Parkinson, 2015), although a recent study reported small but positive effects (Blair & Raver, 2014). This last study, however, focused on embedding support for EF into literacy, mathematics, and science learning activities—so the provision of those activities confounded any EF scaffolding of play. Further, it was conducted in kindergarten, not PK, classes. Researchers need to resolve these divergent findings; for example, they may find that the TotM approach is measurably effective (on certain instruments) only with intensive and extensive supports for implementation.

Further, other researchers claim that there is little or no evidence that pretend play is crucial to building EF or other competencies (Chien et al., 2010; Lillard et al., 2013), so that aspect of the approach may be contraindicated and it may be that high-quality academic learning activities are accounting for the positive effects in all studies, including the present one. This is a reasonable conclusion given all the available evidence, and, if supported by additional studies, would imply a radical restructuring of TotM's theoretical basis and implications for practitioners.

These recent studies provide a different context for interpreting the results of our study. First, regardless of these cautions about the effectiveness of the TotM approach, the comparatively smaller effects of the BBSEF than the BB intervention on all measures, especially mathematics achievement, suggest that implementing both approaches may have interfered with a focus on the mathematics activities, perhaps including nuanced but important ways that the fidelity measures did not capture.

The second point is a caveat concerning implications for future research and development. The present study and other recent research suggest that the TotM-based scaffolding used in this study may be challenging to implement and not efficacious. To the extent that this is true, it leaves open the possibility that alternative high-quality play-based approaches may be less challenging to implement and more effective when synthesized with subject-matter curricula (see, e.g., Sarama et al., 2017).

Recent research indicates that the causal evidence that interventions designed to develop EF increase achievement is weak or missing (Clements et al., 2016). Further, early mathematics competencies predict later mathematics achievement (as well as later reading achievement; Duncan et al., 2007) and early EF does not (once early mathematics is factored in), but early mathematics predicts later EF (Watts et al., 2015). These studies, along with the results of the present study, suggest a unique approach: High-quality mathematics education may have the dual benefit of teaching an important content area and developing at least some EF competencies. An even more intentional development of mathematics curricula based on recent research on EF may do both even more effectively (e.g., Banse et al., in press; Joswick et al., 2019). If confirmed with additional research, the implications for practice are substantial, especially given that this type of intentional instruction in small groups using research-based teaching strategies is more effective than other approaches (Chien et al., 2010) and rarely employed by early childhood teachers (K. E. Diamond et al., 2013).

References

Banse, H. W., Clements, D. H., Sarama, J., Day-Hess, C. A., & Joswick, C. (in press). Strategies for supporting executive function development: Practical takeaways for early childhood teachers. *Young Children*.

Barnett, W. S., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., & Burns, S. (2008). Educational effects of the tools of the mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, *23*(3), 299–313. https://doi.org/10.1016/j.ecresq.2008.03.001

Barnett, W. S., Yarosz, D. J., Thomas, J., & Hornbeck, A. (2006). *Educational effectiveness of a Vygotskian approach to preschool education: A randomized trial*. National Institute of Early Education Research, Rutgers, The State University of New Jersey.

Berends, M., Kirby, S. N., Naftel, S., & McKelvey, C. (2001). *Implementation and performance in New American Schools: Three years into scale-up*. RAND.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*(4), 327–336. https://doi.org/10.1016/j.lindif.2011.01.007

Blair, C., Protzko, J., & Ursache, A. (2011). Self-regulation and early literacy. In S. B. Neuman & D. K. Dickinson. (Eds.), *Handbook of early literacy research* (Vol. *3*, pp. 20–35). New York, NY: Guilford.

Blair, C., & Raver, C. C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. *PLoS One*, *9*(11), e112393. https://doi.org/10.1371/journal.pone.0112393

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Bodrova, E., Germeroth, C., & Leong, D. J. (2013). Play and self-regulation: Lessons from Vygotsky. *American Journal of Play*, *6*(1), 111–123.

Bodrova, E., & Leong, D. J. (2001). *Tools of the mind: A case study of implementing the Vygotskian approach in American early childhood and primary classrooms*. International Bureau of Education.

