Effects of Prekindergarten Curricula: *Tools of the Mind* as a Case Study

Kimberly T. Nesbitt
University of New Hampshire

Dale C. Farran
Vanderbilt University

Lynn S. Liben
Series Editor

Monographs of the Society for Research in Child Development
Vol. 86, No. 1, 2021, Serial No. 340

Effects of Prekindergarten Curricula: *Tools of the Mind* as a Case Study

Contents

Monographs of the Society for Research in Child Development

Lynn S. Liben, Editor

SRCD

Effects of Prekindergarten Curricula: *Tools of the Mind* as a Case Study

Kimberly T. Nesbitt[1] *and Dale C. Farran*[2]

Abstract Research demonstrates that children's participation in quality early childhood care and education often has immediate positive effects on their social-emotional, self-regulation, and achievement outcomes. Most of the research on the impacts of early child care and education has focused narrowly on the United States, but advocacy for economic and social investment in early childhood care and education to support future children's growth and well-being now exists on an international scale.

The longer-term outcomes from prekindergarten programs have not been as strong. To improve children's long-term outcomes, one suggested strategy is an intentional, scripted curriculum. Our goal in this monograph is to provide a fully integrated and comprehensive account of a large-scale, longitudinal, field-based randomized control trial of the *Tools of the Mind* (Internal consistency of the *Tools*) prekindergarten curriculum that occurred in the United States. Our intent is twofold. First, we examine the impact of the *Tools* curriculum itself, addressing both the potential impacts of the curriculum to improve prekindergarten quality and children's academic, executive function, self-regulation, and social outcomes. Second, we consider the broader question of whether the use of intentional, scripted curricula during early education can, more generally, enhance both short- and long-term outcomes in children.

Developed from a Vygotskian framework, *Tools* focuses on equipping children with cognitive tools for learning that they can then apply to the task of acquiring and sustaining academic knowledge as well as behavioral competencies. Thus, *Tools* is an integrated, comprehensive curriculum, not a

[1] Human Development and Family Studies, University of New Hampshire, [2] Peabody College, Vanderbilt University.

Corresponding author: Kimberly T. Nesbitt, Human Development and Family Studies, University of New Hampshire, Pettee Hall 217, Durham, NH 03824, email: kimberly.nesbitt@unh.edu

Citation Information: Nesbitt, K. T., & Farran, D. C. (2021). Effects of prekindergarten curricula: *Tools of the Mind* as a case study. *Monographs of the Society for Research in Child Development*, *86*(1).

https://doi.org/10.1111/mono.12425

supplementary one. The *Tools* approach follows from a socio-cultural perspective on child development that emphasizes children's acquisition of skills and cultural tools in collaboration with knowledgeable others.

The methodology of the 4-year longitudinal cluster randomized control trial is described in detail. We provide comprehensive information about recruitment, randomization of treatment condition, child assessment instrumentation and procedures, as well as observational assessments, including fidelity of implementation and teacher and child classroom behaviors.

We provide results comparing 32 classrooms assigned to the *Tools* condition and 28 assigned to the business-as-usual control condition for children's academic, executive function, self-regulation, and social gains from prekindergarten to the end of first grade. Developers of the curriculum specifically expected to see benefits on these measures. There were no positive effects for *Tools* on any of the outcomes.

The lack of expected curriculum effects required careful consideration and raised more general questions about how curriculum experiences manifest themselves in assessed skills. As a first step to understanding the findings, we focused on teachers who were implementing *Tools* and examined the degree to which the curriculum was delivered as intended and the relations between fidelity of implementation and children's outcomes in prekindergarten. Results indicated a wide variation in observed fidelity of implementation but no consistent associations between fidelity of implementation and any child outcomes.

In terms of more general practices and interactions associated with positive student outcomes, developers of the curriculum hypothesized that implementing *Tools* would enhance classroom practices and teacher–child interactions. Among the aspects they expected to be affected were the amount of non-instructional behaviors, teacher-led and child-directed activities, teacher and child talk, social learning interactions, classroom emotional climate, quality of teacher instruction, and children's level of involvement. Teachers varied as much within treatment and control classrooms as they did between conditions on most of the aspects examined. We found no differences between experimental conditions on most practices and interactions.

Curricula vary in scope and content, but they are universally intended to change classroom processes in ways that in turn will facilitate the development of targeted skills. For this mediational hypothesis to hold, the targeted classroom processes must be associated with child outcomes. We examined the associations between the classroom processes and children's prekindergarten and kindergarten gains and found support for their importance in early childhood classrooms. These findings demonstrate the value of identifying strategies to enhance these classroom practices and interactions.

We situate the findings of our study within the larger context of early childhood education expansion policies and practices, and we offer a set of lessons learned. The study we report is a single evaluation of a single curriculum, yet we hold that the lessons learned are general and shed light on understanding why evaluations of curriculum have yielded such mixed results.

I. Curriculum as a Tool to Improve Prekindergarten Quality: Introduction and Literature Review

International comparisons of academic achievement place the United States in the middle, below many other industrialized countries; the "abysmal educational attainment and test score performance of many disadvantaged students" (Ladd, 2012, p. 204) is held to be at least partly to blame. The connection between educational achievement and family income grew substantially stronger in the United States beginning in the 1980s (Chmielewski & Reardon, 2016). The skill gaps between children from high- and low-income families are evident at kindergarten entry and grow larger over the school years (Bradbury et al., 2018; Valentino, 2017). This finding led many to assert the importance of the experiences young children from poor families have in their early years before formal school entry (Mackey et al., 2015). Enriched early education experiences are proposed as a potential solution (Ladd, 2012).

Young children's participation in early childhood care and education has immediate positive effects on children's cognitive, social-emotional, self-regulation, and achievement outcomes (Burchinal et al., 2016; Keys et al., 2013; Lipsey et al., 2018; Yoshikawa et al., 2013). Most of this evaluation research examined the effects of prekindergarten programs, ones that tend to focus largely on academic learning. With the exception of the study by Lipsey et al., the studies cited above are quasi-experimental. The extensive randomized control trial evaluating Head Start found only a few immediate, positive effects from the program (Puma et al., 2012).

Although most of the research on the impacts of early child care and education has focused narrowly on the United States, a growing body of research supports positive associations between early childhood educational experiences and positive child outcomes in other countries, especially in ones that also show an income-achievement gap such as Chile (Leyva et al., 2015) and Portugal (Abreu-Lima et al., 2013). Advocacy for economic and social investment in early childhood care and education to support future children's growth and well-being now exists on an international scale. Adopted by all United Nations Member States, the United Nations' 2030 Agenda for Sustainable Development (United Nations General Assembly, 2015) set the goal that all children will "have access to quality early childhood development, care and pre-primary education so that they are ready for primary education" (p. 17).

The goal of supporting access to quality early childhood care and education emphasizes the need to define quality by identifying the strategies and approaches that contribute to positive outcomes for children. Recently in the United States, the provision of an intentional, scripted curriculum was proposed as one of the strategies, perhaps the most important one (Yoshikawa et al., 2013). Our intention in this monograph is to provide a fully integrated and comprehensive account of a large-scale, longitudinal, field-based randomized control trial of the *Tools of the Mind* prekindergarten curriculum that occurred in the United States. The goal of the work is not only to report potential impacts of the curricula to improve prekindergarten quality and children's academic, executive function, self-regulation, and social outcomes; it is also to expand understanding of whether early childhood education curricula, in general, can provide the desired short- and long-term positive effects for children.

Early Education and School Readiness in the United States

In 2017, about 69% of the 4-year-old children in the United States participated in some form of early childhood education center-based program, which includes public-funded programs such as Head Start, state- and/or Title I-funded prekindergarten, and non-profit and for-profit child care centers (Institute of Education Sciences, 2019). Over the past 20 years, the number of children in center-based programs increased, and they also shifted in where they received care, with many states now educating a large portion of their 4-year-olds in state-funded prekindergarten programs (Friedman-Krauss et al., 2018).

Most U.S. states currently offer some form of voluntary prekindergarten for children from low-income families, with Florida, Georgia, and Oklahoma offering universal prekindergarten for all 4-year-old children. Among 3- to 5-year-olds enrolled in preschool programs, the percentage attending full-day programs increased from 47% in 2000 to 56% in 2017 (McFarland et al., 2019). As in years past, higher-income families enrolled their children in center-based care more frequently than lower-income families. Among higher-income families, center-based programs tended to be privately operated and focused on providing care for children whose parents work or are in school or whose parents want their children to have a socialization experience before formal school. These programs are year-round and offer extended hours of service. Children from lower-income families are more likely enrolled in publicly funded programs such as Head Start and, more recently, state-funded prekindergarten programs. These programs operate 3–6 hours a day, 9 months of the year, essentially operating under a school-year calendar.

Publicly funded programs for the most part target low-income families, funded as compensatory education experiences with the immediate goal of improving school readiness and the longer-term expectation of closing the achievement gap between children from poor and higher income families (Farran & Nesbitt, 2019; Halpern, 2013). Participation in formal prekindergarten improves some aspects of school readiness at kindergarten entry (Duncan & Magnuson, 2013; Gormley et al., 2005), but longer-term effects are mixed and a matter of some debate (Lipsey et al., 2018; Phillips et al., 2017). Although prekindergarten programs may improve basic pre-reading skills, their influence on complex language skills, mathematics, self-regulation, and social skills is less clear (Gormley et al., 2005; U.S. Department of Health and Human Services, 2005, 2010).

Given these mixed results, policymakers and advocates issued recommendations for improving the quality of prekindergarten and Head Start programs to obtain stronger and more lasting effects. The adoption of an evidence-based curriculum is at the top of many of the recommendations. For example, the first pillar of Head Start's *Framework for Effective Practice* is the implementation of a "developmentally appropriate research-based curricula for group care settings" (https://eclkc.ohs.acf.hhs.gov/teaching-practices/article/framework-effective-practice). The second of the 10 standards for accreditation by the National Association for the Education of Young Children (NAEYC) is the implementation of a curriculum (https://www.naeyc.org/accreditation/early-learning/interested). In 2017, the New America Foundation listed a quality curriculum as one of the indispensable practices for high-quality prekindergarten (Sharpe et al., 2017).

This focus on curriculum is relatively new to early childhood education. Well into the 1960s, a maturationist theory of development dominated early childhood education (Saracho, 2015). Further, early childhood educators believed that children might become frustrated and even turned off to school if they were provided with too much instruction before they were ready. For example, in 2005, Bennett argued against specified curricula for early childhood programs:

> In the early childhood context, the traditional sense of curriculum is inappropriate, namely, *a plan of instructional activities or lesson plans to be carried out by staff in order to inculcate skills and/or pre-defined subject content.* From an early childhood perspective, too much place is given in this definition to content, and the underlying methodology is unsuited to young children's manner of learning (p. 7, author italics).

Moreover, Murray (2015) argued that traditional principles and values determined early childhood pedagogy and the involvement of the government in supporting early childhood education as an intervention for children from poor families led to a *schoolification* of the field. The schoolification focus is reflected in the New America Foundation's assertion that preschool should have an equal footing with other grades in public education and that curricula play a key role in gaining that footing (Sharpe et al., 2017).

Early Childhood Education: The United States in Context

Countries around the world perceive early childhood education and care (ECEC) differently, so differently that it can be difficult to find commonalities or even to have a discussion. Many countries adopted national formal curriculum frameworks that guide the experiences for young children (McLachlan et al., 2010). Hong Kong has the Guide to Pre-Primary curriculum, New Zealand has Te Whāriki, the United Kingdom has the Early Years Foundation Stage. Other national frameworks like that seen in Sweden are less specific but instead embrace important values and goals more broadly. Einarsdottir and Wagner (2006) describe the main principles in the approach to early childhood by the Nordic countries (Denmark, Sweden, Norway, Iceland, and Finland):

> *Child and family policies are based on Nordic ideology and traditions, emphasizing democracy, equality, freedom and emancipation, solidarity through cooperation and compromise, and a general concept of the "good childhood," or what life should be like for all children… Nordic people generally view childhood as important in its own right, not simply [as] a platform from which to become an adult* (pp. 4, 6).

The distinction is between early childhood as a state of being rather than a state of becoming. In a similar vein, the Organization for Economic Cooperation and Development (OECD) asserted that there were two main early childhood traditions, the Nordic (with a focus on the child as an active constructor, or learner-centered) and the Anglo-Saxon (with a focus on preparing the child, or pre-primary) (Brodin & Renblad, 2014). Using different terms, Bernstein (1996) contrasted the *competence* model, which gives children some control over the pacing and content of what is being learned with the *performance* model with its emphasis on knowledge of subject matter. These dualities frequently conflict with each other and push ECEC policies in different directions over time. For example, New Zealand, Sweden, and Finland recently moved toward including more overt educational goals in their early childhood frameworks (Alcock & Haggerty, 2013; Cochran, 2011; Fonsén et al., 2019). In Canada, however, ECEC policies have moved in the opposite direction (Brodin & Renblad, 2014).

Similar conflicts can be found in the ways that U.S. educators approach early education and care, but they are not reflected in a national policy toward ECEC. The United States has a fragmented ECEC system with no agreed-upon definition of quality, perhaps because there is no universally agreed-upon perspective toward the early childhood period (Kamerman &

Gatenio-Gabel, 2007). There was one attempt in the late 1960s to create a unified set of policies for what should constitute quality care for U.S. children. It, however, was opposed so strongly by both the political left and right that it was abandoned (Zigler et al., 2009). The United States has not reconciled the two major functions of care and education in early childhood (Farran & Nesbitt, 2019; Kamerman & Gatenio-Gabel, 2007). The debates about the focus of various alternative curricula reflect the lack of reconciliation about the intended function of U.S. early childhood education programs.

Early Childhood Curricula in the United States

As outlined above, unlike many other developed countries, the United States has no early childhood national framework or uniform vision. Given the fractured early childhood system and the lack of vision, U.S. early childhood curricula take many different forms. The most rigorous and comprehensive evaluation of prekindergarten curricula to date in the United States is the Preschool Curriculum Evaluation Project (PCER) which launched 14 randomized trials of different curricula around the country (Preschool Curriculum Evaluation Research Consortium, 2008).

Using the curricula included in PCER, Jenkins et al. (2018) proposed placing them into two broad categories: (1) global, developmental and (2) skill-specific or academically oriented. (These distinctions mirror the competence/performance distinctions outlined before.) Jenkins et al. characterized global curricula like *Creative Curriculum* and *High Scope* as focusing on a holistic view of development that involves a comprehensive set of domains including social/emotional, physical, language, and cognition. Global curricula also tend to focus on project-based investigations and discovery learning and much less on didactic, teacher-led instruction. On the other hand, skill-specific curricula target particular academic content areas such as literacy (e.g., *Literacy Express, Opening the World of Learning*), mathematics (e.g., *Building Blocks for Math, Pre-K Mathematics*), or self-regulation (*Tools of the Mind*).

Other early childhood theorists proposed similar distinctions among types of curricula. For example, Bennett (2005) characterized curriculum differences as being between a social-pedagogic approach and a pre-primary approach. Wood and Hedges (2016) argued for a distinction between domain-specific—the traditional focus of early childhood education (social, emotional, cognitive)—and discipline-specific (e.g., literacy, numeracy)—a newly emerging focus. In a review of curricular effects on literacy outcomes, Chambers et al. (2016) created a dichotomy similar to Jenkins et al. (2018) using the terms *developmental-constructivist* in place of *global* and *comprehensive* in place of *skill-specific*. The term comprehensive may indeed be more appropriate because it captures broad-based curricula like *Tools of the Mind* that explicitly focus on all aspects of children's development and teach specific skills in deliberate and planful ways.

Although global curricula continue to be the most frequently used in Head Start and prekindergarten settings (Nguyen et al., 2019), the push for the inclusion of more academic content and/or more academically focused curricula is growing (Chambers et al., 2016; Wood & Hedges, 2016). Yoshikawa et al. (2013) characterize content-based curricula that are intentionally focused on academic areas (e.g., literacy and mathematics) as the strongest hope for improving outcomes in early childhood education.

Even within these broad groupings, curricula differ in many aspects, some of which instantiate a profoundly different vision for how children learn and for the role of the teacher. In 2015, the National Center on Quality Teaching and Learning (NCQTL; 2015) of the Office of Head Start created an extensive review of available preschool curricula. The 16 curricula included in the report differed a great deal from one another, ranging from *Creative Curriculum,* a global, developmental-constructivist approach, to *Let's Begin with Letter People*, a highly scripted literacy-focused curriculum. The report evaluated curricula on 13 dimensions, one of which was whether there was research evidence of effectiveness. Most curricula were rated as having no evidence or minimal evidence. Only *Opening the World of Learning* was rated as having solid evidence of effectiveness. The rating for *Tools of the Mind* indicated partial evidence of effectiveness.

At present, there is scant rigorous evidence indicating that different prekindergarten curricula produce significantly different effects for children (Jenkins et al., 2019). For example, most of the curricula tested in the PCER evaluation had a literacy or general developmental focus (with one focused on math). Overall, compared with business as usual, 10 of these curricula showed no statistically significant impacts at the end of the preschool year on any of the student-level outcomes of reading, phonological awareness, language, or mathematics. None of the prekindergarten curricula had statistically significant positive impacts on social skills or problem behaviors. No curriculum outperformed the control classrooms on all child outcomes; only two showed significant differences on even one skill measured in kindergarten. The report concluded that generally, no curriculum stood out as notably more effective than any of the others (Preschool Curriculum Evaluation Research Consortium, 2008).

Jenkins and colleagues reviewed the PCER results as well as curriculum evaluations from several other large-scale preschool and Head Start studies (Jenkins et al., 2019). They found that curricula differed from each other in terms of the number of math and literacy activities, as well as ratings on the Early Childhood Education Rating System (ECERS) and the Classroom Assessment Scoring System (CLASS). The differences, however, were not consistently found in all the classrooms reporting using the same curriculum. Moreover, none of the curricula had consistent effects on children's school readiness that were different from any of the others. Jenkins et al. characterize these findings as showing "distinctions without differences."

Despite the lack of evidence differentiating specific curricular effects, most school systems require the implementation of an identified curriculum in their prekindergarten classrooms. Slavin observed that district administrators rarely use research evidence as a basis for the curriculum chosen: "If they do consider evidence, it is often to ask whether a given program is based on accepted principles rather than whether the program itself has been evaluated in comparison with a *control group*" (Slavin, 2020, p. 21, author italics). State-funded prekindergarten programs often provide a list of approved curricula from which districts are to choose. A currently popular prekindergarten curriculum on the lists is *Tools of the Mind* (Bodrova & Leong, 2007), one of a few curricula recommended for facilitating self-regulation as well as academic skills (Diamond & Lee, 2011; Hughes, 2011). *Tools* is on the NCQTL list of recommended curriculum choices for Head Start. As we outline below, there are reasons to believe that a curriculum focused on executive function and self-regulation would produce generally positive results in many areas of development.

Executive Function and Self-Regulation as a Curriculum Focus

Successful transition into formal schooling for young children and subsequent academic success requires a variety of competencies, including most obviously the early literacy and numeracy skills that provide the foundation for reading and mathematics. Another critical area of competency is the ability to engage in and benefit from the kinds of learning tasks intrinsic to school-based instruction, including attending to speech that conveys information, completing exercises that require planning, problem solving, application of knowledge, practicing acquired skills, and remembering and following rules and instructions (Cooper & Farran, 1988; Howse et al., 2003; McClelland & Morrison, 2003). These latter skills enable children to focus on and benefit from the educational material and learning opportunities provided in school settings (Blair & Razza, 2007). Longitudinal studies demonstrated that these self-regulation skills have an independent relation with long-term academic success, separate from early academic skills (e.g., Duncan et al., 2007; Fuhs et al., 2014; Moffitt et al., 2011; Schmitt et al., 2017; Spivak & Farran, 2016).

Self-regulation is related to the neurocognitive concept of executive function. Executive function is an umbrella term that refers to goal-directed cognitive skills such as the modulation of attention in response to changing demands, the active manipulation of information in the mind, and the inhibition of interfering information, all of which are essential for adapting to a formal school environment (Hughes, 2011). For the current study, the term executive function refers specifically to certain cognitive skills that include tasks like inhibitory control, the ability to shift attention and not perseverate, and working memory. Self-regulation, on the other hand, involves control of one's behaviors and emotions in a social context, including the context of an early childhood classroom (Jones et al., 2016). Clearly, these two sets of competencies are related to each other and likely develop in tandem. We will use both terms throughout the paper. When discussing learning and cognition, we will use the term executive function. When discussing behaviors, we will use the term self-regulation.

Research indicates that both self-regulation and executive function skills show rapid improvement in the preschool years (Carlson, 2005; Garon et al., 2008), but children from low-income homes often lag behind their peers from middle- to high-income families in both types of skills (e.g., Howse et al., 2003; Noble et al., 2007, 2015). There is, unfortunately, less understanding of exactly how young children, especially those who come from poverty, develop self-regulation and executive function skills to be successful in school.

Current educational research suggests that self-regulation and executive function skills are critically correlated with the development of early academic skills (e.g., Bull et al., 2011; Fuhs et al., 2014; Welsh et al., 2010) as well as other positive life outcomes (e.g., Caspi et al., 1998; Kern & Friedman, 2009; Moffitt et al., 2011). Thus, promoting these skills is identified as a potentially fruitful target for intervention for children who are at-risk for academic failure (Ursache et al., 2012). Emerging research suggests that not only are executive function skills at school-entry important for the development of academic skills, but growth in executive function skills may be associated with growth in academic skills (e.g., Fuhs et al., 2014; McClelland et al., 2007; Schmitt et al., 2017; Welsh et al., 2010). The issue discussed next is whether those skills are susceptible to intervention.

Curricula Targeting Executive Function and Self-Regulation

Correlations between measures of executive function and school readiness and indications that executive function skills may be affected by prekindergarten experiences (e.g., Fuhs et al., 2013; Raver et al., 2011) have led to the development of interventions and curricula to support these skills in young children. As summarized in this section, there is a growing body of work evaluating the efficacy of different approaches which yields promising, yet inconsistent findings. Most of these approaches focus on self-regulation and the control of behaviors and not on learning-related skills associated with executive function. Many of these efforts are not comprehensive curricula but are instead add-on sets of activities and practices to be incorporated into whatever general curriculum the classroom is using. The specific components of the various approaches provide interesting information about what the curriculum developers believe to be the foundation for the development of the skills.

One intervention approach for encouraging children's self-regulation skills focuses on improving the quality of interactions among teachers and children as a means of making classrooms more supportive and responsive. The *Incredible Years* (Webster-Stratton et al., 2001) is an add-on curriculum that utilizes this approach by focusing on the improvement of children's social-emotional behaviors (e.g., helping children understand feelings, get along with friends, learn anger management and problem solving, and learning how to behave at school).

The efficacy of the *Incredible Years* curriculum to support children's executive function and academic skills was evaluated as part of the *Chicago School Readiness Project* (*CSRP*; Raver et al., 2008). The CSRP was a randomized control trial of 35 Head Start classrooms. The add-on intervention provided various kinds of training to teachers who had been assigned to the training condition. In particular, these teachers received training on behavior management strategies adapted from the *Incredible Years* curriculum, mental health consultations, and were provided with support for children in their classrooms who were exhibiting high-levels of disruptive behaviors. The evaluation found significant positive effects on one measure of inhibitory control (Peg Tapping) and one measure of self-regulation (Balance Beam). Moreover, *CSRP* found significant positive effects on the academic skills of vocabulary, letter knowledge, and mathematics. The latter were largely mediated by improvements in a measure of self-regulation (i.e., a composite of Peg Tapping scores and assessors' ratings of the child's self-regulation during the assessment sessions, see Raver et al., 2011).

Incredible Years was also evaluated as part of the Head Start CARES project (Classroom-based Approaches and Resources for Emotion and Social skill promotion; Morris et al., 2014). Although the randomized control trial comparing Head Start classrooms implementing *Incredible Years* to business-as-usual classrooms did not find or replicate the *CSPR* greater gains on Peg Tapping, the study did find positive effects of treatment on teachers' ratings of children's self-regulation (social-emotional and classroom work-related behaviors).

Similar to the *Incredible Years*, the *Preschool Promoting Alternative Thinking Strategies* (*Preschool PATHS*) curriculum focuses on social-emotional learning by targeting the development of conflict resolution, emotion regulation, empathy, and decision-making skills (Domitrovich et al., 1999), all believed to be components of self-regulation. The *PATHS* developers posit that through these mechanisms, the curriculum will have positive effects on executive function skills. *Head Start REDI* (Research-based, Developmentally Informed; Bierman et al., 2008) and *Head Start CARES* (Morris et al., 2014) evaluated the impact of *Preschool PATHS* on children's executive function and self-regulation skills. Both *Head Start REDI* and *CARES* were large-scale randomized control trials comparing Head Start classrooms receiving training in *Preschool PATHS* to business-as-usual Head Start classrooms. *Head Start REDI* provided teachers with training on *Preschool PATHS* and dialogic reading and found positive effects for intervention classrooms for children's ability to flexibly switch their attention (measured by the Dimensional Change Card Sort [DCCS] task) and for self-regulation skills (measured by assessors' ratings). However, effects for other executive function and self-regulation skills (measured by Peg Tapping, Backward Word Span, and Balance Beam) were not found (Bierman et al., 2008).

Head Start CARES evaluated the impacts of *Preschool PATHS* enhancement as an add-on to the regular curriculum and found positive effects for teacher ratings of learning-related behaviors (measured by the Cooper-Farran Behavioral Rating Scale), social behaviors (Social Skills Rating Scale), emotion knowledge (Facial Emotions Task), and social problem solving (Challenging Situations Task). It is hard to draw strong conclusions about the efficacy of *Preschool PATHS* given the variations across studies in the intervention, types of assessments, and children's outcomes.

Second Step Early Learning Program (*SSEL*; Committee for Children, 2011, also referred to as *Second Step Social-Emotional Learning*) is another widely used add-on curriculum focused on emotion and behavior regulation. *SSEL*'s approach is focused on social-emotional learning. It provides 28 weeks of 5-min lessons to support children's skills in learning, empathy, emotion, management, and social skills with a goal of preparing children to listen, pay attention, manage behavior, and get along together. Randomized control trials of *SSEL* have found positive effects for Head Start children's gains in executive function (as measured by a composite of three executive function assessments; Upshur et al., 2017; Wenz-Gross et al., 2018).