Bodrova, E., & Leong, D. J. (2005). Self-regulation as a key to school readiness: How early childhood teachers can promote this critical competency. In M. Zaslow & I. Martinez-Beck (Eds.), *Critical issues in early childhood professional development* (pp. 203–224). Paul H. Brookes.

Bodrova, E., & Leong, D. J. (2006). The development of self-regulation in young children: Implications for teacher training. In M. Zaslow & I. Martinez-Beck (Eds.), *Future directions in teacher training* (pp. 203–224). Brooks Cole.

Bodrova, E., & Leong, D. J. (2007a). Play and early literacy: A Vygotskian approach. In K. A. Roskos & J. F. Christie (Eds.), *Play and literacy in early childhood: Research from multiple perspectives* (2nd ed., pp. 185–200). Lawrence Erlbaum Associates.

Bodrova, E., & Leong, D. J. (2007b). *Tools of the mind: The Vygotskian approach to early childhood education* (2nd ed.). Pearson/Merrill Prentice Hall.

Bruner, J. S. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*(1), 36–41. https://doi.org/10.1111/cdep.12059

Campbell, P. F., & Silver, E. A. (1999). *Teaching and learning mathematics in poor communities: A report to the Board of Directors of the National Council of Teachers of Mathematics*. National Council of Teachers of Mathematics.

Chien, N. C., Howes, C., Burchinal, M., Pianta, R. C., Ritchie, S., Bryant, D. M., Clifford, R. M., Early, D. M., & Barbarin, O. A. (2010). Children's classroom engagement and school readiness gains in prekindergarten. *Child Development*, *81*(5), 1534–1549. https://doi.org/10.1111/j.1467-8624.2010.01490.x

Clements, D. H. (2007). Curriculum research: Toward a framework for "research-based curricula. *Journal for Research in Mathematics Education*, *38*(1), 35–70. https://doi.org/10.2307/30034927

Clements, D. H., Fuson, K. C., & Sarama, J. (2017). The research-based balance in early childhood mathematics: A response to Common Core criticisms. *Early Childhood Research Quarterly*, *40*, 150–162. https://doi.org/10.1016/j.ecresq.2017.03.005

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*(2), 136–163. https://doi.org/10.2307/30034954

Clements, D. H., & Sarama, J. (2007/2013). *Building blocks* (Vols. 1–2). McGraw-Hill Education.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*(2), 443–494. https://doi.org/10.3102/0002831207312908

Clements, D. H., & Sarama, J. (2011). *TEAM—Tools for early assessment in mathematics*. McGraw-Hill Education.

Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math: The learning trajectories approach* (2nd ed.). Routledge.

Clements, D. H., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, *36*, 79–90. https://doi.org/10.1016/j.ecresq.2015.12.009

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-based Early Maths Assessment. *Educational Psychology*, *28*(4), 457–482. https://doi.org/10.1080/01443410701777272

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127–166. https://doi.org/10.5951/jresematheduc.42.2.0127

Cobb, P., McClain, K., Lamberg, T. D., & Dean, C. (2003). Situating teachers' instructional practices in the institutional setting of the school and district. *Educational Researcher*, *32*(6), 13–24. https://doi.org/0.3102/0013189X032006013

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, *318*(5855), 1387–1388. https://doi.org/10.1126/science.1151148

Diamond, K. E., Justice, L. M., Siegler, R. S., & Snyder, P. A. (2013). *Synthesis of IES research on early intervention and early childhood education* (NCSER 2013-3001). National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncser/pubs/20133001/pdf/20133001.pdf.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). American Guidance Service.

Farran, D. C., Lipsey, M. W., & Wilson, S. J. (2011, November). *Curriculum and pedagogy: Effective math instruction and curricula*. [Paper presentation]. Early Childhood Math Conference, Berkeley, CA, USA.

Gallimore, R., & Stigler, J. (2003). LessonLab: Evolving teaching into a profession. *TechKnowLogia*, *5*(1), 32–34. [Mismatch]

Germeroth, C., Bodrova, E., Day-Hess, C. A., Barker, J., Sarama, J., Clements, D. H., & Layzer, C. (2019). Play it high, play it low: Examining the reliability and validity of a new observation tool to measure children's make-believe play. *American Journal of Play*, *11*(2), 183–221.