Finally, an alternative approach to increase children's self-regulation is a shorter-term (8 weeks) intervention which focuses on having children participate in music and movement games in small groups. Activities target children's ability to regulate their behavior, for example, in a freeze game and in the game, red light, green light. In randomized control trials, the *Red Light, Purple Light Circle Time Games* intervention (*RLPL*) demonstrated positive gains in behavioral self-regulation among children enrolled in Head Start (McClelland et al., 2019; Schmitt et al., 2015), especially for children with lower entering skills (Tominey & McClelland, 2011). Compared with control children in business-as-usual classrooms, children participating in the *RLPL* intervention made greater gains on the Head-Toes-Knees-Shoulder (HTKS; McClelland et al., 2019; Schmitt et al., 2015; Tominey & McClelland, 2011) and the DCCS (Schmitt et al., 2015). To date *RLPL* has been investigated by only its developers; future research by others seems warranted.

Tools of the Mind Curriculum

As we indicated, the curricula just described are add-ons or insertion curricula which in some way expand the ongoing primary curriculum. Teachers are asked to implement these add-on activities for 5–10 min during the day or for part of the year. *Tools of the Mind* (*Tools*, Bodrova & Leong, 2007), on the other hand, is a complete curriculum with multiple activities focused on math, drama, and literacy, each of which also contains an executive function aspect. Tools has a Vygotskian framework and focuses on equipping children with cognitive tools for learning that they can then apply to the task of acquiring and sustaining academic knowledge and skills. These skills are commonly referred to as executive function skills, that is, skills associated with higher-order cognitive thinking that facilitate planning and goal-directed behavior. Tools aims to improve these skills by providing frequent, structured opportunities for children to practice regulating their cognition, behaviors, and emotions in the classroom context. This approach follows from a socio-cultural perspective on child development that emphasizes how children acquire skills and cultural tools (e.g., spoken and written language, pretend play, the use of numbers, diagrams, and maps) in collaboration with knowledgeable others (e.g., Behne et al., 2008; Bodrova & Leong, 2017; Rogoff et al., 2005).

In the *Tools* approach, teachers model and use tactics such as concrete mediators (e.g., pictures or symbols), language (both speech and writing), and shared activities to scaffold children's learning. In the *Tools* approach, however, the tactics, mediators, forms of talk, and activities teachers use are designed to be part of what the student learns. The instructional tactics begin as external supports for behavior and mental activities (like memory) and then guide children to use these activities internally. *Tools* emphasizes that teachers use scaffolding techniques to help children internalize the learning tools at the center of the curriculum. That is, children use the mediators introduced by the teacher and then create their own. Children apply self-talk and writing and use shared activities and dramatic play in ways that help them attend, self-monitor, solve problems, plan, and remember. Because the theory of change behind *Tools* is so clearly articulated, it facilitates research that can examine the specific behaviors the curriculum is designed to affect in the classroom.

First implemented in prekindergarten classrooms in 1993, *Tools* has been substantially revised on the basis of field experience over the past 25 years. A kindergarten version has since been developed (Blair & Raver, 2014; Blair et al., 2018). As a prekindergarten curriculum, the focus grew from 40 original activities to 60 or more scripted Vygotskian-based activities designed to promote children's self-regulatory skills and cognitive development.

A primary focus of *Tools* is the facilitation of mature pretend play. Children are supported through a daily play planning activity to identify their social role or character for pretend play (e.g., being the doctor at the hospital play center). The goal is that eventually this pre-planning would include formulating what the character might say and do and how the character would interact with other characters in the play scenario. Working with the teacher or a teacher's assistant in small groups, children record the plan for their play with drawings, marks, letter-like forms, and words. Children must later adhere to selected roles, and the curriculum encourages teachers to use recorded play plans to remind children of their roles and to encourage children to assist each other in the maintenance of roles. The play planning is focused on helping children be purposeful in their play, rather than reactive. The plan serves as a mediator of their later behavior.

Tools consists of an array of activities explicitly to support the development of children's literacy, mathematics, science, and self-regulation skills. Activities such as paired buddy reading, practice in drawing shapes and graphemes, and a dynamic array of storybook reading approaches support literacy development. Calendar and weather graphing, puzzles and manipulatives, making collections and patterning, and games to foster knowledge of numbers, colors, shapes, and science support the development of mathematics. A unique aspect of *Tools'* academically focused activities is the deliberate integration of elements that require children to learn through the Vygotskian (1987) principle of the "the transition from interindividual (intermental) or shared to individual (intramental)" (Bodrova & Leong, 2015, p. 373). An example is having children check each other's work in partnered activities. In addition, the *Tools* curriculum contains an array of activities with the primary focus of developing executive function and self-regulation skills, including activities to focus attention and engage in regulating one's fine and gross-motor skills, as well as a variety of freeze games and turn-taking activities.

Tools is similar to a constructivist curriculum in room arrangement, materials, and a balance among whole group, small group, and center-based activities. However, *Tools* differs from other constructivist approaches in the prescriptive and intentional role of the teacher in the classroom. Thus, Chambers et al. (2016) characterized it as comprehensive, not constructivist. The teacher's role is specifically prescribed for each major type of activity during the day (e.g., morning meeting, storybook reading, center-based time) through a series of delineated *steps* to be followed. There is a schedule for each day; some activities are enacted daily whereas others occur twice a week and alternate with a paired activity. This organization of the *Tools* curriculum, therefore, reflects Weiland et al. (2018) goals for early childhood curricula—to contain specific instructional content and to be intentional and highly scripted.

Tools is not a curriculum that can be taken off the shelf and implemented. Effective use of the curriculum depends on the depth of teacher understanding of socio-cultural principles of children's learning and development and a reconceptualization of the teacher's role in facilitating children's development. For example, the *Tools* calendar is different from the usual matrix calendar, requiring teachers to create a linear reflection of the days posted around the walls of the classroom. Similarly, the alphabet is not taught sequentially in *Tools*. Rather letters are grouped conceptually and taught in clusters. The *Tools* developers therefore strongly recommend 2 years of professional development workshops together with in-classroom coaching, and their training packet is set up for this level of teacher contact.

The concepts behind the *Tools* approach are appealing to early childhood educators, especially those who have been concerned about the didactic nature of many early childhood classrooms (Hirsh-Pasek & Golinkoff, 2003; Miller & Almon, 2009). Until a few years ago there was only one small study of the effectiveness of the curriculum described by Barnett et al. (2008) and Diamond et al. (2007). Yet *Tools* received enormous attention in the popular press, featured in the *New York Times*, the *Wall Street Journal*, on National Public Radio, and a popular press book by Tough (2012), to name a few. School systems in Washington, DC, New Jersey, Chicago, and the entire country of Chile received training from *Tools* staff to implement the approach.

Other Evaluations of *Tools of the Mind*

Concurrent with and then subsequent to our evaluation of the curriculum described in detail in the current monograph, eight separate randomized control trials evaluated the effect of either the full *Tools* curriculum or sections of it on the development of preschool and kindergarten children. Table 1 presents a summary of these eight studies (note that some studies have multiple publications). As is evident from the table entries, the studies varied widely with regard to the grade-level targeted (preschool vs. kindergarten), the extent to which *Tools* was implemented (e.g., full curriculum vs. a pretend-play add-on to another curriculum), the characteristics of children, classrooms, and schools (e.g., socioeconomic status, ethnicity, primary or home language, geographic location), methodological details (e.g., sample sizes, characteristics of the counterfactual condition(s), outcomes assessed, documentation of fidelity of implementation), and in the dissemination of results (peer-review publication or conference presentation).

With such variation in study design, it is not surprising that the reported impacts of *Tools* on child outcomes have been mixed and have provided only minimal evidence of effects which extend beyond the intervention year (e.g., a significant prekindergarten effect that extends into kindergarten and first grade). In fact, the inconsistency in findings, as summarized below, led the Head Start Early Childhood Learning and Knowledge Center to provide a rating of *Minimal Evidence* for the indicator of *Evidence Base for Child Outcomes* (https://eclkc.ohs.acf.hhs.gov/curriculum/consumer-report/curricula/tools-mind). Although such variability in outcomes makes answering the key question regarding the effectiveness of *Tools* difficult to answer, a systematic examination

TABLE 1
Study Methodology of Previous Randomized Control Trials of *Tools of the Mind*

Publication/ Reference	Study Features	Participant Demographics	*Tools* Training	Child/Teacher Outcomes[a]	Fidelity and Classroom Quality Measures[a]
Published evaluations of Tools preschool curriculum					
Barnett et al. (2008)	• Randomization: 16 mixed-aged preschool classrooms from 1 school in New Jersey randomly assigned to *Tools* (7) or control (9) • Counterfactual: School-district developed balanced literacy curriculum	• Children age 3 and 4 at onset of school year (88 *Tools*, 122 control) • 80% of district qualify for FRPL • 93% Hispanic or Latinx • 69% Spanish primary home language • 100% of teachers with at least a bachelor's degree	• All training provided by the *Tools* organization • One 4-day workshop before school year • Three 1-day workshops during school year • Staff visits every 6 weeks	• Assessed in Spanish or English at beginning and end of school year (except SSRS) • Academics: WJ AP, WJ LWI, PPVT, EOWPVT, GRTR, OLPT • SR/EF: WIPPSI Animal Pegs • Social-emotional: SSRS Problem Behaviors[b]	• *Tools*-Specific Fidelity: 50-item environmental features[b] • General Classroom Observation: ECERS[b], SELA, PCI[b], CLASS[b]
Diamond et al. (2007)	• Year 2 follow-up to Barnett et al. (2008)	• Children in second year of preschool (mean age at end of year = 5.1) • 62 Control (1 or 2 years of control curriculum), 85 *Tools* (22 with 1 year of *Tools*, 63 with 2 years of *Tools*)	• See Barnett et al. (2008)	• Assessed at end of school year only • SR/EF: Dots Task[a], Flanker Task[a]	• See Barnett et al. (2008)
Solomon et al. (2018)	• Randomization: 20 YMCA Canada urban (Ontario) daycare centers randomly assigned to *Tools* (10) and control (10) • Counterfactual: YMCA Playing to Learn Curriculum which does not target self-regulation	• Children age 3 and 4 at onset of school year (148 *Tools*, 108 control randomized) • 54% of children in programs receive fee subsidy	• 5 workshops from *Tools* organization spread over the course of the evaluation window • Site visits following workshops • Training on core *Tools* activities	• Assessed at beginning and end of school year • SR/EF: Day/ Night, HTT • Social-Emotional: SDQ, SCBE	• *Tools*-Specific Fidelity: *Tools* Implementation Checklist for 21 *Tools* activities

(*Continued*)

TABLE 1. (*Continued*)

Publication/ Reference	Study Features	Participant Demographics	*Tools* Training	Child/Teacher Outcomes[a]	Fidelity and Classroom Quality Measures[a]
Published evaluations of add-on Tools pretend play and SR activities					
Clements et al. (2020)	• Evaluation of *Tools* pretend play and SR enhancements added to Building Blocks PreK Math Curriculum (BB) • Randomization: 84 full and half-day preschool classrooms across three districts into three conditions (25 BB only, 30 BB w/*Tools*, and 29 BAU) • Counterfactual: BAU and BB only. BAU used standard district practices and curricula, including use of published curricula. • Randomization: 84 full and half-day preschool classrooms across three districts into three conditions (25 BB only, 30 BB w/*Tools*, and 29 BAU) • Counterfactual: BAU and BB only. BAU used standard district practices and curricula, including use of published curricula	• Children age 4 at onset of study (365 BAU control, 391 *Tools*) • 39% Latinx, 18% Asian Pacific Islander, 11% Black, 31% White • 27% English Language Learners • 86% of teachers with at least a bachelor's degree	• Training on BB plus 6 days of training in each of the 2 years of implementation on *Tools* by *Tools* staff	• Assessed at pretest, end of study year, and end of K • Evaluation of year 2 of *Tools* implementation • Academic: TEAM, ECLS PreK/K Math, PPVT, EVT, RBS, PALS • SR/EF: HTKS, Peg Tapping, Digit Span	• *Tools*-Specific Fidelity: MPOT[b] • General Classroom Observation: CLASS, COMET, • Teacher Reports: Perceptions of implementation experience

(*Continued*)

TABLE 1. (*Continued*)

Publication/ Reference	Study Features	Participant Demographics	*Tools* Training	Child/Teacher Outcomes[a]	Fidelity and Classroom Quality Measures[a]
Hsueh et al. (2014)	• Follow-up to Morris et al. (2014) • Evaluation of *Tools* pretend play and AR enhancement on 3-year-olds • Randomization: 56 (155 classrooms) Head Start centers into four conditions (IY, PATHS, 37 *Tools* classrooms, 40 BAU classrooms) across United States • Counterfactual: BAU Head Start program	• Children age 3 ($M = 3.47$ at pretest) in mixed age classrooms serving children age 3 and 4 • 220 BAU control and 241 *Tools*	• Training of *Tools* enhancement in summer before evaluation year	• Assessed at pretest and end of study year • Examination of teacher ratings only • Academic: ARS • SR/EF: BPI, CFBRS WRS • Social-Emotional: CFBRS IS, SSRS, STRS	• See Morris et al. (2014)
Morris et al. (2014)	• Evaluation of *Tools* pretend play and SR enhancements • Randomization: 104 (307 classrooms) Head Start centers into four conditions (IY, PATHS, 76 *Tools* classrooms, 77 BAU classrooms) from across United States • Counterfactual: BAU Head Start program	• Children age 4 ($M = 4.42$ at pretest) • 621 BAU control and 678 *Tools* • ~9 children per classroom participated • 43% Latinx, 33% Black, 16% White • 59% receive food stamps • 62% of teachers with at least a bachelor's degree	• Training of *Tools* enhancement in summer before evaluation year	• Assessed at pretest and end of study year • Academic: WJ AP, WJ LWI, EOWPVT, ARS • SR/EF: HTKS, Pencil Tap, BPI, CFBRS WRS • Social-Emotional: EIS[a], CST, CFBRS IS, SSRS[a], STRS • Kindergarten: BPI, CFBRS WRS, SSRS	• General Classroom Observation: CLASS, TSRS[b]

(*Continued*)

TABLE 1. (*Continued*)

Publication/ Reference	Study Features	Participant Demographics	*Tools* Training	Child/Teacher Outcomes[a]	Fidelity and Classroom Quality Measures[a]
Non-published evaluations of Tools preschool curriculum (results of effects on child/teacher outcomes and classroom measures are not available)					
Hammer et al. (2012)	• No publication of study findings • Randomization: 60 Head Start and school district preschool classrooms in large urban areas in New York and Florida • Counterfactual unknown	• 7 Latino children from each preschool classroom participated in the study ($N = 420$) who were Spanish-English Language Learners (Ns by condition not available)	• Year 1: 4 days of in-service training (2 days prior to the start of school) and received coaching two times a month • Year 2: 4 days of in-service training and continuation of coaching	• Assessed at pretest, end of study year, K, and Grade 1 • Evaluation of Year 2 of *Tools* implementation • Academic: WJ AP, WJ LWI, CELF, EOWPVT • SR/EF: Flanker, DCCS, WM, CBQ, ERC • Social-Emotional: TOCA, SCS, PLBS, PIPPS	• General classroom observation: CLASS, LISn and ELLCO extension
Lonigan and Phillips (2012)	• No publication of study findings • Randomization: 117 Head Start and state preschool classrooms in New Mexico and Massachusetts to four conditions (*Literacy Express, Tools, Literacy Express Tools* Commination, BAU) • Counterfactual: BAU Head Start program	• 2,564 children age 28–73 months at onset of school year (Ns by condition not available) • 52% Latinx, 38% White non-Latino • 25% with IEP	• Not reported	• Evaluation of impact for year 1 and 2 of implementation • Assessed at pretest and end of study year • Academic: TOPEL, BBCS • SR/EF: BRIEF-P, HTKS	• Not reported

(*Continued*)

TABLE 1. (*Continued*)

Publication/ Reference	Study Features	Participant Demographics	*Tools* Training	Child/Teacher Outcomes[a]	Fidelity and Classroom Quality Measures[a]
Published evaluations of Tools kindergarten curriculum					
Blair and Raver (2014)	• Randomization: 29 schools (*Tools* = 16) from 12 school district in United States • Counterfactual: BAU with combination of commercial literacy and mathematics curricula and following of Massachusetts State Standards	• 759 total Kindergarteners (*Tools* = 443) from two consecutive year cohorts • First 6 consented children per classroom in sample • Participating schools ranged from 5% to 92% FRPL eligibility • 15% of the schools considered high-poverty • 100% of teachers with at least a bachelor's degree	• Year 1: 5 days of workshops spread across the school year and in-classroom coaching every other week • Year 2: 3 days of workshops spread across the school year and in-classroom coaching once a month	• Combined data across two cohorts • Assessed at pretest, and end of K and Grade 1 • Academics: WJ AP[a], WJ LWI[a], EOWPVT[a] • SR/EF: RCPM, EF Composite[a] (HF, Flanker, DCCS, BDS, Dot-Probe) • Additional measures: Speed of Processing[a]; stress response physiology[a]	• Not reported
Blair et al. (2018)	• Follow-up to Blair and Raver (2014)	• See Blair and Raver (2014)	• See Blair and Raver (2014)	• Combined data across two cohorts • Assessed at pretest, end of K and Grade 1 • Teacher-reported classroom behaviors: TSCRS[a], SDQ[a], ERC[a], STRS[a], SSRS	• Not reported

(*Continued*)

TABLE 1. (*Continued*)

Publication/ Reference	Study Features	Participant Demographics	*Tools* Training	Child/Teacher Outcomes[a]	Fidelity and Classroom Quality Measures[a]
Diamond et al. (2019)	• Randomization:18 classrooms (*Tools* = 9) from Vancouver and Surrey, Canada • Counterfactual: BAU following district practices • All classroom opted into implementing *Tools* before randomization	• 352 total Kindergarteners (*Tools* = 172) • 15% on subsidized lunch, 50% English as a second language, 11.5% receiving special-needs services • Most teachers with at least a bachelor's degree	• One 3-day workshop before school year • Four 1-day workshops during school year	• Assessed at beginning and end of K • Academics: DRA • SR/EF: Teacher-rated self-control • Social-emotional: Teacher-rated prosocial behaviors • Teacher outcomes: Teacher-rated feelings about teaching	• General classroom observation: Play behaviors, whole group activities and use of rewards and time-out

AP = Applied Problems; ARS = Academic Rating Scale; BAU = Business-as-Usual; BB = Building Blocks; BBCS = Bracken Basic Concept Scales; BDS = Backward Digit Span; BPI = Behavior Problems Index; BRIEF-P = Behavior Rating Inventory of Executive Function-Preschool; CBQ = Children's Behavior Questionnaire; CEFL = Clinical Evaluation of Language Fundamentals; CFBRS = Cooper-Farran Behavioral Ratings Scale; CLASS = Classroom Assessment Scoring System; COEMET = Classroom Observation of Early Mathematics—Environment and Teaching; CST = Challenging Situations Task; DCCS = Dimensional Change Card Sort; DRA = Developmental Reading Assessment; ECERS = Early Childhood Environmental Rating Scale-Revised; ECLE = Early Childhood Longitudinal Study; EIS = Emotions Identification and Situations; ELLCO = Early Language & Literacy Classroom Observation; EOWPVT = Expressive One-Word Picture Vocabulary Test; ERC = Emotion Regulation Checklist; EVT = Expressive Vocabulary Test; FRPL = Free or Reduced-Price Lunch; GRTR = Get Ready to Read; HF = Hearts and Flowers; HTKS = Head-Toes-Knees-Shoulders Task; HTT = Head Toes Task; IEP = Individualized Education Plan; IP = Interpersonal Skills; IY = Incredible Years; K = Kindergarten; LISn = Language Interaction Snapshot; LWI = Letter-Word Identification; MPOT = Mature Play Observation Tool; OLPT = IDEA Oral Language Proficiency Test; PALS=Phonological Awareness Literacy Screening for Preschoolers; PATHS = Promoting Alternative Thinking Strategies; PCI = Preschool Classroom Implementation; PIPPS = Penn Interactive Peer Play Scale; PLBS = Preschool Learning Behaviors Scale; PPVT-III = Peabody Picture Vocabulary Test; RBS = Renfrew Bus Story; RCPM = Raven Colored Progressive Matrices; SCBE = Social Competence and Behavior Evaluation; SCS=Social Competency Scale; SDQ = Strengths and Difficulties Questionnaire; SELA = Supports for Early Literacy Assessment; SR/EF = Self-Regulation and Executive Function; SSRS = Social Skills Rating System; STRS = Student-Teacher Relationship Scale; TEAM = Tools for Early Assessment in Mathematics; TOCA = Teacher Observation of Classroom Adaptation; TOPEL = Test of Preschool Early Literacy; TSCRS = Teacher Social Competence Rating Scale; TSRS = Teaching Style Rating Scale; WIPPSI = Wechsler Preschool Primary Scale of Intelligence; WJ = Woodcock–Johnson; WM = Working Memory; WRS = Work-Related Skills.

[a] Significant difference between *Tools of the Mind* and control classrooms reported by the study, all significant effects reported favored *Tools* classrooms.

[b] Not all studies tested for significant differences between *Tools* and control classrooms, differences are only reported for those studies that evaluated the effects. See Chapter I text for detailed summary of prior findings regarding child outcomes, Chapter III for prior findings regarding *Tools*-specific fidelity measures, and Chapter IV for findings regarding general classroom behaviors/activities.

of the particulars of these studies and their findings can offer important insights.

In general, the two randomized control trials of the kindergarten version of the curriculum provide evidence that *Tools* positively affects children's outcomes. The kindergarten *Tools* is different from the prekindergarten program, including play based on fantasy themes, the pairing of children with rotating classmate study buddies, and weekly one-on-one learning conferences. Compared with business-as-usual kindergarten classrooms, children in *Tools* classrooms made greater gains in literacy, vocabulary, and mathematics during kindergarten (Blair & Raver, 2014) or had higher postkindergarten literacy skills (Diamond et al., 2019). Curriculum differences were not found for teachers' ratings of academic competence (Blair et al., 2018) nor postkindergarten mathematics in a recent evaluation in Canada (Diamond et al., 2019). Academic benefits for children who were studied through first grade were seen in only the ability to sight-identify words (Blair & Raver, 2014).

Evaluations of the kindergarten version of *Tools* found greater gains on direct assessments of executive function (Blair & Raver, 2014). Kindergarten teachers in *Tools* classrooms provided more positive ratings of children's ability to get back to work (Diamond et al., 2019), as well as children's self- and emotion regulation, and student–teacher relationships (Blair et al., 2018). Moreover, across both kindergarten randomized control trials, *Tools* kindergarten teachers reported fewer problem behaviors (Blair et al., 2018; Diamond et al., 2019) compared with control teachers. Whereas first-grade teachers reported fewer aggression and conduct problems for *Tools* children, the teacher did not rate children from kindergarten *Tools* classrooms more positively in self- and emotion regulation or student–teacher relationships (Blair et al., 2018).

Findings of significant positive effects for *Tools* over control classrooms have been less prevalent in evaluations of the prekindergarten version of the curriculum (i.e., the curriculum for children ages 3 and 4). Table 1 reports findings from the six randomized control trials (yielding eight publications) of the preschool partial or full *Tools,* not including the research reported in this monograph. Of these six randomized control trials, two published peer-reviewed findings of *Tools* as a stand-alone curriculum (Study 1: Barnett et al., 2008; Diamond et al., 2007; Study 2: Solomon et al., 2018); two published findings of *Tools* as a more narrow enhancement focused on pretend play and self-regulation activities (Study 3: Clements et al., 2020; Study 4: Hsueh et al., 2014; Morris et al., 2014), and two sets of findings reported at conferences each concerning implementation of *Tools* as a stand-alone curriculum (Study 5: Hammer et al., 2012; Study 6: Lonigan & Phillips, 2012).

In general, these rigorous evaluations (conducted for the most part either simultaneously or after the completion of the study we describe in the current monograph) found limited evidence that the *Tools* prekindergarten curriculum has significant positive impacts on children's academic, self-regulation, or socio-emotional skills. This conclusion holds whether *Tools* was implemented as a stand-alone full-day curriculum or was the source of play and self-regulation activities which were infused into another curriculum. No reports of preschool randomized control trials have provided evidence of significant benefits on academic outcomes (literacy, mathematics, or vocabulary). However, in one of the early randomized trials of the full *Tools* curriculum (Barnett et al., 2008), teachers in the *Tools* classrooms rated their students as displaying significantly fewer problem behaviors than did teachers in control classrooms. Separately, one of the evaluations of *Tools* as an add-on (Morris et al., 2014) provided evidence that children in *Tools* classroom made greater gains in their knowledge of emotions compared with children in control classrooms. Neither of the two studies that examined the long-term effects of prekindergarten *Tools* into kindergarten or first grade reported differences in outcomes by intervention condition (Clements et al., 2020; Morris et al., 2014).

In addition to examining effects on children's outcomes, prior evaluators of *Tools* observed aspects of classrooms' social contexts and teacher practices, including fidelity of implementation (Barnett et al., 2008; Clements et al., 2020; Solomon et al., 2018) and the quality of the learning environment (Barnett et al., 2008; Clements et al., 2020; Diamond et al., 2019; Hammer et al., 2012; Morris et al., 2014). These observational data were used primarily to provide descriptive summaries (rather than quantitative comparisons) of ways that *Tools* was implemented and ways that *Tools* classrooms were similar to, or distinct from, control classrooms.

In summary, other evaluations of *Tools* have left uncertainty about the implementation and impact of the curriculum. The project we describe in this monograph is a fully integrated and comprehensive large-scale, field-based, randomized control trial of the *Tools of the Mind* prekindergarten curriculum. We provide data on how well the full curriculum was implemented, and address the impact of the *Tools* prekindergarten curriculum on learning and development in children, and on general practices and interactions in classrooms.

Current Study and Overview of the Monograph

We designed the work described in this monograph as a way to advance understanding of the impact of an early childhood curriculum on prekindergarten children's academic, executive function, self-regulation, and social outcomes. Our intent was to contribute to an understanding of why past research about the effectiveness of *Tools* has yielded inconsistent findings. In addition, we use the study of *Tools* as a case study to help identify challenges faced when trying to understand why curriculum studies, in general, often show limited immediate and long-term effects on children's outcomes.

In Chapter II, we describe the methodology of the study's randomized control trial. We also present the results of the primary statistical tests of curriculum effects (i.e., *Tools* compared with control classrooms) on gains in children's academic, executive function, self-regulation, and social skills from prekindergarten to the end of first grade.

In Chapter III, we explain the instrument we designed to document the fidelity of implementation of the *Tools* curriculum. Focusing on teachers who were randomly assigned to implement *Tools*, we investigate the degree to which the curriculum was delivered as intended and examined the associations between fidelity of implementation and children's academic, executive function, and social outcomes in prekindergarten.

In Chapter IV, we first describe how we collaborated with the developers of *Tools* to (a) identify and (b) test hypotheses about expected curriculum effects on classroom processes. Included are hypotheses about the amount of non-instructional behaviors, teacher-led and child-directed activities, teacher and child talk, social learning interactions, as well as about the quality of classroom emotional climate, teacher instruction, and children's level of involvement. As a secondary focus, we investigate whether the identified classroom processes are associated with gains in children's academic, executive function, self-regulation, and social skills combining treatment and control classrooms.

In Chapter V, we situate the findings from this randomized control trial of the *Tools* curriculum within the larger context of early childhood education policies and practices. We offer a series of "lessons learned" to help guide future research and policy.

II. Evaluating the Impact of a Prekindergarten Curriculum on Child Outcomes

In the current chapter, we focus on the methodology and results of a randomized control trial evaluation of the *Tools of the Mind* (Bodrova & Leong, 2007) prekindergarten curriculum. The study took place in five school districts in two states. The aim of the *Tools* curriculum is to enhance children's self-regulation and executive function skills within an instructional context that promotes basic academic and social skills and prepares children for kindergarten and beyond. The developers of the curriculum suggest that the mental tools children learn from the curriculum will equip them to learn more effectively in subsequent grades (Bodrova & Leong, 2017). To investigate the effectiveness of *Tools* in achieving these aims, we asked whether:

1. children in *Tools* classrooms made greater gains in academics (vocabulary, literacy, scientific knowledge, and mathematics) during the prekindergarten year than children in business-as-usual control classrooms (hereafter referred to as *control* classrooms).
2. children in *Tools* classrooms made greater gains in executive function, self-regulation, and social-emotional skills during the prekindergarten year than children in control classrooms.
3. curriculum effects were sustained to the end of kindergarten and first grade.
4. subgroups of children (as defined by gender and age) responded differently to the *Tools* curriculum.

Methods

Experimental Design

Recruitment for the study occurred in two Southern states in the United States. With the assistance of one of the *Tools* developers, researchers solicited school districts through one-on-one contacts and meetings with district personnel. Districts selected for recruitment had an eligible public prekindergarten program and a willingness to participate. Four school districts in one state and one larger school district in the second state participated in the study. Funding for the prekindergarten programs in the participating schools came from state and/or Title I. All families had to meet the income guidelines for free or reduced-price lunch in order to enroll their children. Four of the school districts were located in suburban and rural areas surrounding a large city, and one district was urban.

We employed a randomized block design to test the effectiveness of the *Tools* curriculum compared with the practices and curricula occurring in classrooms in the participating school systems. Schools were the unit of randomization because it was advantageous for conducting *Tools* professional development if all the prekindergarten teachers within a school were trained together and encouraged to support each other during implementation. This scheme was also intended to minimize interactions between experimental and comparison teachers that might have compromised the evaluation. To facilitate random assignment, we grouped schools into blocks and within each block assigned half the schools to implement *Tools* and half to the comparison control condition (with slight variations due to the uneven number of schools and classrooms in some districts). Each of the four smaller districts were stand-alone blocks and we divided the 22 schools in the large, urban district into five blocks based on the number of classrooms in each school. Although there is variability among districts (e.g., suburban or rural), randomization occurred within school districts; thus, across district variations would not impact the randomization process. Moreover, differences between districts based on children's demographics and academic, executive function, self-regulation, and social-emotional skills, were controlled for in the analytic models with the inclusion of a random effect for randomization block.

This research was conducted over 2 years, with randomization occurring in the summer prior to the first year. We randomly assigned schools to intervention and comparison conditions in the summer of 2009. The curriculum evaluation occurred in the 2010–2011 school year. The Vanderbilt University Institutional Review Board approved all procedures used in this research study. We obtained informed consent from all participating teachers who provided information about the children's classroom behaviors in a series of surveys. Parental consent was obtained for all participating children, and children assented at each assessment. With the cooperation of the school districts, we followed and re-assessed consented children at the end of kindergarten and again at the end of first grade.

Tools of the Mind *Professional Development*

As dictated by the curriculum developers, training of teachers and practicing with the curriculum occurred during the 2009–2010 school year. Training and in-classroom coaching continued the second year of full implementation and data collection. A sub-contract to the curriculum developers supported training and coaching for *Tools*. Because the randomization of curriculum condition occurred at the school level, *Tools* training took place with all the teachers in assigned schools.

Certified *Tools* trainers conducted all workshops. The central office for *Tools of the Mind* in Denver, Colorado, has a cadre of experienced *Tools* trainers that the office assigns to various districts implementing the curriculum. This project benefited from the assistance of several of these certified trainers, one of whom worked extensively with the coaches during the implementation year. The research project reimbursed the districts for the cost of hiring substitute teachers to allow teachers to attend the workshops. During the practice year, teachers participated in four workshops spread across the school year. The first 2-day workshop occurred before the start of the school year, whereas, the other three were 1 full day each and spread across the school year. In the full implementation year, teachers attended an additional three 1-day workshops spread over the course of the school year. Following each workshop, teachers completed surveys to identify their needs. These surveys were shared with *Tools* trainers and coaches.

In-classroom and remote coaching (emails and phone calls) was delivered by five *Tools*-trained coaches supplemented the workshops. Three school districts had their own coach; the remaining two districts shared a coach. Districts hired coaches, and project funds supported them. Coaches participated in all the workshops and had separate online consultations with *Tools* developers and trainers (monthly consultations with one of the trainers) and access to a specific coaching manual and training materials. Accompanying each workshop, coaches had dedicated time with the *Tools* trainers for technical assistance, including trainers attending coaching meetings with their teachers. In addition, following each workshop coaches completed a survey to identify their needs.

The intervals between coaching visits ranged from 2 to 4 weeks. The average amount of individual coaching received by each *Tools* teacher (predominantly occurring in classrooms) was 39.22 hr ($SD = 9.65$) in the practice year and 43.95 hr ($SD = 8.19$) in the full implementation year. Teachers also had online access to videos of all *Tools* activities.

During the implementation year, teachers evaluated the quality of the training they received at the onset and end of the year on a 1 (*not helpful*) to 5 (*very helpful*) scale. On average teachers had a positive view of the training at the onset of the year ($M = 3.81$, $SD = 0.82$), with 22 of the 32 teachers rating the training as a 4 or better. There were, however, 10 teachers who rated the training as either a 2 or 3. At the end of the year, training ratings were higher ($M = 4.28$, $SD = 0.85$) with 26 teachers rating the training as a 4 or better and 6 teachers rating the training with a score of 2 or 3.

Participants

Of the 60 teachers participating in the study, we randomly assigned 32 to the *Tools* condition and 28 to the control condition. The control classrooms used a variety of curricula, with the modal one being *Creative Curriculum* (Dodge et al., 2002). With only one exception, all teachers were female. Teachers averaged 12 years of teaching experience, including an average of m years in prekindergarten or preschool classrooms. All teachers had at least a bachelor's degree and were licensed; over half had completed coursework toward or obtained a master's degree. In addition, each classroom had at least one assistant. Teachers' salaries were commensurate with those in the K-12 systems.

In the 60 classrooms, 877 children (498 Tools; 379 control) were age-eligible for prekindergarten and consented to participate in the study in the fall of 2010. The consent rate in *Tools* classrooms was 88% and the consent rate in control classrooms was 76%. Unfortunately, as information about non-consented children was not available, we cannot know if non-consented children varied in important ways from consented children, including if potential selection bias varied across conditions. Consented children in *Tools* and control classrooms were similar on key demographics and pretest assessments (see Randomization Check in the Results section of the current chapter).

Table 2 presents the demographics for the 877 consented children. Overall, the sample of children was diverse in terms of ethnicity and language background. The school districts described 30% of the children as coming from homes where English was not the primary language (hereafter referred to as *home language*). All study classrooms were in elementary schools and the children in the prekindergarten classrooms generally attended the same elementary school for kindergarten and first grade.

TABLE 2

Child Demographics by Curriculum Condition

	Tools of the Mind		Control	
Variable	*n*	%	*n*	%
Male	261	53	218	58
White	192	39	157	41
Black	145	29	86	23
Hispanic	118	24	95	25
Asian	32	6	21	6
Multi-racial	4	1	16	4
Other Minority	7	1	4	1
Home Language[a]	140	28	117	31
Individualized Education Plan[b]	68	14	58	15
Free and Reduced-Price Lunch[c]	329	86	293	88
	M	*SD*	*M*	*SD*
Age (months) at pretest[d]	54.18	3.55	54.54	3.74
Age (months) at posttest[e]	61.52	3.50	62.00	3.70

Note. Tools n = 498 children nested within 32 classrooms. Control condition *n* = 379 children nested within 28 classrooms.

SD = standard deviation.

[a] Home Language coded as English or not English.

[b] Individual Education Plans were for additional supports for learning difficulties. Information about focus of plan was not provided by study districts.

[c] Missing for 116 *Tools* children and 46 control children; percentages reflect percent of non-missing cases.

[d] *n* = 494 *Tools* condition and 372 control condition.

[e] *n* = 467 *Tools* condition and 349 control condition.

Attrition during the study was minimal. All *Tools* teachers in the evaluation year of the project also participated in the practice year (i.e., there were no *Tools* teachers that were new to the project at the start of the evaluation year). In addition, no *Tools* or control teachers left during the evaluation year. The attrition of children over the course of the study was low and similar across *Tools* and control classrooms. Of the consented children, 866 had pretest scores on one or more direct assessments of achievement or executive function. Pretest teachers rated classroom behavior for 862 children. The consented children who did not receive either a pretest or a teacher report in the fall of 2010 had either withdrawn from the school prior to the assessment period or refused to complete one or more of the assessments. In the spring of 2011, 816 children had at least one direct assessment of achievement or executive function and teachers rated 821 children.

In the spring of 2012 (when most children were completing kindergarten), 810 children completed follow-up assessments and teachers rated classroom behaviors of 811 children. In the spring of 2013 (at the end of most children's first-grade year), 778 children completed assessments and teachers rated 779 children. There were no statistically significant differences in attrition by condition. Assessed children at the end of prekindergarten and kindergarten did not differ significantly on any baseline variable from children who were not assessed. At the end of first grade, assessed children had significantly higher baseline scores on one achievement measure (Spelling) and significantly lower baseline scores on another achievement measure (Applied Problems) than non-assessed children.

Measures

To assess the effects of the curriculum, we used a battery of standardized child achievement measures, direct assessments of executive function and self-regulation skills, and teacher ratings of classroom learning behaviors (self-regulation) and social skills. During the year prior to full implementation and evaluation, researchers met with the *Tools* developers and chose assessment measures to reflect aspects of development the developers felt would most likely be affected by the *Tools* experience. Because of the curriculum's strong focus on self-regulation and executive function skills and the lack of a single measure validated to capture those skills, we used a battery of executive function tasks and teacher ratings of classroom behaviors. The developers were interested in standardized measures of literacy, mathematics, and language; in addition, they requested a measure of children's academic (scientific) knowledge.

All assessments were administered by trained staff in English, the language used for all classroom instruction. Assessments were given in a quiet area away from the classroom and were divided into two sessions of approximately 25-min, each on separate days. Testing took place at four times: at the beginning and end of prekindergarten, at the end of kindergarten, and again at the end of first grade. Within each session, measures were in a fixed order for all children, with all executive-function and self-regulation assessments coming before all academic assessments. Specifically, one session included Peg Tapping, HTKS, Copy Design, and Woodcock–Johnson III subscales of Oral Comprehension, Applied Problems, Quantitative Concepts, and Picture Vocabulary. The other session included DCCS, Corsi Blocks, and Woodcock–Johnson III subscales of Letter-Word Identification, Academic Knowledge, and Spelling. Session order depended on availability of assessor, children, and classrooms; preliminary analyses showed no performance differences in relation to session order.

Academic Achievement

Woodcock–Johnson III tests of achievement. Seven subscales came from the *Woodcock–Johnson III Tests of Achievement* (Woodcock et al., 2001). Letter-Word Identification is an assessment of basic emergent reading skills and required children to identify and pronounce letters and words by sight. Spelling measured children's prewriting skills, such as drawing lines and tracing, writing letters, and spelling orally presented words. Oral Comprehension is an assessment of listening comprehension and measured children's ability to understand a short passage read aloud by the examiner by providing a missing word based on the syntactic and semantic cues provided in the sentence. Picture Vocabulary measured expressive vocabulary as children were asked to say aloud the name corresponding to a picture. Academic Knowledge has three sections and measured children's factual knowledge of science, social studies, and humanities. Two subscales assessed math skills, Applied Problems and Quantitative Concepts. Applied Problems measured children's ability to solve small numerical and spatial problems presented verbally with accompanying pictures of objects. Quantitative Concepts measured children's understanding of number identification, sequencing, shapes, and symbols and in a separate section to manipulate the number line.

Internal consistency of all Woodcock–Johnson III subscales was established through split-half reliability of the assessment's dichotomously coded variables (correct or incorrect). Reliability for all subscales is high with a minimum 84% agreement between the two split-halves of the test (Woodcock et al., 2001). Additionally, test–retest reliability after a 1-day delay is high for all subscales with at minimum a correlation of .83 between the two times.

Executive Function and Self-Regulation

Because a primary focus of *Tools* is to support the development of executive function skills and self-regulation, we included a battery of assessments. Several of these are ones used in prior studies of the effects of add-on self-regulation curricula.

Copy Design. The Copy Design task (Osborn et al., 1984) measured regulation and integration of motor movements (i.e., visual-motor integration); children were asked to copy eight simple geometric shapes that were increasingly complex. Each design had two trials and total scores could range from 0 to 16 with higher scores indicating more accurate copies. All coders of the task established interrater agreement for the eight designs at each time point (Cohen's $\kappa s > .60$). Copy Design test–retest reliability has been previously established at $r = .72$ (Lipsey et al., 2017).

Corsi Blocks. The Corsi Blocks task (Corsi, 1972) measured working memory. In this task, children must recall the order in which an examiner points to a series of blocks on a board in an irregular order. The task assessed both forward (repeat the pattern exactly as the examiner demonstrated) and backward memory span (reverse the pattern given by the examiner). Children had two attempts to complete each pattern within a given trial, and there were two trials for both the forward and backward parts of the task. The final score was the longest pattern or span a child could correctly repeat (possible range = 0–10). Reliability for a verbal variation of the task (i.e., backward digit span) was established at $r = .73$ (Lipsey et al., 2017).

DCCS. The DCCS (Zelazo, 2006) assessed children's attention shifting capabilities. The task required children to sort picture cards by features depicted on the cards, first by color (red vs. blue color), and then according to shape (star vs. truck). If children were able to make the switch between sorting rules, they next sorted a set of cards that had either a black border around the card or no border. If the card had a border, children sorted cards by color; if the card had no border, they sorted by shape. Sort rules were taught orally and through demonstration. Zelazo's recommended four-point scoring was used (0 if children did not pass the color sort; 1 if they passed the color sort but not the shape sort; 2 if they passed the shape sort; 3 if they passed the advanced border sort). Test–retest reliability following a 2- and 3-week delay with prekindergartners was established at $r = .48$ (Lipsey et al., 2017).

HTKS. HTKS (Ponitz et al., 2009) assessed self-regulation, including the ability to respond in a way that was opposite to an examiner's request. HTKS required children to respond to two oral prompts, "touch your head" and "touch your toes," then do the opposite in response to those prompts (i.e., touch their heads when the assessor said "touch your toes"). Six practice trials with feedback were given followed by 10 test trials. For children who responded correctly to five or more of the test trials, two new requests were added with inverse actions required ("touch your shoulders" and "touch your knees."). Four practice trials with feedback were given followed by 10 test trials. Final scores for the task were the sum of children's performance on the six practice items and the 20 testing items, with children receiving 0s for incorrect responses, 2s for correct responses, and 1s for self-corrections (possible range = 0–52). Test–retest reliability was previously established on the task at $r = .80$ (Lipsey et al., 2017) and interrater reliability at $\kappa = .79$ (McClelland et al., 2014). Lower than perfect agreement of interrater reliability for the HTKS primarily reflects disagreements in the coding of self-corrections.

Peg Tapping. The Peg Tapping task (Diamond & Taylor, 1996) measured children's inhibitory control. The task required children to tap a wooden peg once when the examiner tapped twice or tap twice when the examiner tapped once. Children received two practice trials with feedback followed by eight practice trials to successfully respond to the request. If successful, 16 test trials without feedback were given; if unsuccessful, the task was stopped and a score of −1 was assigned. Test trials were scored 0 for incorrect responses and 1 for correct (possible range = −1 to 16). Test–retest reliability for peg tapping was established at $r = .80$ (Lipsey et al., 2017).

Teacher-Rated Classroom Behaviors

Cooper-Farran Behavior Rating Scales (CFBRS). Prekindergarten teachers rated the children's social skills and classroom behavioral competencies in the fall (after 6 weeks of school) and again at the end of the school year. Kindergarten and first-grade teachers rated the same skills in the late spring. To capture self-regulation as evidenced in the classroom, teachers reported on children's behaviors in the classroom using the CFBRS (Cooper & Farran, 1991). The CFBRS contains two subscales with items rated from 1 to 7 using behavioral anchors distinctive to each item. The Work-Related Skills subscale includes 16 items about independent work and compliance with and memory for instructions. The Interpersonal Skills subscale includes empathic and respectful behavior toward teachers and peers. Higher scores on the measure indicate more positive behavior exhibited in the classroom. Estimates of internal consistency among subscale items ranged from $\alpha = .90$ to .95.

Analytic Strategy

As represented in Equation (1), to estimate curricular effects, we used three-level nested regression models, with children at Level 1 (children$_{ijk}$), classrooms at Level 2 (classroom$_{jk}$), and randomization blocks at Level 3 (block$_k$) in SPSS Version 22. The dichotomous variable of curriculum condition was entered at the classroom level with *Tools* being the reference group ($\gamma_{010} \times$ condition$_{jk}$). All analyses of achievement outcomes used the Woodcock–Johnson W-scores, which are IRT scaled but not adjusted for age. All other outcomes remained in their raw score form. Each impact model accounted for pretest scores ($\gamma_{100} \times$ pretest$_{ijk}$), age at pretest ($\gamma_{200} \times$ age$_{ijk}$), the interval between assessments ($\gamma_{300} \times$ interval$_{ijk}$), gender ($\gamma_{400} \times$ gender$_{ijk}$), home language ($\gamma_{500} \times \text{lang}_{ijk}$), and IEP status ($\gamma_{600} \times \text{iep}_{ijk}$) at the student level of the model. The pretest, age, and time interval covariates were grand-mean centered. Gender ($0 = \textit{male}$), home language ($0 = \textit{English}$), and IEP status ($0 = \textit{no IEP}$) covariates were dichotomous.

$$\begin{aligned}\text{Posttest}_{ijk} = {} & \gamma_{000} + \gamma_{100} \times \text{pretest}_{jk} + \gamma_{200} \times \text{age}_{ijk} + \gamma_{300} \times \text{interval}_{ijk} \times \gamma_{400} \\ & \times \text{gender}_{ijk} + \gamma_{500} \times \text{lang}_{ijk} + \gamma_{600} \times \text{iep}_{ijk} + \gamma_{010} \times \text{condition}_{jk} \\ & + U_{00k} + U_{0jk} + r_{ijk}. \end{aligned} \tag{1}$$

To test for potential difference in curriculum effects by child characteristics, we estimated the interactions between pretest, age, gender, ELL, and IEP status and experimental condition (Equation 2: $\gamma_{110} \times \text{moderator}_{ijk} \times \text{condition}_{jk}$). We report the results for each outcome variable separately.

$$\begin{aligned}\text{Posttest}_{ijk} = {} & \gamma_{000} + \gamma_{100} \times \text{pretest}_{jk} + \gamma_{200} \times \text{age}_{ijk} + \gamma_{300} \times \text{interval}_{ijk} \\ & + \gamma_{400} \times \text{gender}_{ijk} + \gamma_{500} \times \text{lang}_{ijk} + \gamma_{600} \times \text{iep}_{ijk} \\ & + \gamma_{010} \times \text{condition}_{jk} + \gamma_{110} \times \text{moderator}_{ijk} \\ & \times \text{condition}_{jk} + U_{00k} + U_{0jk} + r_{ijk}. \end{aligned} \tag{2}$$

Cohen's *d* standardized mean difference effect sizes indicated the magnitude of condition difference. The standardized effect size is the mean difference between the two conditions relative to the variability observed (i.e., the difference between the *Tools* and control classroom's covariate-adjusted means divided by the pooled standard deviation of *Tools* and control classrooms) and is interpreted as the mean difference between conditions in proportion to the standard deviation.

Results

Randomization Check

Tables 3 and 4 present the descriptive data on the cases available for the achievement outcomes and non-academic outcomes (executive function skills and teacher reports of classroom self-regulation and social behaviors) at each time point. For the Woodcock–Johnson (WJ) subtests, we present the data as standardized scores, with means of 100 and standard deviations of 15 (from the WJ-III manual). These scores are age adjusted. All analyses, as noted earlier, used W-scores, with age included as a covariate.

Prior to investigating treatment effects, we conducted analyses on baseline variables as a randomization check. Tools and control classrooms were not statistically significantly or even marginally different ($p < .10$) in terms

TABLE 3

Standard Score Means (*M*), Standard Deviations (*SD*), and Sample Sizes (*N*) for Academic Outcomes by Curriculum Condition and Test Time

	Fall Prekindergarten		Spring Prekindergarten		Spring Kindergarten		Spring First Grade	
Variable	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)
Tools of the Mind condition								
Letter-Word	492	91.57 (12.91)	465	100.03 (10.87)	459	107.61 (11.27)	443	108.13 (11.32)
Spelling	492	80.21 (12.23)	465	88.61 (13.92)	459	99.44 (13.90)	443	99.00 (15.05)
Academic Knowledge	492	87.45 (18.03)	465	92.69 (14.57)	459	94.58 (12.45)	443	94.53 (11.48)
Oral Comprehension	492	90.03 (13.25)	465	93.91 (13.87)	459	97.07 (12.93)	443	98.62 (11.73)
Picture Vocabulary	492	91.90 (20.71)	465	95.58 (13.79)	459	94.88 (11.47)	443	95.67 (10.48)
Applied Problems	492	92.84 (15.53)	465	98.63 (12.11)	459	100.17 (12.71)	443	99.58 (13.13)
Quantitative Concepts	492	85.74 (11.87)	465	92.28 (13.16)	459	96.69 (11.34)	443	94.89 (11.45)
Control condition								
Letter-Word	369	89.97 (13.23)	348	100.33 (11.81)	351	108.22 (12.15)	335	108.06 (12.80)
Spelling	369	78.03 (12.59)	348	86.59 (15.18)	351	100.34 (14.65)	335	100.18 (15.95)
Academic Knowledge	369	86.04 (18.25)	348	92.54 (14.99)	351	93.44 (13.52)	335	94.17 (12.77)
Oral Comprehension	369	89.06 (12.99)	348	93.92 (15.05)	351	96.37 (14.35)	335	97.73 (12.51)
Picture Vocabulary	369	91.53 (20.18)	348	95.89 (13.91)	351	94.60 (10.98)	335	94.94 (10.91)
Applied Problems	369	91.78 (14.74)	348	97.77 (12.89)	351	100.64 (12.52)	335	98.74 (12.87)
Quantitative Concepts	369	83.91 (11.99)	348	91.69 (13.20)	351	97.77 (12.20)	335	93.92 (12.39)

Note. Descriptive statistics for *Tools* condition provided in top panel and control condition in bottom panel. Woodcock–Johnson W-scores were used at the analytic variable for all models conducted in the study, standard scores (average standard score is 100 with a *SD* of 15) are reported to assist in the interpretation and generalizability of study findings.

TABLE 4

Means (*M*), Standard Deviations (*SD*), and Sample Sizes (*N*) for Non-Academic Outcomes by Curriculum Condition and Test Time

	Fall Prekindergarten		Spring Prekindergarten		Spring Kindergarten		Spring First Grade	
Variable	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)
Tools of the Mind condition								
Copy Design	492	1.09 (1.59)	465	5.22 (2.81)	459	7.81 (2.93)	443	9.16 (3.06)
Corsi Forward Span	492	2.53 (1.25)	465	3.06 (1.16)	459	3.95 (1.11)	443	4.58 (1.14)
Corsi Backward Span	492	1.15 (1.15)	465	1.57 (1.32)	459	2.66 (1.38)	443	3.69 (1.27)
DCCS	492	1.32 (0.58)	465	1.66 (0.59)	459	1.95 (0.62)	443	2.52 (0.91)
HTKS	492	10.49 (13.61)	464	22.41 (17.23)	459	36.37 (13.65)	443	43.68 (9.69)
Peg Tapping	493	4.38 (5.80)	465	9.43 (5.62)	459	13.32 (3.99)	443	14.66 (2.67)
Interpersonal Skills	492	5.18 (1.07)	472	5.45 (1.05)	459	5.59 (1.02)	442	5.61 (1.06)
Work-Related Skills	492	4.47 (1.17)	472	4.98 (1.22)	459	4.95 (1.15)	442	4.83 (1.25)
Control condition								
Copy Design	369	1.01 (1.52)	348	4.79 (2.82)	351	7.70 (2.91)	335	9.53 (2.87)
Corsi Forward Span	370	2.51 (1.26)	348	3.12 (1.12)	351	3.93 (1.10)	335	4.60 (1.08)
Corsi Backward Span	369	1.17 (1.13)	348	1.60 (1.37)	351	2.83 (1.33)	335	3.78 (1.32)
DCCS	371	1.29 (0.59)	348	1.64 (0.57)	351	1.97 (0.56)	335	2.55 (0.92)
HTKS	369	9.62 (12.19)	348	21.19 (17.14)	351	36.17 (14.27)	335	44.70 (8.08)
Peg Tapping	369	4.32 (5.77)	348	9.17 (5.96)	351	13.12 (4.16)	335	14.98 (1.99)
Interpersonal Skills	370	5.36 (1.07)	349	5.47 (1.07)	352	5.64 (1.05)	337	5.68 (1.05)
Work-Related Skills	370	4.67 (1.13)	349	5.02 (1.14)	352	4.97 (1.19)	337	4.97 (1.26)

Note. Descriptive statistics for *Tools* condition provided in the top panel and control condition in the bottom panel. Means (*SD*s) are unadjusted raw scores for each measure. Range of possible scores: Copy Design = 0–16; Corsi Forward and Backward Span = 0–10; DCCS = 0–3; HTKS = 0–52; Peg Tapping = −1 to 16; Interpersonal and Work-Related Skills = 1–7.

DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders.

TABLE 5

Test of Baseline Equivalence

Variable	Cohen's *d Effect Size*	*B* (*SE*)
Dichotomous variables		
Male	−.101	−0.047 (0.034)
White	−.082	−0.040 (0.039)*
Black	.137	0.072 (0.048)*
Hispanic	−.023	−0.006 (0.035)
Home Language[a]	−.066	−0.002 (0.045)
Individualized Education Plan[b]	−.029	−0.010 (0.024)
Free and Reduced-Price Lunch	−.060	−0.016 (0.031)
Continuous variables		
Age at pretest	−.148	−0.355 (0.250)
Letter-Word	.123	1.703 (2.497)
Spelling	.176	1.286 (1.919)
Academic Knowledge	.078	0.498 (2.197)
Oral Comprehension	.074	0.415 (1.327)
Picture Vocabulary	.018	1.229 (2.211)
Applied Problems	.070	1.086 (2.623)
Quantitative Concepts	.154	0.617 (1.067)
Copy Design	.051	0.085 (0.107)
Corsi Forward Span	.016	0.001 (0.099)
Corsi Backward Span	−.018	−0.040 (0.090)
DCCS	.051	0.011 (0.040)
HTKS	.067	0.478 (0.978)
Peg Tapping	.010	0.094 (0.456)
Interpersonal Skills	−.168	−0.185 (0.120)
Work-Related Skills	−.174	−0.208 (0.133)

Note. Tools of the Mind was the reference group in each model as such positive coefficient indicate that *Tools* classrooms had higher percentages of children (for dichotomous demographic variable) or higher baseline scores (for a continuous assessments and age variables) compared with control classrooms.
DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders; *SE* = standard errors.
[a]Home Language coded as English or not English.
[b]Individual Education Plans were for additional supports for learning difficulties. Information about focus of plan was not provided by study districts.
*$p < .05$.

of teachers' level of education, years teaching, or years teaching prekindergarten. As indicated in Table 5, there were no statistically significant differences ($p < .10$) between the children in *Tools* and control classrooms in terms of gender, home language, IEP status, age, and proportion of children on free or reduced-price lunch (descriptives provided in Table 2). There were, however, small but statistically significant differences for ethnicity, with slightly higher proportions of Black children in *Tools* classrooms; whereas, the control condition classrooms had larger proportions of White children.

Finally, randomization checks of all baseline assessments and teacher ratings (see Table 5 for estimates and Tables 3 and 4 for descriptives), indicated no differences ($p < .10$) between children in *Tools* and control classrooms at the onset of the study. Estimates of standardized mean difference effect sizes (Cohen's *d*) are also provided in Table 5. Effect sizes ranged from an absolute value of 0.010 (Peg Tapping) to 0.174 for (Work-Related Skills). No estimates fell outside of the What Works Clearinghouse Standards' requirements to satisfy baseline equivalence (What Works Clearinghouse, 2020). All analyses reported below employ baseline assessment scores, age, gender, home language, and IEP status as covariates. Ethnicity was not included in the final analytic models as a covariate as it was found to be colinear with other model covariates (e.g., pretest assessment performance, IEP status).

Curriculum Effects

Table 6 provides the results of treatment and control comparisons on the academic outcomes and Table 7 for non-academic outcomes. The tables present the regression coefficients and standard errors for the treatment effects, as well as effect sizes. These coefficients show the differences in residualized achievement gains (i.e., posttest scores controlling for pretest performance) at the end of prekindergarten, end of kindergarten, and end of first grade. *Tools* was the reference group for each model and as such, positive coefficients indicate children in *Tools* classrooms made greater residualized gains compared with children in control classrooms. Negative coefficients indicate the differences favored children in control classrooms. Standardized mean difference effect sizes (Cohen's *d*) were computed using the adjusted means reported in Tables 3 and 4 as well as the unadjusted pooled standard deviation of the scores at the respective time point.

Academic Achievement

Across the different achievement outcomes and the three test times, few significant effects for curriculum condition are evident (Table 6). Those that are significant favor the control condition. Recall that children in the control classrooms were slightly older. Their pretest scores on many of the WJ-III subtests, however, were somewhat lower than those of the children in the *Tools* classrooms. Adjusting for age and pretest scores demonstrated that control classroom children *gained* more over the year in several areas.

At the end of the prekindergarten year, children in the control condition gained significantly more in Oral Comprehension than children in *Tools* classrooms. The effect was small with children in control classrooms on average having standard scores 1.45 points larger on Oral Comprehension at the end of the prekindergarten year compared with children in *Tools* classrooms.

At the end of kindergarten, children who had been in control classrooms in prekindergarten gained more on Letter-Word Identification and Quantitative Concepts than children in *Tools* classrooms. Children from control classrooms gained on average 1.99 points on Letter-Word standard scores ($d = -0.16$) and gained 2.46 more points on Quantitative Concepts scores ($d = -0.21$) by the end of kindergarten compared with children from *Tools* classrooms.

TABLE 6

Academic Achievement Curriculum Effects at the End of Prekindergarten, Kindergarten, and First Grade

	End of Prekindergarten			End of Kindergarten			End of First Grade		
Variable	*B*	*SE*	*d*	*B*	*SE*	*d*	*B*	*SE*	*d*
Letter-Word	−2.866	1.832	−.125	−4.399*	1.638	−.172	−3.085	1.858	−.110
Spelling	0.736	2.204	.028	−3.639†	1.825	−.158	−4.097*	1.731	−.171
Academic Knowledge	−0.828	0.961	−.048	0.421	0.850	.030	−0.232	0.883	−.019
Oral Comprehension	−1.596*	0.736	−.096	−0.587	0.833	−.040	0.415	0.775	.032
Picture Vocabulary	−1.079	0.648	−.071	−0.234	0.584	−.020	0.084	0.653	.008
Applied Problems	−0.277	1.107	−.013	−1.499	0.953	−.087	0.248	1.043	.015
Quantitative Concepts	−1.124	1.005	−.075	−2.675*	0.871	−.210	−0.404	0.820	−.031

Note. *Tools of the Mind* was the reference group in each model as such positive coefficient indicate children in *Tools* classrooms made greater residualized gains (posttest scores controlling for pretest) compared with children in control classrooms. Coefficients are unstandardized regression coefficients from multilevel regression models that controlled for pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest. Estimates of effect sizes of curriculum differences provided are Cohen's *d* estimates standardized mean difference.

SE = standard errors.

*$p < .05$.

†$p < .10$.

TABLE 7
NON-ACADEMIC CURRICULUM EFFECTS AT THE END OF PREKINDERGARTEN, KINDERGARTEN, AND FIRST GRADE

	End of Prekindergarten			End of Kindergarten			End of First Grade		
Variable	*B*	*SE*	*d*	*B*	*SE*	*d*	*B*	*SE*	*d*
Copy Design	.433†	0.222	.154	.056	.205	.019	–.399*	.197	–.130
Corsi Forward Span	–.093	0.076	–.082	–.081	.078	–.073	–.068	.076	–.062
Corsi Backward Span	–.079	0.092	–.060	–.219*	.096	–.161	–.083	.094	–.064
DCCS	–.027	0.044	–.046	–.051	.048	–.085	–.051	.075	–.056
HTKS	–.211	1.075	–.012	–.413	.973	–.030	–.887	.693	–.098
Peg Tapping	.097	0.408	.017	.189	.267	.047	–.302†	.175	–.126
Interpersonal Skills	.107	0.087	.101	.038	.096	.037	.020	.100	.019
Work-Related Skills	.085	0.110	.072	.102	.093	.087	–.033	.095	–.026

Note. Tools of the Mind was the reference group in each model as such positive coefficient indicate children in *Tools* classrooms made greater residualized gains (posttest scores controlling for pretest) compared with children in the control classrooms. Coefficients are unstandardized regression coefficients from multilevel regression models that controlled for pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest. Estimates of effect sizes of curriculum differences provided are Cohen's *d* estimates standardized mean difference.

DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders; *SE* = standard errors.

*$p < .05$.

$^{\dagger}p < .10$.

Finally, at the end of first grade, children from the control classrooms gained more across the first-grade year in Spelling than those coming from *Tools* classrooms, with children from control classrooms gaining on average 2.62 standard scores points ($d = -0.17$) by the end of first grade compared with children from *Tools* classrooms. Previously found group differences on other measures were not sustained.

Executive Function and Self-Regulation

Table 7 presents the results for the executive function measures and teacher ratings of classroom behaviors. Again, there are few statistically significant differences in gains across the year, and those that are significant favor children in the control condition. There are no significant effects on gains on any of the measures at the end of the prekindergarten year. At the end of kindergarten, a significant negative effect appears for Corsi Backward Span indicating that control group children made significantly greater gains on this measure over children who had participated in *Tools* classrooms. However, the effect was small and was equivalent to children from control classrooms having backward spans 0.22 items larger than children from *Tools* classrooms ($d = -0.16$). At the end of first grade, control children made significantly larger gains on the Copy Design task over *Tools* children, equivalent to correctly reproducing 0.39 more shapes ($d = -0.13$).

Analysis of teacher ratings of Interpersonal and Work-Related Skills yielded no statistically significant differences between ratings of children in the *Tools* and control conditions at any time point.

Differential Effects of Curriculum

To examine whether there were differential impacts for certain subgroups of children (moderation by pretest, gender, home language, IEP status, and age), we ran the same series of multilevel regression models described above for our main effects analyses but included a condition by subgroup interaction term to identify any differential effects. These analyses produced no consistent findings for any outcome or subgroup at any time point (end of prekindergarten, kindergarten, and first grade).

What is notable about the subgroup analyses is the lack of consistency in findings across the models. To illustrate this, we report the p-values for each of the interaction terms across all outcomes and test times in Table 8. Only 6% of interactions are significant which is approximately the rate at which we would expect a false positive effect (Type 1 error rate at $p < .05$). Examination of the pattern of significant interactions indicates child characteristics did not moderate any specific outcomes No particular child characteristics moderated condition effects. Lastly, although condition effects at the end of prekindergarten (7 of 14 significant interactions) and kindergarten (5 significant interactions) were more likely, within a given time point there was still no clear pattern of differential condition effects.

TABLE 8

Estimates of Statistical Significance for Condition by Subgroup Interactions

	End of Prekindergarten					End of Kindergarten					End of First Grade				
Variable	Pretest	Gender	HL	IEP	Age	Pretest	Gender	HL	IEP	Age	Pretest	Gender	HL	IEP	Age
Letter-Word	.660	.078†	.900	.532	.012*	.692	.484	.043*	.861	.500	.930	.394	.184	.691	.530
Spelling	.274	.308	.353	.357	.243	.811	.911	.211	.347	.312	.852	.192	.832	.603	.508
Academic Knowledge	.176	.897	.236	.831	.858	.891	.565	.815	.693	.622	.198	.989	.335	.168	.445
Oral Comprehension	.183	.949	.872	.493	.156	.003*	.206	.830	.284	.108	.147	.291	.843	.567	.860
Picture Vocabulary	.078†	.183	.280	.378	.428	.850	.540	.230	.320	.165	.254	.039*	.993	.423	.190
Applied Problems	.819	.326	.892	.358	.692	.943	.739	.378	.551	.922	.335	.516	.939	.867	.963
Quantitative Concepts	.616	.202	.845	.630	.245	.793	.539	.742	.907	.911	.564	.329	.474	.973	.441
Forward Span	.733	.906	.012*	.827	.979	.120	.893	.100	.547	.207	.543	.156	.610	.247	.497
Backward Span	.882	.037*	.082†	.655	.402	.070†	.883	.689	.388	.274	.885	.353	.388	.342	.453
DCCS	.668	.041*	.855	.214	.671	.027*	.180	.248	.155	.556	.395	.605	.985	.044*	.635
Copy Design	.072†	.578	.356	.785	.267	.795	.119	.843	.036*	.321	.369	.449	.238	.266	.810
HTKS	.536	.105	.420	.093†	.970	.641	.995	.994	.322	.645	.159	.237	.933	.639	.606
Peg Tapping	.305	.877	.102	.219	.528	.339	.908	.021*	.808	.981	.429	.719	.299	.259	.468
Interpersonal Skills	.003*	.651	.374	.037*	.827	.114	.658	.449	.838	.155	.576	.309	.340	.675	.119
Work-related Skills	.164	.566	.550	.038*	.270	.715	.785	.321	.146	.236	.889	.588	.305	.191	.241

Note. The p values associated with the coefficients for the condition × subgroup interactions from multilevel models. Full details of subgroup analyses are available upon request from the first author.

DCCS = Dimensional Change Card Sort; HL = home language (English or not); HTKS = Head-Toes-Knees-Shoulders; IEP = Individual Education Plan (additional supports for children with learning difficulties).

$^{*}p < .05$.

$^{\dagger}p < .10$.

Full regression models for the subgroup analyses are available upon request from the first author.

Discussion

This study involved a highly scripted and intentional curriculum focused on facilitating gains across important academic and behavioral areas. As was evident in the intense training received by teachers, the implementation of the curriculum followed the steps decided upon by the developers. These are the same types of steps and intensity of professional development Weiland et al. (2018) recommend for successful curriculum adoption. Yet, contrary to expectation, we did not find positive effects for *Tools* on any of the outcomes on which the developers specifically expected to see benefits, and, surprisingly, we found negative effects on several outcomes for students in the *Tools* classrooms through the end of first grade. Although findings that children in *Tools* classrooms did not differ significantly in terms of their academic, executive function, self-regulation, and social skill gains at the end of the prekindergarten year were contrary to expectations, they are consistent with the general trend of findings from the other randomized control trials of the prekindergarten *Tools* curriculum.

With regard to academic skills, other preschool evaluations also found null effects for direct assessments of vocabulary, literacy, and mathematics with small mean difference effect sizes between treatment and control classrooms (Barnett et al., 2008; Clements et al., 2020; Morris et al., 2014). Situated in the context of other randomized control trials of *Tools*, the lack of differences in academic gains in the current study may not have been so unexpected, although our current study was far more intensive in its professional development and coaching and involved a larger number of classrooms and children.

Regarding the curriculum effects on the study's assessments of executive function, self-regulation, and social skills at the conclusion of the implementation year, the current study also did not find the expected positive effects for children in *Tools* classrooms compared with children in control classrooms. There was a trend toward greater growth on visual-motor integration (coordination of visual perceptual abilities and fine motor control), as measured by Copy Design, for children in *Tools* classrooms ($d = 0.15$), although this finding was reversed in first grade. For other direct assessments of executive function, there were no differences between conditions, and effect size differences were small (less than 0.08 *SD*).

Teacher ratings of children's work-related learning behaviors (e.g., self-regulation in the context of the classroom) and social skills showed similar null effects. These impacts at the conclusion of the implementation of the curriculum add to the mixed findings of other experimental evaluations of *Tools*. Our executive function findings are consistent with findings of those other studies of the prekindergarten curriculum that established baseline equivalence and/or included controls for pretest performance (Clements et al., 2020; Morris et al., 2014; Solomon et al., 2018).

A key attribute of the current study was the longitudinal follow-up of the children past the year of implementation into kindergarten and first grade. Only one other randomized control trial has reported delayed effects and it was for the kindergarten version of the *Tools* curriculum (Blair et al., 2018; Blair & Raver, 2014). Following children into subsequent grades is key to a curriculum approach whose theory of change involves providing children with the mental tools to become good independent learners. The current study's evaluation of the prekindergarten version of *Tools* found that by the end of kindergarten, small gains in favor of control children had increased, with control children exhibiting significantly greater gains on two achievement subtests (Letter-Word Identification and Quantitative Concepts) and one executive function assessment, the Corsi Backward Span, a measure of working memory. Significant differences in favor of the control group were also seen at the end of first grade on Spelling and Copy Design. Looking across test times, assignment to a *Tools* classroom tended to negatively impact children's performance on Letter-Word Identification (*d*s > −0.11 at all times) and Spelling (*d*s > −0.16 at end of kindergarten and first grade).

Taken with the previously summarized differences of the curriculum effects at the conclusion of the year of implementation, the developmental appropriateness of the curriculum for prekindergarten children could be of concern. As there are too few studies to make conclusions about whether differences in outcomes for the prekindergarten and kindergarten versions of *Tools* are due to the differences in the curriculum itself or due to characteristics of the evaluations (e.g., populations, methodology, fidelity of implementation), these are not questions that can be addressed here. Nonetheless, our findings raise awareness of the need to consider the appropriateness of the curriculum for 4- and 5-year-old children and to suggest that a closer examination of the differences between the two versions of *Tools* would be warranted.

To help address the question of variability in effects for whom, a series of subgroup analyses were conducted to test differential impacts based on pretest performance, gender, age, home language, and IEP status. The subgroup analyses, unfortunately, did not help us understand the findings. Treatment impacts were not consistently found across individual subgroups on any similar outcome measures, with several showing opposite impacts for the same subgroup on different outcomes. A smaller prior randomized control trial of *Tools* found differential effects suggesting that prekindergarten *Tools* impacts were greater for children with higher levels of hyperactivity/inattention (Solomon et al., 2018). Differential effects must be considered with great care when the main effect between two conditions (e.g., *Tools* and control classrooms) is null. If there is then moderation of treatment effects by a group characteristic (e.g., gender) the finding indicates that not only does one group potentially benefit from the intervention, the other group would have to be detrimentally impacted.

As summarized here, the primary interest for this evaluation and for the majority of other *Tools* evaluations, involved child outcomes. However, it is also important to consider the impacts of a curriculum on teachers' perceptions of their practices and teaching. For example, Diamond et al. (2019) asked teachers at the end of the school year to report their feelings about teaching and found teachers in *Tools* classrooms were significantly more likely than teachers in control classrooms to indicate that they were extremely excited about teaching and to more strongly look forward to the next school year. The current study did not ask teachers about their feelings about teaching and cannot make this contrast between *Tools* and control classrooms.

In this study, *Tools* teachers were asked about their impressions of *Tools* at the end of the full implementation year. In general, *Tools* teachers rated the ease of implementing the curriculum at the midrange ($M = 3.09$, $SD = 1.06$) on a 1 (*very difficult*) to 5 (*very easy*) scale. They tended on average to be somewhat more positive about the effectiveness of Tools ($M = 3.78$. $SD = 1.10$) on a 1 (*least effective*) to 5 (*most effective*) scale at the end of the full implementation year. For each rating, there was, however, quite a range of responses, with some teachers being quite positive and others just the opposite.

As the field moves more strongly in the direction of advocating for the use of a scripted early childhood curriculum, considerations of how to define and evaluate the effectiveness of a curriculum are critical. This is particularly important in cases of null findings on child outcomes. Considering that teachers tend to have little voice in the adoption of a curriculum within their school or district, it is important in evaluations to consider the impact implementation will have on both teachers and children. Teachers' voice is one of the six characteristics Weiland et al. (2018) identified as contributing to a successful curriculum. There is a bit of a conundrum in incorporating teacher voice in a randomized control trial to determine how effective a curriculum is. In these studies, teachers rarely have a voice in determining the curriculum. We discuss this point in much more detail in Chapter V.

Lastly, the findings of our evaluation of *Tools of the Mind* raise more general questions about how curriculum experiences manifest themselves in assessed skills. *Tools* has several very specific activities that one would expect to be related to children's gains. For example, Graphics Practice involves children drawing various shapes under the direction of the teacher, with the goal of developing eye-hand coordination as well as shape knowledge. We found nonsignificant but positive effects on the outcome most closely connected to this skill, Copy Design, at the end of prekindergarten. The positive effects were short-lived, however—group differences were gone by the end of kindergarten and reversed by the end of first grade. At no time did we find effects for practicing shape drawing on the mathematics assessment that included shape knowledge (Quantitative Concepts*)*. Similarly, in Make-Believe Play Planning, which children did daily, children wrote, "I am going to…" and completed the stem with a drawing (or perhaps a word) saying what role they would adopt in later play. One of the words assessed in Spelling is "to." The word is read to children in the context of a sentence and they are to write it. A hand count of the number of correct responses revealed no difference on this item between children in the *Tools* classrooms and children in control classrooms.

In the subsequent chapters, we discuss how and why the curriculum did not have its intended effects. In Chapter III we explore the structure of the curriculum and its expectations of teachers and report the degree to which the curriculum was delivered as intended (i.e., fidelity of implementation of the *Tools* curriculum). In Chapter IV we explore practices and interactions associated with positive child outcomes that the curriculum developers expected to be affected by the implementation of *Tools*, and in Chapter V we suggest some general lessons derived from our work.

III. Fidelity of Curriculum Implementation in Treatment Classrooms

> *Yet whereas conventional wisdom in policy analysis often locates null results in implementation failure, we have no estimates of the extent to which this is true, particularly in recent, rigorous trials of educational interventions* (Hill & Erickson, 2019; p. 590).

Of increasing importance in understanding the effects of an educational intervention is the fidelity with which the intervention is implemented. As Jenkins et al. (2019) noted, teachers who were supposed to be using the same curriculum often had very different classroom practices. In the case of a randomized control trial in which teachers and coaches are trained in the use of a specific curriculum or approach and then monitored for their implementation, we would expect less variation. Nevertheless, teachers are not machines, and there will be some variability no matter how carefully developers and researchers attempt to achieve uniformity. Since the 1990s when the shift toward evidence-based practices in education began (Connolly et al., 2018), much attention has been paid to the fidelity of implementation (Century et al., 2010; Stains & Vickrey, 2017), also known as intervention fidelity (Nelson et al., 2012)

As described in Chapter II, contrary to our expectations, children who had been randomly assigned to *Tools of the Mind* (Bodrova & Leong, 2007) prekindergarten classrooms did not fare better than their peers who were in business-as-usual control classrooms. These results require careful consideration, and thus it is the aim of Chapter III to understand the degree to which *Tools* teachers properly executed specific practices of the *Tools* prekindergarten curriculum and to examine whether the degree of execution was related to children's gains in academic, executive function, and social skills. *Tools* is substantively different from other early childhood curricula. It has many new activities for teachers to learn, and each activity is intended to be implemented according to steps outlined in curriculum manuals. Perhaps teachers varied in the degree to which they carried out these activities and steps, variations that might account for the overall null findings.

Fidelity of Implementation Reported in Prior Evaluations of *Tools*

Previous randomized control trials of *Tools* considered fidelity of implementation primarily by creating short, general rating scales that addressed whether classrooms appeared to be implementing *Tools* principles or some specific aspects of the curriculum (Barnett et al., 2008; Clements et al., 2020; Diamond et al., 2007, 2019; Morris et al., 2014; Solomon et al., 2018). These scales did not, however, provide detailed assessments of whether the entire curriculum was being implemented. The *Tools* developers themselves had not created a fidelity instrument, and thus each research team developed fidelity measures independently. As a consequence, there was great variability in how investigators defined and measured fidelity. No prior study examined the associations between curriculum fidelity and children's outcomes.

More specifically, in the evaluation of the preschool version of *Tools* implemented in New Jersey (see Barnett et al., 2008; Diamond et al., 2007), the fidelity measures examined first, how fully *Tools* classrooms provided the materials identified in the curriculum and second, whether particular *Tools* procedures were followed. With respect to materials, these researchers reported that by the end of the implementation year, all *Tools* classrooms in the study provided materials required for full implementation. With respect to procedures, the fidelity evaluation focused exclusively on whether teachers followed curriculum requirements to limit large-group meetings to no more than 8–10 min, and to direct questions to the group as a whole rather than to individual children (thus promoting group, rather than individual talk). Data showed that, indeed, *Tools* classrooms devoted less time to whole-group meetings and included more group-directed teacher questions than did control classrooms. Other *Tools* evaluations focused on different elements of the curriculum in their assessments of curriculum fidelity. For example, Clements et al. (2020) used the *Mature Play Observation Tool* (MPOT; Germeroth et al., 2019) to measure the extent to which mature make-believe play occurred, that is, play in which children created imaginary situations, took on explicit roles, and used objects symbolically. The eight-item MPOT rating scale captured both children's play actions and teachers' attempts to facilitate children's mature play. MPOT ratings of mature play were higher for classrooms implementing *Tools* make-believe play compared with control classrooms. Diamond et al. (2019) also assessed play, and reported descriptively that control classrooms had play that was unlike *Tools*' make-believe play; play in control classrooms was not scripted. Again, as with other evaluations of *Tools*, these evaluations did not address the degree to which the implementation fidelity of the *Tools* curriculum was related to child outcomes.

Current Study

The reporting of *Tools* fidelity data in previous work has been mainly descriptive and subjective in nature. There has yet to be an examination of whether variation in the fidelity of implementation of the entire curriculum package is related to child outcomes. In order to differentiate within the 32 classrooms enacting *Tools*, this chapter focuses on two fidelity elements—dosage and adherence (Mendive et al., 2016). Other data were collected to measure horizontal fidelity (or process fidelity; Century et al., 2010), that is, aspects of the curriculum that might differentiate classrooms using *Tools* from those using another curriculum. We postpone discussion of those differentiation data until Chapter IV. In the current chapter we:

1. explain the process of developing and quantifying our measure of the fidelity of curriculum implementation;
2. describe the duration of time (i.e., dosage), as well as the number of activities and steps implemented and the number of time-appropriate activities and steps completed (i.e., adherence) in *Tools* classrooms; and
3. examine associations between (a) curriculum dosage and adherence and (b) growth in children's skills across the prekindergarten year in academic, executive function, and social-emotional domains.

Methods

Participants

For the purposes of examining fidelity of implementation of the *Tools of the Mind* curriculum, this chapter focuses exclusively on the 32 classrooms assigned to the *Tools* condition and the 498 children who were enrolled in those classrooms. It is important to note that the fidelity system was also used in control classrooms where we observed almost no *Tools* activities. The one activity occasionally observed in control classrooms was the freeze game (freezing movement when music stopped playing), a common early childhood activity not unique to *Tools*.

Quantifying Fidelity of Implementation

At the time of this study's randomized control trial, *Tools* was being implemented across the United States when an instrument to assess how well teachers were implementing the curriculum did not exist. When this project began, the *Tools* developers had constructed one detailed fidelity instrument which focused solely on play behaviors during center-based activities. The North West Regional Laboratory had done some previous work on a fidelity instrument covering a broader range of activity periods.

The measures that already existed had limitations. They did not give equal coverage to all *Tools* activities, did not include the activities that had recently been added to the curriculum, and they did not clearly distinguish between instructional features that might be distinctive to the *Tools* curriculum and those that were relevant to *Tools* but more generic. In addition, the response format for most of the items was yes or no, which limits sensitivity to classroom variation. We go into some detail about the development of the fidelity measure and its composition to illustrate the complexity of determining fidelity when a curriculum is full day, includes numerous activities, and is highly scripted. The PCER (Preschool Curriculum Evaluation Research Consortium, 2008) discovered and reported how few curricula included a fidelity measure so that it was difficult to determine if teachers were actually enacting any of the curricula tested. We wanted to avoid that problem.