Germeroth, C., & Sarama, J. (2017). Coaching in early mathematics. *Advances in Child Development and Behavior*, *53*, 127–168. https://doi.org/10.1016/bs.acdb.2017.04.003

Glasgow, C., & Cowley, J. (1994a). *Renfrew bus story test (North American edition)*. Learning Tools LLC.

Golinkoff, R. M., Hirsh-Pasek, K., & Singer, D. G. (2006). Why play = learning: A challenge for parents and educators. In D. G. Singer, R. M. Golinkoff, & K. Hirsh-Pasek (Eds.), *Play = learning: How play motivates and enhances children's cognitive and social-emotional growth* (pp. 3–12). Oxford University Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *The early childhood environment rating scale: Revised edition*. Teachers College Press.

Heck, D. J., Weiss, I. R., Boyd, S., & Howard, M. (2002, April). *Lessons learned about planning and implementing Statewide Systemic Initiatives (SSIs) in mathematics and science education* [Paper presentation]. *American Educational Research Association*, New Orleans, LA. www.horizon-research.com/public.htm.

Invernizzi, M., Sullivan, A., Swank, L., & Meier, J. (2004). *PALS pre-K: Phonological awareness literacy screening for preschoolers* (2nd ed.). University Printing Services.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, *85*(4), 512–552. https://doi.org/10.3102/0034654314561338

Joswick, C., Clements, D. H., Sarama, J., Banse, H. W., & Day-Hess, C. A. (2019). Double impact: Mathematics and executive function. *Teaching Children Mathematics*, *25*(7), 416–426.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the pre-kindergarten year. *The Elementary School Journal*, *104*(5), 409–426. https://doi.org/10.1086/499760

Lahiri, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *International Statistical Institute Bulletin*, *33*(2), 133–140.

Lewis Presser, A., Clements, M., Ginsburg, H., & Ertle, B. (2015). Big math for little kids: The effectiveness of a preschool and kindergarten mathematics curriculum. *Early Education and Development*, *26*(3), 399–426. https://doi.org/10.1080/10409289.2015.994451

Lillard, A. S., Lerner, M. D., Hopkins, E. J., Dore, R. A., Smith, E. D., & Palmquist, C. M. (2013). The impact of pretend play on children's development: A review of the evidence. *Psychological Bulletin*, *139*(1), 1–34. https://doi.org/10.1037/a0029321

Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., & Wilson, S. J. (2017). Learning-related cognitive self-regulation measures for prekindergarten children: A comparative evaluation of the educational relevance of selected measures. *Journal of Educational Psychology*, *109*(8), 1084–1102. https://doi.org/10.1037/edu0000203

Lonigan, C. J., & Phillips, B. M. (2012, March). *Comparing skills-focused and self-regulation focused preschool curricula: Impacts on academic and self-regulatory skills*. [Paper presentation]. Society for Research on Educational Effectiveness, Washington, DC, USA.

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in Psychology*, *5*, 599. https://doi.org/10.3389/fpsyg.2014.00599

Morris, P., Mattera, S. K., Castells, N., Bangser, M., Bierman, K., & Raver, C. (2014). *Impact findings from the Head Start CARES demonstration: National evaluation of three approaches to improving preschoolers' social and emotional competence* (OPRE Report 2014-44). Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Neuenschwander, R., Röthlisberger, M., Cimeli, P., & Roebers, C. M. (2012). How do different aspects of self-regulation predict successful adaptation to school? *Journal of Experimental Child Psychology*, *113*(3), 353–371.https://doi.org/10.1016/j.jecp.2012.07.004

Pankratz, M. E., Plante, E., Vance, R., & Insalaco, D. M. (2007). The diagnostic and predictive validity of the Renfrew Bus Story. *Language, Speech, and Hearing Services in Schools*, *38*(4), 390–399. https://doi.org/10.1044/0161-1461(2007/040)

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS) manual, pre-K*. Teachstone Training.