The first year of the project was spent working with the *Tools* curriculum developers and their national training staff to create a comprehensive instrument to measure the fidelity of implementation in the classrooms. Project staff attended all curriculum-training sessions the first year; one staff member, a former teacher, became thoroughly familiar with the curriculum. We strived to develop an instrument that appeared valid to *Tools* developers as well as the curriculum's experienced trainers. *Tools* was a dynamic curriculum; the activities and their implementation were supposed to change during the year. The curriculum provided a timeline to determine when each of the 61 activities should be implemented across the year. Each of the activities involved between 3 and 12 steps to carry it out appropriately; the steps prescribed the way the activity was to be implemented, and those also changed during the year. For example, as children became familiar with an activity, the teacher eliminated easier steps to enact the activity and added other extensions.

During fall of the pre-evaluation year, project staff had several daylong meetings with the developers and trainers. During those meetings, staff also visited classrooms where teachers experienced with the curriculum were currently implementing it. A preliminary draft of the fidelity instrument was designed and shared with one of the *Tools* developers and several trainers in mid-winter in a meeting in Asbury Park, New Jersey. Because of the complexity of the curriculum, from the outset, the fidelity measure was designed for a tablet computer. The developers chose the New Jersey location as a test site because it provided access to a large group of classrooms implementing *Tools*. *Tools* trainers and the developer were given the opportunity to try out the system and to provide important feedback.

Similar to other curriculum developers (Century & Cassata, 2016), *Tools* developers had a difficult time deciding which were the core components of the curriculum and which were of lesser priority. Eventually, all activities, together with the steps that were prescribed to implement them, were included in the fidelity measure. However, the use of make-believe play to build self-regulatory skills in children was a central focus of the curriculum. Thus, the one main aspect the developers identified as being both critical and unique to *Tools* was the presence of a defined make-believe play theme visible in all centers. It was intended to encourage purposeful interactions and high-level dramatic play complete with defined roles and role speech. In our analyses, we examine the implementation of the make-believe play associated activities separately.

In addition, the developers asserted that individualization based on a child's zone of proximal development would be evident through the use and subsequent withdrawal of physical mediators and the presence of differentiated levels of scaffolding over time (changing implementation steps).

The curricula were designed to encourage children to work in pairs and to use private speech to guide their actions. In classrooms implementing *Tools,* children should not be engaged in rote copying and worksheet activities, have assigned seating, or be subject to external behavioral reinforcement contingencies.

We created a detailed plan for the curriculum to allow us to track the expected changes in implementation across the year. Project staff traveled to New Jersey for 3 days to share the fidelity instrument development with *Tools* trainers. On the first day, we introduced the observation instrument to the developer and the staff, following which everyone spent 2 days beta-testing the measure by observing in six classrooms where teachers had been implementing *Tools* for 5–6 years. Suggestions and ideas to further revise the measure were discussed and incorporated after the observations.

Fidelity of Implementation Instruments

Documenting the implementation of a curriculum as complex as *Tools* would not have been possible without the assistance of digital recording. The FileMaker Pro® database software on tablet computers served as the platform for the *Tools of the Mind Fidelity Instrument* (Vorhaus & Meador, 2010). An observer, familiar with the curriculum, noted when a *Tools* activity began and captured the time on the tablet. The observer then clicked to the page describing the activity selected. All steps of the activity were listed along with behaviors the developers said "should not" occur such as worksheets (hereafter termed *should-nots*). The observer selected the steps observed along with any should-nots. Because of the flexibility of the database entry screens on the tablet, observers moved with ease among activities and back to the main classroom observation page. The length of each activity was captured at the same time. The fidelity instrument yielded information about how much of the day *Tools* was implemented, which activities were enacted, and with how many steps and should-nots.

The fidelity instrument required coders to record concrete, observable behaviors reflective of curriculum implementation. Codes were designed to capture both dosage (i.e., how much time of the day was spent delivering the curriculum) and adherence (i.e., whether the structure and sequence of curriculum activities were followed, see Outhwaite et al., 2019). Because codes explicitly tapped the complexity of the curriculum, it was necessary for coders to be highly familiar with the curriculum itself. Thus, all fidelity coders participated in *Tools* curriculum training and received additional and extensive training from research staff prior to each round of observations. Although coders' knowledge of the curriculum therefore made it impossible to blind them to condition, the instrument's focus on explicit behaviors rather than high-inference codes likely avoided (or at least minimized) effects of coder bias.

Tools activities were organized into five time blocks: (1) large group, (2) literacy and story lab, (3) math and science, (4) across-the-day activities, and (5) make-believe play. Large group activities were implemented in a whole-group setting as children engaged in discussions with peers to solve the mystery questions, do the calendar activity and weather graphing, share and tell, and review the class schedule for the day. The literacy and story lab block consisted of activities to support phonemic awareness, vocabulary, letters, and the turn-taking roles of reader and listener, as well as an interactive reading activity. The math and science block contained activities aimed at supporting children's memory, number sense, and spatial awareness as well as an understanding of scientific knowledge and the scientific method. The across-the-day block included activities implemented throughout the day to support children's attention, self-regulation, movement, and community. Last, to encourage purposeful interactions and high-level, dramatic play complete with defined roles and role speech, the make-believe play block included not only children's engagement in make-believe play but play planning beforehand that used a scaffolded writing process.

The *Tools Fidelity Instrument* was embedded in another instrument developed for prior studies. The *Narrative Record* (Farran et al., 2010) was used in the study to capture the overall organization of the classroom environment. The *Narrative Record* provided a continuous record of data about the progression of episodes that classify the classroom in terms of pedagogy and academic content (see Chapter IV for detailed information on the *Narrative Record*). With the *Narrative Record* as the base, observers first decided whether a *Tools* activity was occurring and which type within one of the five time blocks. Any time a *Tools* activity took place, the observer marked what activity was happening and then was given access to a list of steps, mediators, and should-nots for that activity. The observer marked each step completed, marked any mediators that were observed, as well as whether anything occurred that would negate or violate the purpose of the activity. When the activity concluded, the observer returned to the Narrative Record home page.

In the *Narrative Record* system, an episode begins when 75% of children in the classroom participate in a new learning setting or a change in academic content. The episode ended, and a new episode began when there was a shift in the learning setting or academic content for 75% of children in the classroom. Because it is continuously recorded, the *Narrative Record* captured the entire observational period with no breaks in coding and thus was able to capture the amount of time classrooms engaged in *Tools* activities as well as time spent in other non-*Tools* activities.

Data Collection Procedures

Daylong observations took place three times during the implementation year, specifically, in the fall, mid-winter, and spring. We observed all classrooms in both *Tools* and control conditions. Observers of implementation fidelity participated in a weeklong intensive training session before the first observation. Two days of re-training took place before each of the subsequent observations. Reliability estimates were calculated from two observers spending a full day in 23% of the classrooms (41 total reliability observations), with different classrooms being chosen at each time. We calculated estimates for reliabilities for whether an activity occurred, and the use of mediators using Cohen's κ; the averages of those estimates across classrooms and across times were .954 for the occurrence of an activity and .904 for the use of mediators. The least reliable estimates involved the Attention Focusing Activities ($\kappa = .590$); these activities were short and could occur at any point throughout the day. Other activities had a nearly perfect agreement among observers. The *Tools* curriculum was distinctive; once trained, observers had little difficulty seeing the activities and recording information about how they were carried out.

Intraclass correlations (ICC) calculated reliabilities for the continuous variables of fidelity. Based on guidance from McGraw and Wong (1996), we conducted a two-way random effects model (raters randomly assigned to a reliability session as part of our larger population of raters), with single rater/measurement and absolute agreement parameters. ICCs for the number of steps carried out correctly and the number of should-nots observed for an activity was .978 and .930, respectively. Interrater reliability for time spent in *Tools* activities had an ICC = .966.

Analytic Strategy

As represented in Equation (3), to estimate the effects of fidelity of implementation on child outcomes, we employed three-level nested regression models with children at Level 1 (children$_{ijk}$), classrooms at Level 2 (classroom$_{jk}$), and randomization blocks at Level 3 (block$_k$) in SPSS Version 22. We entered the fidelity of implementation analytic variables at Level 2 ($\gamma_{010} \times \text{fidelity}_{jk}$). All analyses of achievement outcomes used the Woodcock–Johnson W-scores, which are IRT scaled but not adjusted for age. All other outcomes remained in their raw score form. Each impact model accounted for pretest scores, age at pretest, the interval between assessments, gender, home language, and IEP status at the student level in the model. The pretest, age, and time interval covariates were grand-mean centered. The status covariates of gender (0 = *male*), home language (0 = *English*), and IEP (0 = *no IEP*) were dichotomous covariates. The results for each outcome variable are reported separately. As analyses are at the classroom level, significance was based on the classroom sample ($N = 32$), which is a relatively small sample, thus associations with p values of .10 or better are discussed.

$$\begin{aligned}\text{Posttest}_{ijk} = {} & \gamma_{000} + \gamma_{100} \times \text{pretest}_{jk} + \gamma_{200} \times \text{age}_{ijk} + \gamma_{300} \times \text{interval}_{ijk} \times \gamma_{400} \\ & \times \text{gender}_{ijk} + \gamma_{500} \times \text{lang}_{ijk} + \gamma_{600} \times \text{iep}_{ijk} + \gamma_{010} \times \text{fidelity}_{jk} \\ & + U_{00k} + U_{0jk} + r_{ijk}. \qquad (3)\end{aligned}$$

To estimate the impact of the two indicators of fidelity, we multiplied the unstandardized coefficient (B) for the indicator from the multilevel regression models by the standard deviation of the indicator and then divided by the standard deviation of the child outcome ($ES = (B_{\text{fideltiy}} \times SD_{\text{fidelity}})/SD_{\text{child\_outcome}}$); see National Institute of Child Health and Human Development Early Child Care Research Network & Duncan, 2003 for an overview of the procedure). The standardized effect size (*ES*) calculated the change in a child's outcome in standard deviation units when fidelity increased by one standard deviation.

Results

Fidelity of Implementation Descriptives

Time in Tools *Instruction (Dosage)*

Presented in Table 9 are the means and standard deviations from the *Narrative Record* portraying how time was spent in the *Tools* classrooms across the three observations. Table 9 includes the *Tools* activities grouped by the major type, the time teachers devoted to non-*Tools* instructional activities, transitions, meals, and nap. Mixed Tools were *Tools* activities that teachers appeared to have adapted or added to in order to make it their own version.

The standard deviations presented in Table 9 indicate that there was variability across classrooms in the allocation of classroom time. For example, at the first observation, although on average classrooms spent 14.01 min in make-believe play centers, the *SD* was 10.81 min. Similarly, for the literacy time block, the mean was 19.03 min with an *SD* of 9.25. The overall time in the day that classrooms spent in *Tools* activities varied from 5% to 54% at Observation 1, from 8% to 41% at Observation 2, and 9% to 42% at Observation 3. Clearly, some teachers chose not to implement much of the curriculum, whereas other teachers were closer to providing a full dose of the curriculum.

The *Tools* curriculum manuals did not clearly describe how to organize the day. For instance, only a small amount of time was indicated for transitions among activities although transitions were clearly necessary given the number of activities contained in the curriculum. Also difficult for teachers was the large number of manuals they received across the year and the fact that information about how time should be spent varied across different manuals. Research staff went through the curriculum manuals from each time point to determine (roughly) how much time teachers should devote to each of the major blocks in the curriculum. Table 9 provides a comparison of this rough estimate of the time prescribed by the curriculum and time actually observed in classrooms. An examination of the values from the manual and actual time spent indicates that classrooms on

TABLE 9

Classroom Schedule in Minutes as Designated by the *Tools* Manuals and as Observed

	Observation 1			Observation 2			Observation 3		
	Manual 1 and 2	Observed Mean	Observed *SD*	Manual 3	Observed Mean	Observed *SD*	Manual 4	Observed Mean	Observed *SD*
Large Group	30	18.22	13.64	30	16.04	6.86	30	21.31	12.32
Literacy	30	19.03	9.25	30	20.77	14.81	30	18.97	11.09
Math/Science	10	0.87	4.58	10	4.47	7.48	10	3.15	7.37
Make-Believe Play Planning	15	8.95	5.69	15	11.09	6.18	15	12.02	5.58
Make-Believe Play Centers	25	14.01	10.81	45	20.54	13.82	45	18.51	11.35
Make-Believe Play Practice	15	3.48	7.43	0	4.29	6.72	0	2.19	3.57
Tools Transitions	10	12.91	5.03	10	10.86	4.81	5	9.95	4.53
Across the Day Activities[a]	–	18.61	14.40	–	17.43	12.90	–	19.67	15.21
Free Choice Centers	60	25.70	22.59	35	29.92	17.81	35	19.01	19.12
Non-*Tools* Instruction	–	32.77	26.38	–	25.98	23.16	–	25.75	17.26
Non-*Tools* Transitions	–	39.48	17.92	–	33.89	11.11	–	36.37	18.60
Meal/Nap/Out	165[b]	165.98	45.87	185[b]	164.73	36.35	190[b]	173.10	40.47

Note. $n = 32$. Values reported represent minutes of a 6-hr (360 min) school day.

SD = standard deviation.

[a] Reflects the time in the school day not allocated to other activities by the *Tools of the Mind* manuals.

[b] Although included for implementation in the *Tools* manuals, across the day activities were to be integrated throughout the day and manuals do not specify durations of time to be spent on these activities.

average spent less time than prescribed implementing *Tools* activities, described in more detail below.

The three observations corresponded to the three training sessions in the implementation year. Across the three observations, the *Tools* manuals indicated approximately half of the 6-hr day should be devoted to *Tools* activities (54.2%, 48.6%, and 47.2% at Observations 1, 2, and, 3, respectively). Classrooms on average were actually observed spending less than a third of their time in *Tools* activities (28.7%, 32.8%, and 29.2% at Observations 1, 2, and 3, respectively). In terms of absolute minutes, in a 6-hr day the difference between expected and observed enactment of *Tools* activities was 91.8 min at Observation 1, 57.0 min at Observation 2, and 64.9 min at Observation 3. The data in Table 9 show that although generally classrooms were engaging in less time on *Tools* activities than prescribed, some classrooms were dedicating time to *Tools* activities consistent with overall time commitment expectations.

Fidelity of Activity Implementation (Adherence)

When we conducted this study, the curriculum contained 61 separate activities, with varying schedules of implementation across the year. Some were twice a week, some every day. Some occurred later in the year, some earlier and then not again. Each activity also had a series of 3–13 steps with various numbers of mediators to be used and should-nots to be avoided. Thus, in addition to the amount of time spent enacting *Tools* activities (dosage), the quality of the implementation could be calculated by examining the number of (1) time-appropriate *Tools* activities (activities indicated as appropriate to implement by the curriculum at the given observation point in the school year), (2) all *Tools* activities (irrespective of whether or not it was indicated as appropriate to implement at the given observation point in the school year), (3) time-appropriate steps, (4) all steps, (5) mediators, and (6) should-nots.

With all these metrics, choosing what scores to summarize to determine variations in implementation fidelity was not simple. We could simply count the number of activities we observed being implemented and the number of steps enacted. We were, however, well aware of the differences among the activities in their difficulty to prepare for and implement. For example, implementing Make Believe Play Centers required extensive preparation. Teachers had to organize their classrooms around a central theme (e.g., restaurant, health clinic, grocery store). This in turn required removing props related to previous themes, incorporating new props that facilitated play around the new theme, and reorganizing centers around new scenarios related to the theme. In contrast, the activity of Weather Graphing required only the development of a weather graph at the onset of the school year; implementation merely involved teachers guiding children to update the graph each day. Project staff highly familiar with the curriculum estimated the difficulty to implement each of the activities. A simple three-level system was devised—easy, medium, and difficult.

Table 10 provides a list of all the *Tools* curriculum activities by observation period. The first column in Table 10 designates whether the activity was deemed easy, medium, or difficult to prepare and implement. Further, information in this table indicates when during the year an activity was to be implemented and its alignment with classroom observations, as well as the expected steps for the activity for a given observation period. Finally, the number of predefined actions that should not occur for a given activity is provided.

To create a fidelity score that might approximate a measure of the quality of implementation, we combined our behavioral data into a weighted fidelity score for each observation. As we have said, the curriculum prescribed a timeline for the implementation of activities during the year as well as changes in the steps of the activity. We also independently determined the complexity involved in implementing an activity. This newly created weighted score adjusted for the difficulty level of each activity and the time-appropriateness of the steps enacted. The curriculum developers themselves did not specify what constituted quality of implementation; this weighted score is our approximation of an index of the quality of implementation. It captured the efforts teachers were expending on the more difficult items and enacting the curriculum steps appropriately. The weighted fidelity score significantly correlated with the number of activities and steps implemented across all three observations ($r = .79$–$.92$), meaning that generally if teachers implemented the curriculum they did so with fidelity. Detailed information on the creation of the weighted fidelity score is provided in Meador et al. (2015).

Table 11 provides descriptive statistics for the various metrics of fidelity of implementation. Across observations, the number of activities was relatively consistent (e.g., time-appropriate activities ranged from 12.47 at Observation 1 to 13.81 at Observation 2). However, variability existed across the 32 *Tools* classrooms with as few as 3 time-appropriate activities observed in one classroom and as many as 20 in another. It is important to note that none of the classrooms in the study implemented all 22 time-appropriate activities prescribed by the *Tools* manuals.

We observed a similar pattern in the use of mediators and developer-identified should-nots, with relatively stable averages across observations (30.75–32.47 for mediators and 3.19–4.78 for should-nots) and once again wide variability among individual classrooms (range = 7–48 mediators used at a given observation and 0–12 for should-nots). With regard to steps, the average number implemented tended to be lower at Observation 1 (e.g., 45.25 time-appropriate steps at Observation 1 and 58.31 steps at Observation 3), but as with the other metrics, variability existed among classrooms with as few as 8 time-appropriate steps observed in one classroom and as many as 86 in another. It should be noted that the total number of steps observed is by definition related to the actual number of activities the teacher carried out.

TABLE 10

TOOLS ACTIVITIES BY IMPLEMENTATION DIFFICULTY, TARGETED TIME OF SCHOOL YEAR, STEP IMPLEMENTATION, AND ACTIVITY SHOULD-NOTS

		Observation 1[a]		Observation 2		Observation 3		
Activities	Level of Difficulty[b]	Enactment Window[c]	Steps[d]	Enactment Window	Steps	Enactment Window	Steps	Should-Nots[e]
Large Group								
Mystery Question	E	1	1–5					6
Mystery Shape	E	2	1–4	2, 3	1–6			6
Mystery Word	E			3	1–3	3, 4	1–7	6
Mystery Numeral	E			3	1–3	3, 4	1, 3–7	6
Mystery Pattern	E					4	1–6	6
Mystery Letter	E					4	1–4	6
Mystery Rhyme	E					4	1–4	6
Timeline Calendar	E	1, 2	1–5	2, 3	1–7	3, 4	1–8	6
Weather Graphing	E	1, 2	1–3	2, 3	1–3	3, 4	1–3	2
Message of the Day	M	1, 2	1–6	2, 3	1–7	3, 4	1–8	8
Message of the Day Write Along	D					4	1–7	8
Share the News	E	1, 2	1–6	2, 3	1–4, 7	3, 4	1–4, 8	3
Share and Tell	E	1, 2	1–5	2, 3	1–5	3, 4	1–5	3
Tally	E					4	1–4	0
Write Along a Familiar Song/Finger Play	D					4	1–5	5
Make a Rhyme	M					4	1–5	2
Take Away Sounds	M					4	1–7	2
Class Schedules	E	1, 2	1–3	2, 3	1–3	3	1–3	0
Literacy								
Graphics Practice	M	1, 2	1–9	2, 3	1–8	3, 4	1–8, 11–13	5
Buddy Reading	M	1, 2	1–6	2, 3	1–9	3, 4	1–5, 7–10	5

(*Continued*)

TABLE 10. (*Continued*)

		Observation 1[a]		Observation 2		Observation 3		
Activities	Level of Difficulty[b]	Enactment Window[c]	Steps[d]	Enactment Window	Steps	Enactment Window	Steps	Should-Nots[e]
Elkonin Boxes 1: Jumping the Sounds	D					4	1–5	4
Elkonin Boxes 2: Token Game	D					4	1–4	4
I have who has Letters	E			3	1–8	3, 4	1–8	4
Story Lab: Active Listening	E	1, 2	1–6	2, 3	1–6	3, 4	1–6	4
Story Lab: Connections	E	1, 2	1–5	2, 3	1–5	3, 4	1–5	3
Story Lab: Vocabulary	D	1, 2	1–6	2, 3	1–6	3, 4	1–6	4
Story Lab: Learning facts	D	2	1–5	2, 3	1–6	3, 4	1–7	1
Story Lab: Visualization	M	2	1–7	2, 3	1–7	3, 4	1–8	2
Story Lab: Grammar	D			3	1–10	3, 4	1–10	3
Story Lab: Extensions	D			3	1–8, 10	3, 4	1–10	4
Story Lab: Predictions and Inferences	D					4	1–6	1
Math/Science								
Remember and Replicate	M	1, 2	1–8	2, 3	2–9	3	2–7, 9, 10	1
Puzzles and Manipulatives	E	1, 2	1–3					1
Math Memory	M	2	1–8	2, 3	1, 3–9	3, 4	1, 3–13	2
Science Eyes	D	2	1–6	2, 3	1, 2, 4–9	3, 4	1, 2, 4, 5, 7–12	5
Numeral Game	M			3	1–5	3, 4	1, 2, 4–8	2

(*Continued*)

TABLE 10. (*Continued*)

		Observation 1[a]		Observation 2		Observation 3		
Activities	Level of Difficulty[b]	Enactment Window[c]	Steps[d]	Enactment Window	Steps	Enactment Window	Steps	Should-Nots[e]
Venger Drawing	D			3	1–5	3, 4	1–6	0
Attribute Game	M			3	1–4	3, 4	1–6	0
Numberline Hopscotch	M			3	1–4	**3**	**1–6**	2
I have who has Colors	E			3	1–8	3	1–8	3
I have who has Numbers	E			3	1–8	3, 4	1–8	3
I have who has Shapes	E			3	1–8	3, 4	1–8	3
Making Collections	D	2	1–4, 6–12	2, 3	1–3, 5–12	3, 4	1–3, 5–12	0
Patterns with Manipulatives	M					4	1–5	0
Make Believe Play								
Make Believe Play Planning	D	1, 2	1–8, 10	2, 3	1–10	3, 4	1–11	7
Make Believe Play Practice	D	1, 2	1–4	2, 3	1–4	3, 4	1–8	2
Make Believe Play	D	1, 2	1–5	2, 3	1–7	3, 4	1–11	2
Make Believe Play Clean-up	E	1, 2	1–3	2, 3	1–3	3, 4	1–3	3
Tools Transitions								
Pretend Transitions	E	1, 2	1–3	2, 3	1–3	3, 4	1–3	3
Across the Day Activities								
Attention Focusing Activities	E	1, 2	1–5	2, 3	1–5	3, 4	1–6	2
Freeze Game	E	1, 2	1–4	2, 3	1–5	3	1–5	4
Partner Freeze	E					4	1–7	4
Two Step Freeze	M					4	1–4	4

(*Continued*)

TABLE 10. (*Continued*)

		Observation 1[a]		Observation 2		Observation 3		
Activities	Level of Difficulty[b]	Enactment Window[c]	Steps[d]	Enactment Window	Steps	Enactment Window	Steps	Should-Nots[e]
Freeze on Number	M			3	1–4	3, 4	1–5	4
Pattern Movement Game	M	2	1–7	2, 3	1–7	3	1–9	3
Complete and Continue	M			3	1–7	3, 4	1–7	3
Number Follow the Leader	M			3	1–4	3, 4	1–5	2
Community Building Activities	E	1, 2	1–3	2, 3	1–3			0
I have who has Name Game	E	1, 2	1–6	2, 3	1–6	3, 4	1–6	1
Mousetrap	E					4	1–5	2
What are you doing Mr. Wolf?	E					4	1–5	2

[a] Three observations were conducted over the course of the school year, noted enactment window and time-appropriate steps indicate if a given activity was appropriate to implement during a given observation. See Chapter II for more information on the observation schedule.

[b] Level of difficulty designates whether the activity was deemed easy (E), medium (M), or difficult (D) to prepare and implement.

[c] Based on *Tools* manuals, activities were designated as appropriate to implement between August and September (1), October and December (2), January to February (3), and March to April (4).

[d] In addition to activities having windows for appropriate implementation, appropriate steps to be followed varied over the course of the school year. Activities varied in the number of expected steps (range = 3–13 steps).

[e] Should-nots were predetermined actions that should not occur during the implementation of an activity (range = 0–6).

TABLE 11
Fidelity of Implementation by Observation

		Observation 1			Observation 2			Observation 3		
	n	Range	Mean	*SD*	Range	Mean	*SD*	Range	Mean	*SD*
Time-Appropriate Activities[a]	32	4–19	12.47	3.28	3–20	13.81	3.58	5–19	13.75	3.29
All Activities[b]	32	5–22	13.97	3.57	4–22	14.91	3.77	6–20	14.84	3.34
Time-Appropriate Steps[a]	32	12–70	45.25	14.40	8–84	55.75	16.55	15–86	58.31	16.35
All Steps[b]	32	16–78	53.66	16.18	11–95	61.81	18.27	15–91	62.00	17.13
Mediators[b]	32	12–46	30.75	7.96	12–48	32.31	7.84	7–44	32.47	8.21
Should-Nots[b]	32	0–8	4.78	2.57	0–9	3.19	3.06	0–12	4.44	2.91
Weighted Fidelity[a]	32	28–259	154.05	55.14	23–290	180.36	58.42	48–293	164.42	53.42

SD = standard deviation.
[a]Values reported only include activities and steps that were indicated as appropriate to implement by the curriculum at the given observation point in the school year.
[b]Values reported include all activities, steps mediators, and should-nots observed irrespective of whether the feature was indicated as appropriate to implement by the curriculum at the given observation point in the school year.

The highest weighted fidelity score averages were at Observation 2 (180.36), followed by Observation 3 (164.42), and Observation 1 (154.05) with scores among classrooms varying from 28 to 293 across the three observations. Based on the curriculum manuals, if fully implemented, weighted fidelity scores should range from 380 to 460 at Observation 1, 370 to 530 at Observation 2, and 350 to 570 at Observation 3. Thus, on average, classrooms in the study implemented about half of what the developers expected for full implementation. Moreover, although some classrooms had scores approaching 300, none were within the range identified in the manuals as full implementation. In part, these scores reflect the way we calculated the weights. *Tools* teachers could be doing many activities (yielding high activity and step scores) but enacting only the simplest and least demanding ones. To have a higher weighted fidelity score, teachers had to implement more complex activities with fidelity.

Teachers were moderately consistent in the degree to which they implemented across observations with the number of activities implemented at one observation significantly related to the number of activities implemented in the others ($r = .52–.69$). Thus, teachers who implemented fewer activities at one observation tended to implement fewer activities at all the observations and vice versa. Similar associations were observed for steps ($r = .55–.64$), mediators ($r = .56–.68$), and the weighted fidelity score ($r = .43–.68$), but not for the should-nots ($r = .21–.29$).

Associations With Child Outcomes

The variability in the degree to which teachers implemented the *Tools* curriculum could have led to overall null results on effects. To examine if variability in fidelity of implementation was related to child outcomes, we examined the associations between two fidelity variables—the amount of time spent in *Tools* instruction (dose) and the weighted fidelity scores (quality or adherence)—and children's skills at the end of prekindergarten (see Chapter II for a description of outcome measures). Because the mean weighted fidelity score based on all three observations was 502.1, we rescaled the variable by dividing by 100 so that parameter estimates would not be too small when taken out to three decimal points. To interpret the weighted fidelity score in terms of the raw metric, multiple the parameter estimate provided by 100.

Results from the analyses indicated that neither the amount of time spent in *Tools* instruction nor the quality or adherence of implementation of the *Tools* activities was statistically related to children's academic, executive function, or social-emotional outcomes at $p < .05$ (Table 12). There were a few marginal ($p < .10$) associations. For the academic measures, a positive association was found between the amount of time in *Tools* and gains in Applied Problems (*ES* = .072) and a negative association between

TABLE 12

Fidelity of Implementation Effects at the End of Prekindergarten

		Amount of *Tools* Instruction[a]			Weighed Fidelity Score[b]		
Variable	*n*	*B*	*SE*	*Effect Size*	*B*	*SE*	*Effect Size*
Letter-Word	465	10.761	19.726	.037	.281	1.045	.018
Spelling	465	7.451	21.785	.023	.185	1.102	.010
Academic Knowledge	465	−14.507	8.569	−.067	−.577	0.443	−.048
Oral Comprehension	465	0.098	7.529	.001	−.674	0.362	−.059[†]
Picture Vocabulary	465	−1.921	5.704	−.010	−.037	0.290	−.003
Applied Problems	465	19.453	11.078	.072[†]	.331	0.575	.022
Quantitative Concepts	465	4.681	9.765	.024	.109	0.498	.010
Copy Design	465	4.498	2.244	.125[†]	.090	0.118	.046
Corsi Forward Span	465	0.169	0.743	.011	.010	0.038	.012
Corsi Backward Span	465	0.258	0.856	.015	.036	0.043	.039
DCCS	465	0.106	0.412	.014	.023	0.020	.055
HTKS	464	−5.125	11.218	−.023	−.215	0.575	−.018
Peg Tapping	465	2.669	3.473	.037	−.156	0.181	−.040
Interpersonal Skills	472	0.243	0.935	.018	.059	0.049	.080
Work-Related Skills	472	−0.048	1.184	−.003	.019	0.064	.022

Note. Coefficients are unstandardized regression coefficients and standard errors from multilevel regression models that account for nesting of teachers in schools and random assignment blocks. Covariates included in the models were pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest. Standardized effect sizes calculated indicated the change in a child's outcome in standard deviation units when quality increased by one standard deviation.

SE = standard errors.

[a] Proportion of school day in *Tools of the Mind* Instruction.

[b] Weighted fidelity score rescaled to dividing original estimate by 100.

[†] $p < .10$.

the weighted fidelity score and Oral Comprehension. There was also a positive association between the amount of time in *Tools* and children's prekindergarten gains on Copy Design, a measure of visual-motor integration (*ES* = .125). Thus, even though there was variability in implementation, it was inconsistently related to the prekindergarten gains in the outcomes examined. Basically, despite variability across teachers, higher levels of fidelity of implementation did not relate to the gains made by the children.

Make-Believe Play and Literacy Sub-Analyses

According to the developers, make-believe play and literacy activities were the two key aspects of the *Tools* curriculum. A central focus of the curriculum was the use of make-believe play to build self-regulatory skills in children. The presence of a defined make-believe play theme that cut across all centers and encouraged purposeful interactions through defined roles and role speech was essential to the curriculum.

Additionally, *Tools* sought to encourage children to work in pairs and use private speech to guide their actions and develop self-regulation. Reflecting these goals are the literacy activities of Graphics Practice (children practiced fine motor movements while stopping and starting along with the music) and Buddy Reading (children took turns in the roles of reader or listener using visual representations of lips and ears to scaffold roles).

Even though overall fidelity of implementation did not relate to children's outcomes, because of their key role in the curriculum, we explored the relations between the fidelity of the make-believe plan and literacy time blocks and children's skills at the end of prekindergarten. These two core components, if enacted by the teachers, might actually be predictive of effects, effects lost when combined with all the other activities. Totaled across all observations, the mean weighted score for the make-believe play activities was 160.15 (*SD* = 72.83) and 31.04 (*SD* = 13.65) for the literacy time block. Variation existed among the teachers in both these activities (i.e., *SD*s approximately 45% of the mean).

Fidelity of the make-believe play activities (centers, planning, practice, and cleanup) did not relate to children's prekindergarten gains in academic (*p*s > .235), executive function and self-regulation (*p*s > .306) skills, or teacher reports of work-related and social-emotional skills (*p*s > .444).

Fidelity of the literacy activities marginally related to less gain on Oral Comprehension (*B* = –0.082, *SE* = 0.045, *p* = .067, *ES* = –.069) but did not relate to the other measures of academic skills (*p*s > .225). Fidelity of implementation of literacy activities significantly negatively related to executive function measures of Peg Tapping (*B* = –0.067, *SE* = 0.019, *p* = .002, *ES* = –.164) and HTKS (*B* = –0.220, *SE* = 0.060, *p* < .001, *ES* = –.174). Thus, even though the *Tools* literacy activities were designed to support turn-taking and in turn executive function skills, we found an opposite effect: the more faithfully the activities were implemented, the lower the gains on Peg Tapping and HTKS. Associations with the other executive function measures were nonsignificant (*p*s > .358), as were associations with teacher reports of work-related and social-emotional skills (*p*s > .147).

Discussion

Determining whether teachers implemented a curriculum to which they were randomly assigned is not an easy task. In the real world, teachers and schools adopt and use curricula without clear parameters as to what qualifies as an adequate level of implementation to yield the expected positive outcomes for their children. Teachers' training and understanding of best practices in early childhood pedagogy may or may not fit the practices outlined in the adopted curriculum or activities. In a rigorous test of a new curriculum that purports to have positive effects on specific skills, establishing clear criteria to identify full implementation is all the more essential. The goals of this chapter were to describe the fidelity of implementation and to examine the associations between curriculum implementation and children's academic, executive function, self-regulation, and social competencies.

Determining what constitutes implementation is also not simple for researchers and therefore methods have not been uniform across various studies involving full or partial prekindergarten *Tools*. Prior randomized control trials of *Tools* attempted to capture fidelity of implementation, with each research team developing different types of measures. A concurrent evaluation of the preschool version of *Tools* by Solomon and colleagues (2018) is most like the study presented in the current monograph, but they focused on about a third of the activities in the curriculum. Solomon et al. identified 21 *Tool*s activities as core to the curriculum and developed a checklist of elements for each activity necessary for fidelity of implementation (total elements expected ranging from 119 to 160 depending on the time of year observed). Reported as percentages of the elements observed, fidelity in *Tools* classrooms ranged from an average of 48.9% to 58.4% across three observations. These findings, albeit on a smaller number of activities, are very similar to ours. Solomon et al. did not examine the association between these implementation percentages and children's outcomes.

The lack of prior examinations of the associations between fidelity of implementation of *Tools* and child outcomes makes it difficult to place our findings in context. It is apparent, though, that the degree of implementation fidelity is at least similar to if not higher than that reported in other studies. An issue with our results is whether implementation fidelity was too low (i.e., insufficient dosage) to achieve positive effects for the curriculum. Implementation of *Tools* activities varied across classrooms ranging between 3 and 20 at a given observation. Variability of this magnitude should relate to child outcomes if the curriculum is effective.

Several teachers simply refused to give up extant activities they believed worked in order to try out the new and complex *Tools* curriculum. Teachers who had low fidelity of implementation tended to be more experienced; the association between lead teachers' overall years of experience and their weighted fidelity score averaged across the three observations was $r = -.379$, $p = .032$. Children in these teachers' classrooms made strong gains across the year. The process the study school systems followed to participate in the study was generally typical of administrative decisions to adopt a new curriculum. In fact, the training procedures were actually much more extensive than most school systems can afford. As described in Chapter II, training was provided, and curriculum-specific in-class coaching occurred for a full year before full implementation and continued in the implementation year. Nevertheless, teachers make their own decisions about how much to implement any new curriculum.

Even willing teachers experienced myriad implementation difficulties with a curriculum as complex as *Tools*. First, no predefined estimate or criterion for adequate implementation existed. Teachers had no clear benchmark to strive for, and trainers and coaches had no clear benchmark to pursue. As researchers, we did not know if the implementation had been all that could be achieved. Although we created several metrics to quantify fidelity of implementation and were able to create benchmarks based on the *Tools* curriculum manuals, we could not know if full implementation was possible or what developers expected. For example, even if classrooms were to reduce their transitions and cease doing non-*Tools* activities, implementing to the degree outlined in the manuals would have been difficult, perhaps impossible. In the *Tools* manual 10 minutes were allocated for transitions to be done in *Tools*-specific fashion. With the number of *Tools* activities among which children must circulate in a day, plus the usual transitions to get to meals and outside, 10 min was unrealistic. In fact, on average, nearly 53 min of the day was observed in transitions in the first observation, reduced to 46 min in Observation 3 (combining *Tools* and non-*Tools* transitions, Table 8).

Second, over the course of this study, the curriculum grew as developers added activities to meet learning standards in different states or due to requests from other nonstudy classrooms implementing *Tools*. These additions did not seem to reference what was actually possible in a classroom. When we wrote the proposal to fund the evaluation, the curriculum consisted of 40 activities; by the time we were funded, there were 61 activities. An additional 3 were added during the year of training before implementation, followed by another 18 during the year. *Tools* developers considered all 61 activities to be part of the curriculum they wished teachers to implement, and they provided training on all of them. As we have seen, because of the time necessary for classroom routines (meals, naps, outside time) and transitions, the actual time teachers had available to implement a curriculum was substantially less than what is necessary for the implementation of the *Tools* curriculum, even if teachers did no non-*Tools* activities.

As the number of activities grew, their sheer number may have worked against the overall *Tools* focus of developing children's self-regulation. In order to engage with the activities, children received specific and sometimes constant directions from their teachers to comply with what was extensive external regulation. The curriculum directed even make-believe play. It did not emerge spontaneously from the activities of the children as much pretend play does. In fact, the *Tools* developers presented play to teachers as a means of behavioral control; during training, teachers were encouraged to hold children accountable for playing the roles they said they were going to enact during play planning.

Third, most teachers spent some time engaging their children in non-*Tools* activities. Such activities could include a teacher's favorite story and or songs that were not part of the curriculum. These activities also included ones related to holiday celebrations, also not in the curriculum. Teachers with the most years of experience were more likely to include these non-*Tools* materials and were less likely to implement *Tools* activities.

Many acknowledge the importance of teacher ownership (e.g., Fantuzzo et al., 2011; Hill & Erickson, 2019; Weiland et al., 2018). The more teachers are allowed to adapt the curriculum to their own preferences, the more likely they will be to enact at least some version of it. However, McMaster et al. (2014) argue that adaptations by teachers should come *after* they have learned to enact the curriculum as intended and then with support from the developers. It is very difficult to test the effectiveness of an approach if teachers begin to modify it immediately.

Fourth, due to the complexity and dynamic nature of the *Tools* curriculum, Diamond et al. (2019) assert implementation works best for teachers with at least a bachelor's degree. All teachers in the current study had a bachelor's degree, making teacher education less a factor impacting implementation in this study. Moreover, other evaluations of the prekindergarten version of *Tools* also had teachers with a bachelor's degree, ranging from 62% percent in Morris et al. (2014) to 100% in Barnett et al. (2008). As there is no variability in education in the current study, it is unfortunately not possible to test the hypothesis about whether teacher education is linked to the fidelity of implementations. Nonetheless, consideration of the complexity of the curricular expectations and whether teachers have adequate prior knowledge and training to implement the curriculum are important to acknowledge.

Although *Tools* does have unique characteristics that make determining the fidelity of implementation challenging, developing clear criteria for fidelity is an essential step for all curricula. Our work demonstrates a careful and systematic attempt to develop indices of fidelity with full involvement from the developers. We think it can be a model for other efforts as the field moves toward more scripted approaches to early childhood education. Each time a curriculum is adopted it may not need this same level of scrutiny, but rigorous, objective evaluation should occur at least once during the development phase. The detailed data we have on each of the 61 activities and how well teachers enacted them could form the basis for the developers to revise their curriculum.

In the early fall following the evaluation year, we presented the evaluation results in detail first to the developers and then in a more extensive meeting to all of the *Tools* national trainers who worked in the field. We were able to review each activity in detail so that the groups could see where teachers were having difficulties. We presented the trainers with a *Tools Trainers Report* that included summaries of all the results and embedded lists of takeaway thoughts. These were provided in an attempt to help the trainers focus on issues that would profit from their attention.

Feedback about how well a curriculum works is critical, particularly when it may not be achieving what the developers hoped. Curricula are sold to school systems without having been completely tested. Often only a few activities at best have been empirically tested to determine if a curriculum can be implemented within the constraints of current classrooms. In this era of heightened concern about the effects of preschool and prekindergarten programs, it is important to require that curriculum developers implement their full curriculum experimentally and evaluate not only the effectiveness of the curriculum but also the feasibility of faithful implementation. And it is important to revisit the curriculum as necessary once feedback comes in.

Almost all of the scholarly literature on fidelity takes the perspective that *lack* of fidelity is important in explaining null or negative findings. It is important to entertain the possibility, however, that fully implementing the intervention may not lead to the desired effects. In that case, it seems incumbent upon the developers to revisit their theory of change or to reconsider the activities they developed to reflect their theory. The *Tools* curriculum is designed to equip children with cognitive *tools* for learning that they can then apply to the task of acquiring and sustaining academic knowledge and skills. The *Tools* approach is supported by a theory of change hypothesizing that enhancing children's executive function and self-regulation skills will subsequently benefit children's academic knowledge. This theory is supported by correlation research indicating significant associations between the development of executive function and self-regulation skills and academic skills (e.g., Fuhs et al., 2014; McClelland et al., 2007; Schmitt et al., 2017; Welsh et al., 2010). Without clear experimental evidence to support a causal relation, however, this theory of change remains insufficiently tested (e.g., Bailey et al., 2018; Jacob & Parkinson, 2015).

The next chapter examines fidelity from a different lens, with a focus on how *Tools* classrooms differed from control classrooms with respect to more general types of classroom practices and processes. Based on expectations of the developers of the curriculum, these processes should distinguish a classroom carrying out *Tools* from a control classroom as a function of, but in addition to, the curriculum activities themselves.

IV. Classroom Processes and Associations With Child Outcomes

Curricula are created to improve classroom practices and interactions as a means to improve child outcomes. It stands to reason then that a curriculum should be associated with changes in targeted classroom processes and that those classroom processes will be associated with child outcomes. In this chapter, we first describe how we collaborated with the developers of *Tools* to identify and test hypotheses about expected curriculum effects on classroom processes. Second, we investigate whether the identified classroom processes are associated with gains in children's academic, executive function, self-regulation, and social skills.

Aim 1. Differentiation of Classroom Processes by Condition

The implementation of an intervention can be measured in many different ways; in Chapter III, we reported on elements of fidelity designed to examine dosage and adherence to the *Tools of the Mind* (Bodrova & Leong, 2007) curriculum in the intervention classrooms. We indexed dosage by the amount of time teachers devoted to delivering the curriculum; we indexed adherence by observing whether specific activities and steps prescribed by the curriculum were enacted. We now turn to another approach to understanding the impact of a curriculum on classroom practices which asks if classrooms assigned to the intervention differed from classrooms assigned to the control condition on general types of classroom practices or processes (van Dijk et al., 2019). We note that our term "general" is intended to refer to denote practices and processes that are part of virtually any classroom, irrespective of instructional approach (e.g., talk by both teachers and students).

Curriculum evaluators typically give little attention to examining general classroom practices that are not explicitly targeted by the curriculum being evaluated (Missett & Foster, 2015; van Dijk et al., 2019). Variations in general classroom practices could, however, help to explain patterns of effects—positive or negative—on children's outcomes. For example, learning and delivering a new curriculum could be stressful for teachers which may change the classroom climate. This change might serve to lower children's performance, perhaps obscuring what might otherwise have been a positive effect of the curriculum itself.

Differentiation in Prior Tools *Evaluations*

Some earlier studies have examined differentiation between *Tools* and control classrooms. In the first randomized control trial of the *Tools* preschool curriculum, Barnett et al. (2008) evaluated whether *Tools* and control classrooms differed in the global quality of general instruction by using the *CLASS* (Pianta et al., 2008) and the *Early Childhood Environmental Rating Scale-Revised* (*ECERS-R*; Harms et al., 1998). Each of the systems was used once, always in the morning of the school day. For the *CLASS* dimension of productivity, differences favored *Tools* classrooms. This dimension, termed Organization in newer versions of CLASS, tapped how efficiently teachers manage children's time, including efficient transitions and teacher preparation of materials. *Tools* and control classrooms did not differ on any of the other subscales or the overall *CLASS* score. Morris et al. (2014) also used *CLASS* to assess the *Tools* make-believe play enhancement in Head Start classrooms but found no differences between *Tools* and control classrooms for any of the subscales or the overall score.

When looking at only the morning routines, Barnett et al. (2008) found significant differences in the *ECERS-R* ratings favoring *Tools* classrooms over control. The differences were driven by the subscales of Language-Reasoning (encouraging children to communicate and teachers using language to develop reasoning skills) and Activities (provision of various activities across a variety of developmental domains).

In a later randomized control trial of *Tools*, Morris et al. (2014) conducted a more focused approach to identify specific classroom practices that might differentiate *Tools* from business-as-usual classrooms. In the spring of the school year, teachers were observed to have provided more scaffolding of children's pretend play and peer interactions in *Tools* classrooms compared with control. The investigators found no differences, however, in classroom management or teachers' social-emotional instruction. Two studies showed that less time was spent in whole-group activities in *Tools* than in control classrooms (Barnett et al., 2008; Diamond et al., 2019), and one reported that *Tools* classrooms (unlike control classroom) abstained from the use of rewards and time-outs (Diamond et al., 2019).

Current Study

As described in Chapter III, when looking solely at the classrooms implementing *Tools*, there were no consistent associations between children's gains in any area and either the amount of time the curriculum was implemented (dosage) or the fidelity with which teachers carried out the activities (adherence). In this chapter, we examine whether the curriculum affected general classroom processes that the *Tools* developers hypothesized would distinguish a classroom carrying out *Tools* from a control classroom.

Differentiation Hypotheses Identification Procedure

As described earlier, before classroom evaluations were begun, members of the research team gathered with *Tools* developers and national *Tools* curriculum trainers in a 4-day meeting to develop both the fidelity of implementation system and the plan to assess differences between *Tools* and control classroom practices. Project staff and the curriculum developers and trainers discussed (a) the important aspects of the curriculum that set it apart from other early childhood curricula, and (b) how these characteristics could be measured or quantified. The observation scheme was created by thinking about ways to measure the aspects of the curriculum (i.e., behaviors and materials) which the developers and trainers believed were important to *Tools* and should be present in every classroom enacting *Tools.* During this meeting, researchers also visited two classrooms in Massachusetts being taught by experienced *Tools* teachers.

From this meeting, developers and trainers collectively created and agreed upon specific hypotheses about what they expected would differ between *Tools* and control classrooms. Most of the hypotheses the developers generated involved conditional probabilities (e.g., teachers will be more positive during instruction). The six key hypotheses generated by the curriculum developers are described in turn below.

Hypothesis 1. The pattern of talk for children will be different in *Tools* classrooms from control classrooms. Namely, (1) there will be more instances of child-to-child talk in *Tools* classrooms; (2) children who are talking to each other will be more likely to have a learning focus (all content areas) in *Tools* classrooms; (3) children in *Tools* classrooms will more often talk to themselves; (4) children will more often be observed listening to other children in *Tools* classrooms, and (5) children will talk more during associative interactions in *Tools* classrooms than in control classrooms.

Hypothesis 2. The pattern of talk for teachers will be different in *Tools* classrooms from control classrooms. Specifically, (1) teachers will talk less in *Tools* classrooms than in control classrooms; (2) there will be a better balance between teacher and child talk in *Tools* classrooms; (3) teachers will talk less to children during management in *Tools* classrooms compared with comparison classrooms; and (4) teachers will talk more with children during center time in *Tools* classrooms than in control classrooms.

Hypothesis 3. Children's interpersonal interactions with teachers and peers will differ in *Tools* classrooms compared with control classrooms: (1) children will more often engage in associative and cooperative interactions in *Tools* classrooms; and (2) children will less often engage in parallel interactions in *Tools* classrooms compared with control classrooms.

Hypothesis 4. Children's learning behaviors will differ in *Tools* classrooms compared with control classrooms: (1) overall, children will be more highly involved in *Tools* classrooms; (2) specifically during large group instruction and centers, children will be more involved in *Tools* classrooms; (3) children will less often be disruptive in *Tools* classrooms; and (4) children in *Tools* classrooms will be unoccupied less often than in control classrooms.

Hypothesis 5. The socio-emotional climate in the classroom and teachers' instructional behaviors will be different in *Tools* classrooms compared with control classrooms. In particular, (1) teachers in *Tools* classrooms will have a warmer emotional tone; (2) teachers in *Tools* classrooms will engage in more approving behavior; (3) teachers in *Tools* classrooms will engage in less disapproving behavior; (4) teachers in *Tools* classrooms will engage children in conversations and learning opportunities that encourage inferential thinking rather than encouraging them to retrieve already stored knowledge or exercise scripted basic skills; (5) teachers will engage in fewer management behaviors in *Tools* classrooms than in control classrooms.

Hypothesis 6. The classroom day will be organized differently in *Tools* classrooms compared with control classrooms. Specifically, in *Tools* classrooms there will be (1) less didactic teaching (i.e., teacher-led instruction); (2) more center time; and (3) less time in spent in transitions.

Methods

Participants

To examine differences in classroom processes across curriculum conditions, the analyses in this chapter include all 60 study classrooms (32 *Tools*, 28 control) and the 877 children (498 *Tools*; 379 control) in the study (as described in Chapter II). We observed all children in each classroom but identified only consented children, a procedure that provided a complete picture of classroom interactions.

Procedures

Daylong observations were made by two observers on occasions during fall, mid-winter, and spring of the implementation year. One observer focused on fidelity of implementation of the curriculum as part of creating the *Narrative Record*, described in Chapter III. The second observer focused on the type and quality of interactions among members of the classroom using the *Teacher Observation in Preschool* (TOP, Bilbrey et al., 2011) and the *Child Observation in Preschool* (COP, Farran, 2011) instruments. TOP and COP observers did not participate in training sessions on the curriculum itself. However, as noted earlier, *Tools'* activities are so recognizably different from other early childhood classroom activities that observers could not be blind to condition. Their focus, though, was on classroom practices and interactions that are common in all early childhood classrooms, not unique to *Tools*.

Measures

Teacher observation in preschool and child observation in preschool. The TOP and COP were used to test many of the hypotheses. These instruments have previously been shown to capture unique and important aspects of quality in 4-year-old prekindergarten classrooms (Farran et al., 2017; Fuhs et al., 2013; Spivak & Farran, 2016). Some adaptations to TOP and COP were made for evaluating *Tools*. For example, in both TOP and COP, make-believe play was added as a separate option under the set of schedule codes to distinguish this kind of play from the more general option of center time. Other adaptations involved adding detail to the coding manuals to include guidance for coding interactions that might be unique to *Tools* classrooms (manuals with complete descriptions are available online under Resources at https://my.vanderbilt.edu/toolsofthemindevaluation/).

TOP and COP use a snapshot behavior-sampling procedure to capture teacher and child behaviors. Observers progress through a series of 20 rounds of coding, first coding the primary teacher and then the assistant teacher(s) in what we call a sweep, followed by each individual child in the classroom before returning to the teacher to start another round. For each sweep, a classroom member is located, observed, and then, after a count of approximately 3 s, coded across an array of dimensions. When aggregated, the collection of snapshots provides a picture of how members of a classroom spend their time across the full day. For the current study, we included only the primary teachers' TOP data.

Coding occurred throughout the school day, apart from outdoor recess, meals, and naptime. Continuous coding ensures that individuals will be observed across multiple contexts (e.g., large groups, centers, transitions) and instructional content. Coding options for each dimension are mutually exclusive. We created the analytic variables by first computing the sum of individual scores across the three daylong observations (fall, winter, and spring), and then aggregating them to the classroom level to provide a picture of classroom processes. Aggregated sums of the behavioral count variables were the proportions of sweeps in which the target behavior occurred out of the total number of sweeps observed. The variables of teacher tone (TOP) and level of involvement (COP) are Likert ratings and were averaged across all sweeps observed.

Further, we created scores to capture the conditional probabilities of many behaviors (e.g., children's involvement compared with various types of activities). Most scores were based on summed and aggregated behavioral counts including the probability scores. The variable involving the teachers' level of instruction (TOP) was relevant only when instruction occurred. The level of instruction was scored on a scale ranging from 1 (*interaction with child and activity*) to 4 (*high inferential instruction*) and then scores were averaged across all instances of instruction. A score of 2.0 signified basic instruction (e.g., What color is this? What letter is this?).

Cohen's κ estimated interrater reliability for the dimensions using behavioral counts used in our analyses. Kappa coefficients for the two TOP behavioral counts included in the analyses were .932 for schedule and .860 for teacher task. To estimate interrater reliability for dimensions using rating scales, we estimated intraclass coefficients (ICC; one-way random effects, absolute agreement). TOP ICC estimates were .908 for level of instruction and .799 for teacher tone. Cohen's κ estimates of reliability for the COP dimensions of verbal, interaction state, and type of task were .790, .879, and .846, respectively. For the rating scale variable of level of involvement, the ICC estimate was .887.