Preschool Curriculum Evaluation Research Consortium. (2008). *Effects of preschool curriculum programs on school readiness: Report from the Preschool Curriculum Evaluation Research Initiative* (NCER 2008-2009). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ncer.ed.gov.

Richardson, K. (2008). *Developing math concepts in pre-kindergarten*. Math Perspectives.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge. https://doi.org/10.4324/9780203883785

Sarama, J., & Clements, D. H. (2013). Lessons learned in the implementation of the TRIAD scale-up model: Teaching early mathematics with trajectories and technologies. In T. Halle, A. Metz, & I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 173–191). Paul H. Brookes.

Sarama, J., & Clements, D. H. (2019). *COEMET: The classroom observation of early mathematics environment and teaching instrument*. University of Denver.

Sarama, J., Brenneman, K., Clements, D. H., Duke, N. K., & Hemmeter, M. L. (2017). Interdisciplinary teaching across multiple domains: The C4L (Connect4Learning) curriculum. In L. B. Bailey (Ed.), *Implementing a standards-based curriculum in the early childhood classroom* (pp. 1–53). Routledge.

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*(3), 489–502. https://doi.org/10.1016/j.ecresq.2011.12.002

Schochet, P. Z. (2008). *The late pretest problem in randomized control trials of education interventions* (NCEE 2009-4033). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

University of Chicago School Mathematics Project. (1995/1997). *Everyday mathematics. Preschool. Teacher's resource package*. Everyday Learning Corp.

Unlu, F., Layzer, C., Clements, D. H., Sarama, J., Fesler, L., & Cook, D. (2014, April). *Approaches to incorporating late pretests in experiments: Evaluation of two early mathematics and self-regulation interventions* [Paper presentation]. Annual Conference of the American Educational Research Association, Philadelphia, PA, USA.

van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012). The development of executive functions and early mathematics: A dynamic relationship. *British Journal of Educational Psychology*, *82*(1), 100–119. https://doi.org/10.1111/j.2044-8279.2011.02035.x

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2017). Does early mathematics intervention change the processes underlying children's learning? *Journal of Research on Educational Effectiveness*, *10*(1), 96–115. https://doi.org/10.1080/19345747.2016.1204640

Watts, T. W., Duncan, G. J., Chen, M., Claessens, A., Davis-Kean, P. E., Duckworth, K., Engel, M., Siegler, R., & Susperreguy, M. I. (2015). The role of mediators in the development of longitudinal mathematics achievement associations. *Child Development*, *86*(6), 1892–1907. https://doi.org/10.1111/cdev.12416

Watts, T. W., Duncan, G. J., Clements, D. H., & Sarama, J. (2018). What is the long-run impact of learning mathematics during preschool? *Child Development*, *89*(2), 539–555. https://doi.org/10.1111/cdev.12713

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112–2130. https://doi.org/10.1111/cdev.12099

Weiss, I. R. (2002). *Systemic reform in mathematics education: What have we learned?* [Paper presentation]. Research Presession of the 80th Annual Meeting of the National Council of Teachers of Mathematics, Las Vegas, NV.

Williams, K. T. (2007). *EVT-2: Expressive vocabulary test, second edition. Form B.* Pearson Education.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement* (3rd ed.). Riverside.

Authors

Douglas H. Clements and **Julie Sarama**, Morgridge College of Education, Marsico Institute, University of Denver, Katherine A. Ruffatto Hall 160, 1999 East Evans Avenue, Denver, CO 80208-1700; Douglas.Clements@du.edu and Julie.Sarama@du.edu

Carolyn Layzer, Social and Economic Policy Division, Abt Associates, 10 Fawcett Street, Cambridge, MA 02138; Carolyn_Layzer@abtassoc.com

Fatih Unlu, RAND Corporation, Education and Labor Division, 1776 Main St., Santa Monica, CA 90401; Funlu@rand.org

Lily Fesler, Stanford Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305-3096; lfesler@stanford.edu

doi:10.5951/jresemtheduc-2019-0069

Note: Appendices for this article are available online only at https://pubs.nctm.org/view/journals/jrme/51/3/article-p301.xml?tab_body=supplementaryMaterials