Narrative Record. One observer in the classroom used the *Narrative Record* (with adaptations for use in the *Tools* curriculum evaluation; Farran et al., 2010) to record the amount and timing of a variety of activities throughout the school day. The *Narrative Record* provided a continuous record of episodes that characterized the classrooms in terms of pedagogy and academic content (see Chapter III for a detailed description of the measure). Observers that completed the *Narrative Record* also completed the fidelity of implementation instrument. *Narrative Record* observers were provided training on the *Tools* curriculum in order to complete the fidelity of implementation instrument. As noted, these observers therefore were not blind as to condition. However, the coding of time spent in various types of activities and the record of curriculum activities and steps were objective measures, less subject to observer bias.

We used ICC to estimate interrater reliability for the continuous measure of time from the *Narrative Record*. Based on guidance from McGraw and Wong (1996), we conducted a two-way random effects model (raters were randomly assigned to a reliability session as part of our larger population of raters), with single rater/measurement and absolute agreement parameters. ICCs for variables examined were: large group = .901, small group = .941, centers = .997, and transitions = .956.

Testing Hypotheses With Variables From COP, TOP, and Narrative Record

To test Hypothesis 1, COP codes captured children's verbal and listening behaviors to determine if there was more child to child talk in *Tools* classrooms, more private talk, more children listening to other children, and when talking occurred, whether it happened more in a learning context and during associative interactions. Teacher talk, assessed by TOP, is the focus of Hypothesis 2, including a lower amount of teacher talk overall and teachers talking less to children during behavior management but more during centers.

One focus of the *Tools* curriculum is to increase the occurrence of social learning opportunities (Hypothesis 3). The interaction state dimension of COP codes the proportion of associative interactions (children are interacting with another member of the classroom in the context of an activity or task that does not have predetermined rules) and cooperative interactions (children interacting with another member of the classroom in an activity or task with predetermined rules). Both should happen more often in *Tools* classrooms while parallel interactions should occur less often. Parallel interactions occur when children have the same materials or are engaged in the same activity but not working together or co-creating it. Examples are (1) a large group activity where children are listening to a story and (2) children occupying the block corner but each playing independently with the materials.

Hypothesis 4 predicted that higher levels of children's involvement (overall and in both large groups and centers) would occur in *Tools* classrooms compared with withcontrol classrooms. Level of involvement was coded on a 5-point scale, where 1 corresponded to low engagement (e.g., completely off task, not paying attention to the activity), 3 corresponded to medium engagement (e.g., on task, maintaining eye contact with the teacher, participating but might briefly look around but immediately comes back to the task), and 5 corresponded to high engagement (e.g., intense focus, serious persistence and pursuit of an activity, very difficult to be distracted from the activity). We examined the level of involvement as a conditional probability score—the average rating during a learning opportunity. We also coded whether children were unoccupied or disruptive, both of which should happen less often in *Tools* classrooms.

Hypothesis 5 involves teacher behaviors observed with TOP. The first three predictions within this hypothesis relate to the emotional climate of the classroom and involve three variables from TOP. The first was the teacher's emotional tone that could range from 1 (*extremely negative affect*) to 5 (*vibrant and enthusiastic affect*). The percentage of the observation the teacher spent approving children's behavior (making approving verbal comments, facial expressions, or physical contact with the children to express positivity) and the percentage of the observation the teacher spent disapproving behavior were the second and third variables involved in the overall emotional climate of the classrooms. TOP captured teachers' use of behavior management, predicted to be lower in *Tools* classrooms. Also, in Hypothesis 5, the level of instruction provided by the teachers was predicted to be higher in *Tools* classrooms. The definition of the term "instruction" in an early childhood classroom is broad and inclusive, including typical academic activities as well as more general activities such as art, music, or free play. In short, instruction occurs whenever teachers are engaged with children in a learning opportunity.

Hypothesis 6 was related to the distribution of time in the classroom. The *Narrative Record* provided all variables to investigate this hypothesis. We calculated the amount of time children spent in center activities (when children could choose freely from a variety of available activities), predicted to occur more often in *Tools*. We calculated the amount of time children spent in teacher-led instruction (e.g., whole group and small group instruction) and in transitions, each of which should occur less often in *Tools* classrooms.

Analytic Strategy

As represented in Equation (4), to estimate the effects of the experimental condition (*Tool*s vs. control) on the hypothesized differences (Hypothesized_Difference$_{ij}$), we employed two-level nested regression models, classrooms at Level 1 (classroom$_{ij}$), and randomization blocks at Level 2 (block$_{j}$) in SPSS Version 22. The dichotomous variable of curriculum condition was entered at the classroom level with *Tools* being the reference group ($\gamma_{10} \times$ condition$_{ij}$). All dependent variables examined were differences at the classroom level and averaged across the three observations, with *COP* variables aggregated to the classroom level. As analyses are at the classroom level, significance was based on the classroom sample ($N = 60$), which is a relatively small sample, thus associations with p values of .10 or lower are discussed.

$$\text{Hypothesized\_Difference}_{ij} = \gamma_{00} + \gamma_{10} \times \text{condition}_{jk} + U_{0j} + r_{ij}. \tag{4}$$

Cohen's d standardized mean difference effect sizes estimated the magnitude of condition difference (i.e., the difference between the *Tools* and control classroom's means divided by the pooled standard deviation of *Tools* and control classrooms).

Results

Child and Teacher Talk

Hypothesis 1 involved verbal interactions of children to adults and peers in the classrooms. As presented in Table 13, across both conditions, children talked in approximately one-quarter of the sweeps (25% for *Tools* and 25% for control classrooms). As predicted, children in *Tools* classrooms talked significantly more when there was a content focus ($d = 1.137$); detailed statistics are presented in Table 13. The amount children talked overall, to themselves, and during associative interactions did not differ significantly across classroom conditions. Children also did not vary significantly by condition in the amount of listening observed.

Hypothesis 2 involved verbal interactions of teachers in the classrooms (Table 13). Across both conditions, teachers talked in a majority of the sweeps (73% for *Tools* and 70% for control classrooms) and did not differ significantly by condition. Across both conditions, teachers did not differ in how much they talked overall to children, either during management interactions, or during center interactions. Likewise, when pooling across all locations and activities, we found no significant differences between conditions in ratio of the proportion of sweeps children were observed talking to the proportion of sweeps the lead teacher was observed talking; the ratio was 1 to 3.5 in *Tools* classrooms and 1 to 3.6 in control classrooms.

TABLE 13

Children and Teacher Talking by Condition

	Tools ($n = 32$)		Control ($n = 28$)		Test of Condition Effects		
	M	*SD*	*M*	*SD*	*B*	*SE*	*d*
Hypothesis 1. Child Talk[a]							
Child Talk Overall	.253	.039	.254	.049	−.004	.010	−0.030
Child Talk to Child with Content Focus	.112	.046	.068	.030	.044*	.011	1.137
Child Talk to Self	.065	.016	.064	.019	.002	.005	0.035
Child Listen to Another Child	.075	.022	.065	.016	.009	.006	0.505
Child Talk During Associative Interaction	.493	.081	.535	.108	−.042	.010	0.452
Hypothesis 2. Teacher Talk[b]							
Teacher Talk Overall	.727	.085	.696	.081	.032	.022	0.386
Teacher Talk to Children During Management	.100	.065	.121	.077	−.023	.016	−0.302
Teacher Talk to Children During Centers	.577	.211	.547	.185	.024	.056	0.152

Note. Tools of the Mind was the reference group in each model as such positive coefficient indicate *Tools* classrooms had higher rates of an observed variable compared with the control classrooms. Estimates are the proportion of sweeps across the three observations a given behavior was observed. Coefficients reported are unstandardized regression coefficients from multilevel regression models. Estimates of effect sizes of curriculum differences provided are Cohen's *d* estimates standardized mean difference.
SD = standard deviation; *SE* = standard error.
[a]Variable captured using the Child Observation.
[b]Variables captured using the Teacher Observation Protocol.
*$p < .05$.

Child Interactions and Involvement Levels

Hypotheses 3 and 4 each involved dimensions of children's behaviors (Table 14). As predicted children in *Tools* classrooms significantly more often engaged in both associative ($d = 1.054$) and cooperative interactions ($d = 1.081$) compared with children in control classrooms. However, the amount of time spent in parallel interactions, the most common type of interaction observed, did not differ by condition. Associative interactions occurred in about 10% of sweeps in *Tools* classrooms (compared with 7% of control classrooms sweeps) and cooperative interactions occurred in about 2% of sweeps in *Tools* classrooms (less than 1% in control classrooms).

Contrary to predictions, there were no differences in the proportion of sweeps children were unoccupied or disruptive (Hypothesis 4). Indeed, irrespective of condition, children were unoccupied only about 5% of the time and were rarely observed being disruptive.

Hypothesis 4 also examined levels of children's involvement, shown in Table 14. Overall the children's involvement across the day was the same for children in *Tools* and control classrooms. However, children in *Tools*

TABLE 14

Children's Interactions and Learning Behaviors by Condition

	Tools (*n* = 32)		Control (*n* = 28)		Test of Condition Effects		
	M	*SD*	*M*	*SD*	*B*	*SE*	*d*
Hypothesis 3. Children's Interaction							
Associative Interactions	0.104	.039	0.068	.028	.035*	.010	1.054
Cooperative Interactions	0.021	.018	0.006	.008	.015*	.004	1.081
Parallel Interactions	0.470	.051	0.479	.053	−.014	.015	−0.175
Hypothesis 4. Children's Learning Behaviors							
Overall Level of Involvement[a]	2.379	.167	2.326	.201	.045	.048	0.293
Large Group Level of Involvement[a]	2.692	.179	2.571	.201	.121*	.050	0.545
Centers Level of Involvement[a]	2.924	.166	2.834	.170	.090*	.042	0.649
Unoccupied	0.047	.018	0.045	.023	.001	.005	0.100
Disruptive	0.009	.006	0.007	.006	.003	.001	0.333

Note. Tools of the Mind was the reference group in each model as such positive coefficient indicate *Tools* classrooms had higher rates of an observed variable compared with the control classrooms. Coefficients reported are unstandardized regression coefficients from multilevel regression models. Estimates of effect sizes of curriculum differences provided are Cohen's *d* estimates standardized mean difference. All variables captured using the Child Observation Protocol. Unless noted, estimates are the proportion of sweeps across the three observations a given behavior was observed.

SD = standard deviation; *SE* = standard error.

[a]Level of involvement was scores on a 1–5 rating scale.

*$p < .05$.

classrooms were significantly more involved when they were in centers (*d* = .649) and in large groups (*d* = .545) than were children in the control classrooms.

Teacher Behaviors and Classroom Organization

As the *Tools of the Mind* curriculum aimed to facilitate teachers' ability to effectively scaffold children's learning and development, the developers hypothesized that teachers' classroom behaviors and instructional practices in *Tools* classrooms would differ from control classrooms in key ways (Hypotheses 5 and 6). Data related to these hypotheses are presented in Table 15. We found no significant differences between conditions in teacher behaviors. Teachers in *Tools* and control classrooms did not differ in their positive emotional tone or the amount of their approving and disapproving behaviors or behavior management. Also, in contrast to predictions, the mean level of instruction provided by teachers did not distinguish *Tools* classrooms from control classrooms. In each, the quality of instruction was relatively low, that is, basic skills or below.

Hypothesis 6 related to the organization of the classroom at the level of the entire classroom day. As presented in Table 15, contrary to predictions, *Tools* classrooms spent significantly more time in teacher-led instruction

TABLE 15

Teacher Classroom Behaviors and Instructional Practices by Condition

	Tools ($n = 32$)		Control ($n = 28$)		Test of Condition Effects		
	M	*SD*	*M*	*SD*	*B*	*SE*	*d*
Hypothesis 5. Teacher Behaviors							
Emotional Tone (rating 1–5 scale)[a]	3.433	.158	3.376	.200	.059	.047	0.324
Approving Behavior[a]	0.034	.031	0.046	.029	−.013	.008	−0.415
Disapproving Behavior[a]	0.058	.048	0.059	.055	−.003	.013	−0.020
Level of Instruction (rating 1–4 scale)[a]	1.890	.140	1.891	.116	−.003	.036	−0.009
Behavior Management[a]	0.223	.084	0.228	.085	−.005	−.004	−0.060
Hypothesis 6. Classroom Organization							
Time in Teacher-led Instruction[b]	0.256	.071	0.173	.067	.075*	.013	1.215
Time in Centers[b]	0.117	.044	0.159	.061	−.037*	.012	−0.813
Time in Transition[b]	0.131	.040	0.133	.038	−0.001	.010	−0.023

Note. Tools of the Mind was the reference group in each model as such positive coefficient indicate *Tools* classrooms had higher rates of an observed variable compared with the control classrooms. Coefficients reported are unstandardized regression coefficients from multilevel regression models. Estimates of effect sizes of curriculum differences provided are Cohen's *d* estimates standardized mean difference.

SD = standard deviation; *SE* = standard error.

[a]Variable captured using the Teacher Observation Protocol (unless noted estimates are the proportion of sweeps across the three observations a given behavior was observed).

[b]Variables captured using the Narrative Record (estimates are the proportion of time across the three observations a given instructional approach was observed).

$^{*}p < .05$.

($d = 1.215$) and less time in centers ($d = -0.813$). Teacher-led large group and small group instruction occurred 25% of the day in *Tools* classrooms (compared with 16% for control classrooms). Center activities occurred 12% of the day in *Tools* (compared with 16% for control classrooms). Contrary to predictions, no significant differences were found between classroom conditions with respect to the amount of time in transitions.

Aim 2: Associations Between Classroom Processes and Child Outcomes

The *Tools* developers had proposed hypotheses that focus on the kinds of dynamic, day-to-day interactions which normally occur in early-childhood classrooms (Howes et al., 2008; Pianta et al., 2005; Valentino, 2017). Although our results were not generally consistent with developers' expectations about how *Tools* and control classrooms would differ, other research literature has established links between similar classroom processes and gains in young children's academic, executive function, self-regulation, and social skills (e.g., Hamre et al., 2014; Farran et al., 2017; Fuhs et al., 2013; Keys et al., 2013;

Spivak & Farran, 2016; Weiland et al., 2013). Had the curriculum brought about significant changes in classroom processes in the samples of classrooms we studied, perhaps there would also have been observable curriculum effects on child outcomes.

Despite our failure to find many differences between *Tools* and control classrooms, the data presented in Tables 13–15 show large variations in classroom processes within participating classrooms. It is possible that the developers correctly identified behaviors that are important for children's growth and development even though the *Tools* curriculum did not affect them. Thus, we also conducted analyses to test whether variability in those classroom processes related to child outcomes.

Methods

Because (as just noted) there were few differences in relevant classroom processes in *Tools* versus control classrooms, we pooled data across instructional conditions to examine the association between classroom interaction patterns and child outcomes for the participants as a whole. For parsimony, we examined a subset of the variables that had been identified in previous research as important for child growth across the prekindergarten and kindergarten years (e.g., Christopher & Farran, 2020; Dickinson, 2011; Farran et al., 2017; Justice et al., 2008; Nesbitt et al., 2015).

Instead of testing all possible combinations of child and teacher talk, we focused on the overall levels of talking observed. We created an estimate of the emotional climate of the classroom as an equally weighted composite from the standardized scores (*z*-scores) of teacher tone; the proportion of sweeps a teacher was observed approving behavior, and the proportion of sweeps in which the teacher was observed engaging in disapproving behavior. Tone and approving behavior contributed positively to the composite whereas disapproving behavior contributed negatively so that higher scores indicated a more positive climate. We combined children's participation in cooperative and associative interactions to reflect children's participation in social learning interactions, and we combined children's unoccupied and disruptive to index off-task behaviors.

Analytic Strategy

As represented in Equation (5), to estimate the association between classroom processes (e.g., child and teacher behaviors) on child outcomes, we employed three-level nested regression models with children at Level 1 (children$_{ijk}$), classrooms at Level 2 (classroom$_{jk}$), and randomization blocks at Level 3 (block$_k$) in SPSS Version 22. The child and teacher behaviors were entered at Level 2 ($\gamma_{010} \times$ classroom_process$_{jk}$). All analyses of achievement outcomes used the Woodcock–Johnson W-scores, which are IRT scaled but not adjusted for age. All other outcomes remained in their raw score form. Each impact model accounted for pretest scores, age at pretest, the interval between assessments, gender, home language and IEP status at the student level in the model. The pretest, age, and time interval covariates were grand-mean centered. Gender (0 = *male*), home language (0 = *English*), and IEP (0 = *no IEP*) status covariates were dichotomous covariates. We report the results for each outcome variable separately. As analyses are at the classroom level, significance was based on the classroom sample ($N = 60$), which is a relatively small sample. As we consider these analyses exploratory, we report associations with p values of <.10, although we recognize the need for caution and for replication.

$$\begin{aligned}\text{Posttest}_{ijk} = {} & \gamma_{000} + \gamma_{100} \times \text{pretest}_{jk} + \gamma_{200} \times \text{age}_{ijk} + \gamma_{300} \times \text{interval}_{ijk} \\ & + \gamma_{400} \times \text{gender}_{ijk} + \gamma_{500} \times \text{lang}_{ijk} + \gamma_{600} \times \text{iep}_{ijk} \\ & + \gamma_{010} \times \text{classroom\_process}_{jk} + U_{00k} + U_{0jk} + r_{ijk}. \qquad (5)\end{aligned}$$

To estimate the impact of the classroom processes, we multiplied the unstandardized coefficient (B) for the classroom process indicator from the multilevel regression models by the standard deviation of the classroom process indicator and then divided by the standard deviation of the child outcome ($ES = (B_{\text{classroom\_process}} \times SD_{\text{classroom\_process}})/SD_{\text{child\_outcome}}$); see National Institute of Child Health and Human Development Early Child Care Research Network & Duncan, 2003 for an overview of the procedure). The standardized effect size (ES) was calculated to indicate the change in a child's outcome in standard deviation units when a classroom process variable increased by one standard deviation.

Results

Child and Teacher Talk

Table 16 presents the relation between the overall proportions of sweeps the lead teacher and children spent talking and children's prekindergarten gains in both academic and executive function outcomes. For academic outcomes, the amount teachers talked was not associated with children's gains on any outcome. Child talk, however, was positively related to one of the outcomes, Spelling ($ES = 0.090$).

Higher rates of teachers' talking positively related to children's gains on the HTKS ($ES = .074$). Contrary to expectations, classrooms with more child talk made smaller gains on HTKS ($ES = -.069$), and Corsi Backward Span ($ES = -.092$) but higher gains on DCCS ($ES = .083$).

Teacher Emotional Climate and Instruction Level

As seen in Table 17, the classroom emotional climate positively related to children making larger gains on the academic subscales of Letter-Word ($ES = .107$), Spelling ($ES = .120$), Academic Knowledge ($ES = .082$), and Picture Vocabulary ($ES = .043$), all areas related to literacy development.

TABLE 16
Teacher and Child Talk Effects on End of Prekindergarten Child Outcomes

		Teacher Talk			Child Talk		
Variable	*n*	*B*	*SE*	*Effect Size*	*B*	*SE*	*Effect Size*
Letter-Word	813	−0.763	11.132	−.003	3.488	22.683	.006
Spelling	813	1.908	13.538	.006	55.891*	27.191	.090
Academic Knowledge	813	−1.043	5.969	−.005	0.401	13.323	.001
Oral Comprehension	813	4.251	4.430	.022	−4.462	10.839	−.011
Picture Vocabulary	813	1.592	3.818	.009	0.770	8.587	.002
Applied Problems	813	1.434	6.666	.006	−16.895	16.185	−.032
Quantitative Concepts	813	4.217	5.986	.024	−15.027	12.988	−.042
Copy Design	813	0.132	1.359	.004	1.011	3.146	.015
Corsi Forward Span	813	0.256	0.461	.019	−2.524*	0.994	−.092
Corsi Backward Span	813	0.741	0.561	.047	−0.421	1.291	−.013
DCCS	813	0.506	0.264	.073	1.153†	0.612	.083
HTKS	813	14.947*	6.139	.074	−28.145†	14.153	−.069
Peg Tapping	813	1.723	2.422	.025	−6.171	5.310	−.045
Interpersonal Skills	821	0.035	0.528	.003	0.801	1.083	.032
Work-Related Skills	821	−0.277	0.670	−.020	1.010	1.369	.036

Note. Coefficients are unstandardized regression coefficients and standard errors from multilevel regression models that account for nesting of teachers in schools and random assignment blocks. Covariates included in the models were pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest. Standardized effect sizes calculated indicated the change in a child's outcome in standard deviation units when the process variable increased by one standard deviation.
DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders; *SE* = standard error.
*$p < .05$.
†$p < .10$.

Associations with gains on the mathematics measures were not significant. Significant and moderately positive associations between emotional climate and executive function gains were most robust for Backward Digit Span (*ES* = .089), DCCS (*ES* = .067), and HTKS (*ES* = .060).

The level of teacher's instruction showed a similar pattern of positive associations as emotional climate, but to a somewhat lesser degree. For executive function gains, the effect sizes were largest for Forward Digit Span ($p = .078$, *ES* = .051), HTKS ($p = .046$, *ES* = .054), and Peg Tapping ($p = .031$, *ES* = .074).

Children's Proportion of Day Spent in Centers and Transitions

Table 18 presents the associations between children's prekindergarten gains and the proportion of the day classrooms spent in child-directed instruction (i.e., in centers and small groups with centers) and non-instructional transitions. More time spent in centers (child-directed instruction) related to larger gains on Letter-Word ($p < .001$, *ES* = .168) and Spelling ($p = .093$, *ES* = .089). Yet for Oral Comprehension, child-directed instruction related to less gain ($p = .024$, *ES* = −.071). Time in centers showed no statistically

TABLE 17

Teacher Emotional Climate and Level of Instruction Effects on End of Prekindergarten Child Outcomes

		Emotional Climate			Level of Instructions		
Variable	*n*	*B*	*SE*	*Effect Size*	*B*	*SE*	*Effect Size*
Letter-Word	813	1.137*	.390	.107	8.268	7.414	.044
Spelling	813	1.441*	.475	.120	4.028	9.101	.019
Academic Knowledge	813	0.649*	.214	.082	0.788	4.057	.006
Oral Comprehension	813	0.271	.182	.035	2.128	3.024	.016
Picture Vocabulary	813	0.300*	.146	.043	$4.253^{\dagger}$	2.546	.035
Applied Problems	813	0.108	.274	.011	3.563	4.550	.020
Quantitative Concepts	813	0.309	.227	.045	5.900	3.925	.049
Copy Design	813	0.026	.053	.020	0.747	0.904	.033
Corsi Forward Span	813	0.028	.018	.053	0.470	0.316	.051
Corsi Backward Span	813	0.055*	.023	.089	0.350	0.392	.032
DCCS	813	$0.018^{\dagger}$	.011	.067	0.104	0.189	.022
HTKS	813	$0.479^{\dagger}$	.249	.060	$7.583^{\dagger}$	4.225	.054
Peg Tapping	813	0.076	.094	.029	3.448*	1.563	.074
Interpersonal Skills	821	0.016	.020	.033	−0.047	0.349	−.005
Work-Related Skills	821	0.004	.025	.007	−0.054	0.444	−.006

Note. Coefficients are unstandardized regression coefficients and standard errors from multilevel regression models that account for nesting of teachers in schools and random assignment blocks. Covariates included in the models were pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest. Standardized effect sizes calculated indicated the change in a child's outcome in standard deviation units when the process variable increased by one standard deviation.

DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders; *SE* = standard error.

*$p < .05$.

$^{\dagger}p < .10$.

significant or marginally significant associations with executive function skills or teacher reports of learning-related and social skills. Time spent in transition was positively associated with children's gains on Quantitative Concepts (*ES* = .058) and negatively associated with Copy Design (*ES* = −.094) and executive function.

Children's Involvement Levels, Off-task Behaviors, and Social Learning Interactions

Table 19 presents associations between children's level of involvement, off-task behaviors, and social learning interactions and children's academic and executive function gains. Classrooms with higher levels of involvement most robustly related to the academic measures of Letter-Word (*ES* = .117) and Academic Knowledge (*ES* = .078).

Classrooms with more instances of children engaging in social learning interactions made more gains in Copy Design (*ES* = .099), Corsi Forward (*ES* = .069), and Backward Span (*ES* = .081). In addition, social learning interactions significantly related to children's prekindergarten gains on Letter-Word (*ES* = .121).

TABLE 18
INSTRUCTIONAL TIME EFFECTS ON END OF PREKINDERGARTEN CHILD OUTCOMES

		Time in Centers			Time in Transitions		
Variable	*n*	*B*	*SE*	*Effect Size*	*B*	*SE*	*Effect Size*
Letter-Word	813	51.397*	13.084	.168	−2.270	24.216	−.004
Spelling	813	30.837†	17.836	.089	−26.308	29.683	−.039
Academic Knowledge	813	12.181	8.628	.053	−7.884	13.480	−.018
Oral Comprehension	813	−15.608*	6.817	−.071	−6.618	10.486	−.015
Picture Vocabulary	813	2.112	5.219	.010	10.008	8.355	.026
Applied Problems	813	−5.296	10.518	−.018	1.027	15.694	.002
Quantitative Concepts	813	2.580	8.182	.013	22.318†	12.807	.058
Copy Design	813	−2.366	2.046	−.063	−6.837*	2.972	−.094
Corsi Forward Span	813	−0.017	0.697	−.001	−0.215	1.064	−.007
Corsi Backward Span	813	−1.125	0.824	−.063	−2.100	1.300	−.060
DCCS	813	−0.670	0.412	−.086	−0.261	0.627	−.017
HTKS	813	−12.159	8.960	−.053	14.649	14.472	.033
Peg Tapping	813	−2.087	3.311	−.027	7.916	5.263	.053
Interpersonal Skills	821	0.264	0.707	.019	−0.996	1.134	−.036
Work-Related Skills	821	1.014	0.888	.064	−0.023	1.448	−.001

Note. Coefficients are unstandardized regression coefficients and standard errors from multilevel regression models that account for nesting of teachers in schools and random assignment blocks. Covariates included in the models were pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest. Standardized effect sizes calculated indicated the change in a child's outcome in standard deviation units when the process variable increased by one standard deviation.
DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders; *SE* = standard error.
$^{*}p < .05$.
$^{\dagger}p < .10$.

Last, classrooms with higher rates of children's off-task behaviors made smaller gains in preliteracy skills. Specifically, classrooms with higher levels of children's off-task behavior made smaller gains on Letter-Word (*ES* = −.084), Spelling (*ES* = −.111), and Picture Vocabulary (*ES* = −.039). No associations were found between off-task behavior and the other measures of academic.

Prekindergarten Classroom Processes and Gains in Kindergarten

As we saw accelerated growth following prekindergarten at least to the end of kindergarten for academic skills, in exploratory analyses we examined the possible long-term effects of prekindergarten classroom processes on children's kindergarten outcomes. Summaries in Table 20 show that there was no clear pattern of the long-term effects of children's prekindergarten experiences on the academic, executive function, and social-emotional gains they made in kindergarten.

First, we summarize the associations between prekindergarten processes and academic outcomes. Letter-Word Identification, Oral Comprehension, and Picture Vocabulary were unrelated to any of the processes we measured in their prekindergarten classrooms. Spelling (*ES* = .089) and Academic

TABLE 19

Child Involvement, Social Learning, and Off-Task Effects on End of Prekindergarten Child Outcomes

		Level of Involvement			Social Learning Interactions			Off-Task Behavior		
Variable	*n*	*B*	*SE*	*Effect Size*	*B*	*SE*	*Effect Size*	*B*	*SE*	*Effect Size*
Letter-Word	813	13.301*	4.469	.117	61.773*	24.066	.121	−73.422*	36.559	−.084
Spelling	813	8.635	5.786	.067	29.002	30.726	.050	−110.494*	43.709	−.111
Academic Knowledge	813	6.634*	2.546	.078	13.588	13.695	.036	−25.519	22.213	−.039
Oral Comprehension	813	−1.169	2.091	−.014	−3.471	10.351	−.009	0.412	18.221	.001
Picture Vocabulary	813	1.058	1.654	.014	−4.665	8.761	−.014	−22.637†	13.471	−.039
Applied Problems	813	2.005	3.154	.019	16.152	15.431	.033	1.750	27.349	.002
Quantitative Concepts	813	2.721	2.549	.037	−3.720	13.802	−.011	−11.173	20.760	−.020
Copy Design	813	0.446	0.618	.032	6.206*	3.014	.099	−7.887	5.096	−.073
Corsi Forward Span	813	0.030	0.217	.005	1.753†	1.040	.069	−0.890	1.821	−.020
Corsi Backward Span	813	0.171	0.265	.026	2.426†	1.286	.081	0.623	2.215	.012
DCCS	813	−0.027	0.129	−.009	0.672	0.626	.052	−0.385	1.070	−.017
HTKS	813	2.105	2.901	.025	17.501	14.769	.046	−17.588	24.695	−.027
Peg Tapping	813	0.597	1.059	.021	7.838	5.514	.061	−1.916	8.665	−.009
Interpersonal Skills	821	0.356	0.221	.068	0.690	1.209	.029	0.152	1.760	.004
Work-Related Skills	821	0.455	0.280	.077	1.300	1.532	.049	0.145	2.226	.003

Note. Coefficients are unstandardized regression coefficients and standard errors from multilevel regression models that account for nesting of teachers in schools and random assignment blocks. Covariates included in the models were pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest. Standardized effect sizes calculated indicated the change in a child's outcome in standard deviation units when the process variable increased by one standard deviation.

DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders; *SE* = standard error.

*$p < .05$.

$^{\dagger}p < .10$.

TABLE 20

Classroom Behavior and Activity Effects on End-of-Kindergarten Child Outcomes

Variable	*n*	Teacher Talk	Child Talk	Emotional Climate	Level of Instruction	Time in Centers	Time in Transition	Level of Involvement	Social Learning Interactions	Off-Task Behaviors
Letter-Word	810	.024	.016	.036	−.005	.056	−.046	.032	.035	−.026
Spelling	810	.004	.089*	−.003	.033	.091*	−.044	.084*	.007	−.088*
Academic Knowledge	810	−.016	.054	.064†	.029	.075	.029	.023	.088*	−.031
Oral Comprehension	810	−.004	−.041	.047	.033	−.007	.045	−.022	−.019	.015
Picture Vocabulary	810	−.039	−.024	−.008	.051†	−.037	.037	−.016	−.004	.008
Applied Problems	810	.019	−.050†	.018	.011	−.068*	−.025	−.008	.016	.034
Quantitative Concepts	810	−.024	.011	.037	−.050	.091*	−.024	.012	.066	−.012
Copy Design	810	−.040	.041	.033	.054	.009	−.029	.050	.079†	−.073†
Corsi Forward Span	810	−.038	−.047	−.005	−.040	.098*	.008	−.058	.013	.021
Corsi Backward Span	810	−.009	−.026	.067†	.001	−.029	−.069†	−.003	.113*	−.050
DCCS	810	.024	−.041	.023	.021	−.024	−.002	−.033	.044	−.003
HTKS	810	.053	.061	−.007	.087†	−.159*	.022	−.108*	.087	.024
Peg Tapping	810	−.010	.040	.063	.020	.039	.020	.033	.079	−.049
Interpersonal Skills	811	.023	−.046	.035	−.036	−.065	−.033	.017	−.011	.045
Work-Related Skills	811	.014	−.030	−.005	−.067†	−.044	−.050	−.014	.017	.017

Note. For parsimony only effect sizes and indication of significance (from multilevel regression models that account for nesting of teachers in schools and random assignment blocks. Covariates included in the models were pretest, gender, home language (English or not), Individual Education Plan status (additional supports for children with learning difficulties), age at pretest, and interval from pretest are reported. Standardized effect sizes calculated indicated the change in a child's outcome in standard deviation units when the process variable increased by one standard deviation. Unstandardized regression coefficients and standard errors are available from the first author.

DCCS = Dimensional Change Card Sort; HTKS = Head-Toes-Knees-Shoulders.

*$p < .05$.

$^{\dagger}p < .10$.

Knowledge (*ES* = .064) gains in kindergarten suggested some continuing relations to the prekindergarten classroom experiences. The trend that prekindergarten time in centers was positively associated with both Spelling and Academic Knowledge maintained to a degree for children in kindergarten (*ES* = .091 and .075, respectively).

Similarly, the associations between children's outcomes in classrooms with higher levels of child involvement and more social learning interactions were also maintained into kindergarten. Specifically, the level of children's involvement in prekindergarten classrooms positively related to Spelling gains (*d* = .084) in kindergarten and off-task behaviors negatively related to gains (*d* = −.088). Children from prekindergarten classrooms with more social learning interactions had larger Academic Knowledge gains (*ES* = .088) in kindergarten.

The pattern of associations was inconsistent between the prekindergarten classroom processes and the two measures of mathematics. For example, prekindergarten time in centers negatively related to gain in kindergarten for Applied Problems (*ES* = −.068) and positively related to kindergarten gains in Quantitative Concepts (*ES* = .091).

Second, we examined the associations between those processes and executive function gains in kindergarten. The analysis suggested that the positive relations between prekindergarten emotional climate and executive function gains in some measures did persist in kindergarten for Corsi Backward Span (*ES* = .067). Similarly, prekindergarten social learning interactions positively related to kindergarten gains in Copy Design (*ES* = .079) and Corsi Backward Span (*ES* = .113). Prekindergarten level of instruction positively related to HTKS kindergarten gains (*ES* = .087). However, more time spent in centers in prekindergarten was negatively related to HTKS kindergarten gains (*ES* = −.159).

Lastly, exploration of the effects on teacher ratings of children's interpersonal and work-related skills did not yield significant associations between prekindergarten classroom processes and teacher ratings in kindergarten.

Discussion

Comparing the classroom processes and interactions of the *Tools* intervention to the control classrooms provides insight into the mechanisms that may underlie classroom effects. Working from the *Tools* developers' theory of change, we expected that specific elements would be different in *Tools* classrooms compared with control. However, we did not find extensive differences in interactions and practices between classrooms enacting *Tools* and typical early childhood classrooms (i.e., control classrooms). This finding may help explain the lack of curriculum effects on children's outcomes. In a secondary examination, we explored whether the behaviors identified were associated with children's growth in prekindergarten classrooms and beyond.

Differentiation Between Conditions

Presumably implementing the specific activities contained in any curriculum has purposes beyond just enacting the activities. Curriculum developers and trainers of *Tools* were both clear and concrete about what behaviors they thought would differentiate their classrooms from other typical early childhood classrooms. Having the developers and trainers for a curriculum predict how the implementation of their activities would affect specific classroom behaviors and interaction is unusual in intervention evaluations.

In some respects, their predictions were correct. Where *Tools* classrooms differed most from control classrooms involved behaviors that were most closely connected to the curriculum itself. Children were observed more often listening to other children; sharing the news with your neighbor and buddy reading are key activities in the curriculum. The effects of these activities and others are also reflected in the higher amounts of associative and cooperative interactions in *Tools* classrooms. Although the effects were small, children were more engaged during centers when they were allowed free play (not the make-believe center time) as well as during large group instruction. These differences attest to the implementation of the curriculum described in Chapter III and align with other randomized control trials of *Tools*.

Other interactions, however, did not differ between classrooms in the two conditions. The classrooms did not differ in the amount of teacher talk (with *Tools* teachers actually talking slightly more) and in the overall amount of child talk. The quality of the instruction observed was relatively low and did not distinguish *Tools* classrooms from control classrooms. Also contrary to predictions, there was more teacher-led instruction in *Tools* classrooms than in control classrooms, and teachers in control classrooms engaged in more approving behavior than teachers did in *Tools* classrooms.

One possible explanation for the overall lack of differentiation in the two sets of classrooms may lie with a recently identified issue involving changes in the counterfactual (Lemons et al., 2014). Experimental research often contrasts an intervention with business-as-usual, the counterfactual, or what children would have experienced without the intervention. What constitutes business-as-usual changes over the years and can affect results even for curricula previously established as effective. The recent large-scale evaluation of the *Building Blocks* prekindergarten math curriculum found no differences in math gains between treatment and control classrooms at the end of the prekindergarten year, primarily because of the increased math instruction taking place in the control classrooms (Morris et al., 2016).

The counterfactual may have been a factor in this evaluation. The early childhood classrooms were all prekindergarten programs connected to their school systems. The classrooms were well equipped and taught by licensed teachers with at least a bachelor's degree. All classrooms also each had an educational assistant. The emotional climate of the classrooms was marked by warmth from the teacher and relatively low rates of disapproval with about an even ratio between approval and disapproval. Children were not disruptive and not often unoccupied. The introduction of the curriculum actually brought some changes that might be of concern, more teacher talking, less approving, and more teacher-led instruction. In the original development of the curriculum, a large proportion of the *Tools* classrooms were Head Start classrooms. Much of the development of *Tools* took place in the 1990s prior to the great expansion of prekindergarten by school districts. Thus, the curriculum may not have been as appropriate for the current prekindergarten classrooms and/or the quality of the classrooms to which its effects were compared.

The circumstance of the changing counterfactual presents a dilemma for curriculum developers whose developmental work can take years prior to wide-scale dissemination. The difficulty then arises for developers who may need to reacquaint themselves with the classrooms they want to enhance with their interventions. Lemons et al. (2014) strongly urge curriculum developers to continue to familiarize themselves with the current educational practices to see how they and the developer's instructional program are alike and different. Developers must make sure that what they are offering will continue to be more facilitative for young children than the evolving alternative business-as-usual.

Classroom Processes Associated With Child Outcomes

Curricula are designed to help teachers enhance the quality of children's learning experiences in order to improve child outcomes. Although curricula vary in scope and content focus (e.g., global or developmental, skill-specific, or academically oriented), the expectation remains that implementation of a curriculum will change classroom processes that in turn will facilitate the development of targeted skills. Put another way, improved learning experiences or classroom processes are the mediators between a curriculum and its intended outcomes. For this mediational hypothesis to hold, it is essential that a curriculum create changes in targeted classroom processes and that those classroom processes are then associated with targeted child outcomes.

The *Tools of the Mind* developers had a theory of change that outlined specific hypotheses about how the curriculum would improve the quality of learning experiences for children. However, when evaluated, these differences did not for the most part emerge as predicted between classrooms implementing *Tools* and the control classrooms. Although this aspect of the mediational hypothesis of the curriculum was not supported, there was merit in testing if the processes that the curriculum was hypothesized to change were related to child outcomes.

The ability to identify classroom processes to target for intervention is limited by the ability to reliably define, observe, and measure the specific behaviors of teachers and children. A strength of the current study's approach to capturing classroom processes was the use of a coding scheme that primarily used behavioral counts. Even for those variables that required rating on a scale (teacher tone and children's level of involvement), the scale was anchored to observable behaviors and applied many times to individual children and teachers, not rated across a group of children. In addition, the processes examined captured both global aspects of the classroom experienced by all classroom members (e.g., time in centers and transitions) and a mixture of specific children's (e.g., associative and cooperative interactions, level of involvement in learning) and teachers' behaviors (e.g., talking to children, providing higher-quality instruction).

The behaviors we captured informed the *Tools* developers of areas where the curriculum was not performing as expected. In the winter of the evaluation year, when the first observational data were summarized, we scheduled a meeting with the developers and the trainers to share the interim results. Of particular concern to the researchers was that children appeared to be no more engaged in *Tools* classrooms than the children in control classrooms. Given that this was a randomized control trial and not a scale-up, the developers could have made changes in the coaching and training to address these issues. We urged a mid-year correction. Interestingly, perhaps because the trainers did not visit the control classrooms, they may not have been clear about what needed to change in *Tools* classrooms to make them more engaging. We did not find any differences in the practices observed in *Tools* classrooms following our meeting.

More precise descriptions of classroom interactions are also valuable for administrators and policymakers who can use these data to determine if educational settings fit their visions for high-quality early childhood education. In the United States, as we outlined in Chapter I, the emphasis is increasingly on how early childhood practices and curricula lead to school readiness in children. Although the classroom processes we examined were supported by extant research that looked at relations between processes and outcomes (e.g., Dickinson, 2011; Farran et al., 2017; Hamre et al., 2014; Justice et al., 2008; Keys et al., 2013; Nesbitt et al., 2015; Weiland et al., 2013), we did not find the robust and consistent impacts for classroom processes on child outcomes that we expected. In fact, in some cases, associations even varied across outcome measures that conceptually were capturing the same aspect of child development and knowledge. Despite the fact that our measures were reliable and specific, the effect sizes we reported are not much higher than the effect sizes reported from other recent research on prekindergarten classrooms that used more global and general measures as predictors (see Burchinal, 2018, for a review).

Nonetheless, some aspects of classroom functioning emerged as important to consider. Examination of relations between the prekindergarten classroom processes and children's gains suggested that the emotional climate of the classroom (less disapproving behavior, more approving behavior, positive teacher tone) and teacher's level of instruction were the aspects of the classrooms most strongly associated with children's outcomes. Emotional climate effects were largest for measures of literacy and language and a subset of the executive function measures. The level of instruction was associated with greater gains for HTKS and Peg Tapping. In addition, social learning interactions (associative and cooperative) were generally found to be positively related to children's outcomes with the most consistent impacts for measures of executive function skills. If this research can be corroborated by other investigations, it suggests that efforts to improve classroom processes through such systems as coaching could focus on a smaller number of interactive behaviors no matter what curriculum is also being implemented.

Future research should also consider that variability in behaviors that exist among individual children within a classroom as there can be wide variability in children's experiences within the same classroom (Bratsch-Hines et al., 2019). In fact, prior work has demonstrated that there can be more variability between individual children in a classroom than between classrooms themselves with regard to children's involvement levels, social-learning interactions, and unoccupied-disruptive behaviors, with 53%, 57%, and 84% of the variance occurring within classrooms, respectively (Nesbitt et al., 2015). The focus of this chapter was to understand the impacts of a curriculum on classroom processes. Analyses focused on the average experience of children in a given classroom. However, if progress is to be made in our ability to provide all children the most effective environments and experiences to learn and thrive, we must consider not only how quality at the classroom level is delivered but also how quality is received by individual children within a classroom. This latter will be a challenge to our measurement and analytic skills.

V. Improving Prekindergarten Quality With Curricula: Discussion and Conclusions

In this monograph we explored the effects of a comprehensive curriculum, *Tools of the Mind* (Bodrova & Leong, 2007) on first, children's academic, social, behavioral, and self-regulation outcomes, and second, on teacher behaviors and classroom processes. The *Tools* curriculum is focused on helping children develop executive function skills, especially those related to self-regulation. Just as importantly, it is focused on facilitating the acquisition of mathematics and literacy skills by having children become more responsible for learning and practicing literacy skills through make-believe and thematic play. Beginning in the 2000s, *Tools* received significant national attention, notably as a new type of early childhood approach.

In 2009, Paul Tough and colleagues published a popular piece in the *New York Times Magazine* in which they described both the increasingly pre-academic focus of early childhood education and the backlash that this focus was generating (Tough et al., 2009). In 2016, Weisberg et al. (2016) asserted that the competing trends in early childhood education—the push for both a strong curriculum focus and for children's need for free exploration—could be resolved by using specific types of curricula. They cited *Tools of the Mind* as one. Chambers et al. (2016) characterized *Tools* as comprehensive, by which they meant that it focused on specific skills but also made time for play and discovery. It certainly appeared to the field that *Tools* embodied the best of both of the overarching goals of early childhood education curriculum (Blair & Diamond, 2008).

The research project and the data we reported in this monograph were addressed to evaluating whether *Tools* does, indeed, deliver on the promises it was said to hold. In our research, we provided supports for effective implementation of the curriculum and documented the process and its impact. Most importantly, we worked in close partnership with the developers of *Tools* while we wrote the original proposal and then as we implemented the project after having been funded. All assessments, observation measures, and fidelity systems were chosen in collaboration with the developers and all were beta-tested with national *Tools* trainers before final decisions were made.

It is hard to imagine providing a new curriculum intervention with any more support than was provided to *Tools* teachers in this project. Ours was a randomized control trial, not a scale-up. It was appropriate for the curriculum developers to be closely involved in the evaluation. However, objectivity was fostered by having designed the project as a partnership with independent researchers. By having the evaluation conducted independently, we avoided the well-documented phenomenon of highly inflated effect sizes in developer-run or commissioned evaluations (Wolf et al., 2020).

The classrooms involved in the evaluation came from five school districts in two southern states. One of the *Tools* developers visited with each of the districts and explained the approach in detail before the school systems agreed to participate. All teachers had early childhood education certification and a bachelor's degree or more. Classrooms each had at least one educational assistant for classrooms that enrolled no more than 20 children. Because *Tools* is a complex curriculum, one that requires modifications in the ways the teachers normally taught, a full year of professional development was provided before teachers implemented the curriculum, and before we evaluated the effects. Coaches hired the first year received training from *Tools* staff and participated in all the workshops for the teachers. The coaches provided many hours of in-classroom consultation, offering even more support during the second, implementation year than they had during the first, preparation year.

Briefly, we assessed children's academic achievement in the areas of literacy, language, mathematics, and academic knowledge. We assessed children's executive function and self-regulation skills with a battery of individual measures chosen specifically to tap children's skills of inhibitory control, working memory, and cognitive flexibility. We measured children's social and emotional development through teacher ratings that focused on both learning-related classroom behaviors and social interactions with adults and peers. We were able to obtain pre- and posttest scores on almost all the children. There was very little attrition. We followed most of the sample into kindergarten and first grade. Ultimately, however, we found that almost none of these child measures at any time showed a positive effect of participating in the curriculum; a few outcomes even favored children in the control classrooms.

One possibility to consider is that null effects were due to the curriculum not being implemented well. We addressed this issue in Chapter III. Because the developers of *Tools* had not previously created their own measure of the fidelity of implementation, we worked with them to create such a measure in the first year of the project. As reported in Chapter III, the *Tools* developers did not already have a way to specify precisely what they considered to be high-quality implementation. Thus, we developed a fidelity tool based on two sources. First, we drew on the *Tools* curriculum manuals which outlined specific activities that should be incorporated into the curriculum as well as the progressive steps that should be followed when implementing those activities across the school year. At the time we began to design our fidelity measure, the curriculum provided 40 such activities, but by the time we were ready to collect data, the developers had added 21 more. We integrated these 21 activities into the final fidelity assessment. Second, we drew from the professional-development training materials used to introduce teachers to the *Tools* curriculum, and later to support them as they used it in their classrooms.

Other evaluators of early childhood curriculums with far fewer activities have reported dosage figures comparable to ours—about 60% implementation on average. Unsurprisingly, teachers varied in how much they implemented the curriculum, with none reaching the expectation that 50% of the school day would be spent in *Tools* activities (implementation averaged 29–33% across the three observations). The data we presented show that many *Tools* teachers followed the script even though they may not have achieved the levels of implementation outlined in the manuals. They carried out most of the activities with the steps outlined in the manuals, making changes in steps needed to implement activities as the school year progressed. There were few observed should-nots, that is, behaviors identified by the developers as behaviors to avoid.

Within the constraints of a 6-hr prekindergarten day that also included naps, meals, and outdoor time, most teachers made serious attempts to implement the curriculum. Teachers did not schedule make-believe play for as long as the curriculum manuals indicated, and they spent far more time transitioning their children between activities than the manuals suggested. We note that 3 of the 32 teachers were highly resistant and did not implement very much of the curriculum. Their lack of implementation could not, however, account for the weak impact of *Tools* because the children in these teachers' classrooms actually made strong gains across the year.

We measured the general classroom qualities the developers of *Tools* articulated in both treatment and control classrooms. Our findings are reported in Chapter IV. Classroom observations across the year assessed the classroom practices and interactions the developers thought would differentiate *Tools* classrooms from other early childhood settings. A few of the interactions more closely aligned with the curriculum occurred more often in *Tools* classrooms. For example, there was more evidence of children talking to other children in *Tools* than in control classrooms. However, the major processes such as the level of child involvement, the degree of positivity in the classroom, the ratio of child-to-teacher talk, and the quality of the instruction observed, were not different in tools versus control six skills across the prekindergarten year.

This extensive research project has thus failed to find evidence for the efficacy of the *Tools of the Mind* curriculum. We believe that the lessons learned from our research have implications beyond the validation of a single curriculum effort. They bear on many issues currently facing early childhood education, including the goal of finding ways to provide more effective early education for children from low-income families. In the remainder of this chapter, we offer four lessons learned, and end with brief concluding remarks about directions for future work.

Lessons Learned 1

Rigorous, Independent Evaluation of Curricula

> *To determine if a comprehensive early childhood curriculum with a focus on academic skills might be the strongest hope for improving the outcomes from prekindergarten programs (as suggested by* Yoshikawa et al., 2013*), the field of education must help programs identify or create strong curricula with the potential to create the desired outcomes.*

As we have noted, the United States is practically unique among developed countries in not having a national or unified vision for the important experiences and competencies needed by young children before formal schooling. As more effort is invested in prekindergarten programs, the emerging vision for these programs includes both the development of immediate school readiness skills and long-term effects on school achievement. Although there is considerable evidence that current programs have positive impacts on the former, there is less evidence for accomplishing the latter goal. Evaluations of the curricula currently in widest use—*Creative Curriculum* and *High Scope*—did not find long-term effects and attributed this negative result to the possibility of a heavy focus on discovery learning and too little direct teacher instruction (Jenkins et al., 2018; Nguyen et al., 2019). In response to the weak evidence of longer-term effects for prekindergarten programs and concerns about the widespread use of more global curricula, a chorus of researchers and policymakers in the past few years has proposed that a solution is to require scripted curricula that are explicitly focused on academic content (Jenkins et al., 2018; Phillips et al., 2017; Sharpe et al., 2017; Weiland et al., 2018; Yoshikawa et al., 2016).

If program directors and school administrators want to adopt a more intentional and academically focused curriculum for their prekindergarten programs, it would be difficult for them to find reliable evidence on which curricula are effective. There are two possible sources administrators could use for information. The first is the National Center on Early Childhood Development, Teaching, and Learning (NCQTL) which lists curricula that Head Start recommends for its programs (described in Chapter I). *Tools* is on that recommended list even though it is designated as having a minimal evidence base for child outcomes (https://eclkc.ohs.acf.hhs.gov/curriculum/consumer-report/curricula/tools-mind).

The second source is the What Works Clearinghouse (WWC) funded by the Institute of Education Sciences (https://ies.ed.gov/ncee/wwc/). The WWC provides reviews of research on practices and policies in education that allow administrators and teachers to make informed decisions on effective curricula and practices. The WWC has reviewed numerous early childhood curricula but it does so only if published evaluation results are available or if evaluators and/or developers submit their research for review. Many of the curricula being sold by publishers have not been evaluated in a manner that qualifies for WWC review. To qualify, the research must allow for a conclusion about a causal effect (i.e., evaluated using a randomized control trial or quasi-experimental design with baseline equivalence between conditions). Not since PCER (Preschool Curriculum Evaluation Research Consortium, 2008) has there been an intentional effort by the Institute of Education Sciences to evaluate and compare the effectiveness of early childhood curricula.

In terms of what would be available to an interested program leader or school district administrator, the WWC organizes the curricula on the basis of the strength of the evidence related to each. Of the 16 early childhood curricula and practices currently listed on the website as having some evidence of positive effects, 9 are supplemental sets of activities, not full curricula. Five are complete curricula covering more than one academic discipline of which four remain on the market (*Curiosity Corner*; *Ready, Set, Leap*; *Doors to Discovery*; and *Literacy Express*). The fifth curriculum with positive evidence, *Bright Beginnings,* developed by the Charlotte Mecklenburg, NC school system, underwent revisions and a name change. It became *Opening the World of Learning*.

Publishers are producing and replacing curricula faster than they can be evaluated. For example, *DLM Early Childhood Express* is listed on the WWC site and also appears on the NCQTL list of Head Start recommended curricula. However, at the McGraw Hill website, the listed publisher for *DLM Early Childhood Express*, the curriculum is not included on its list of those available to purchase for prekindergarten. Instead, an entirely new curriculum is listed for prekindergarten, *World of Wonders: Developing Early Learners.* The transformed *Bright Beginnings* curriculum, *Opening the World of Learning* (*OWL*), is on the NCQTL recommended list. Its publisher is supposed to be Pearson, but a school administrator would not find *OWL* listed as available for purchase from Pearson, or anywhere else.

In other words, much work is needed to identify curricula that could be purchased and implemented for which there is sufficient evidence of short- and long-term effects in the areas desired. It is difficult to know how to address this issue, but it is incumbent on early childhood researchers to figure it out. There is a gulf between researcher-developed curricula and publisher-developed curricula (often with assistance from academicians). Curriculum developers such as Leong and Bodrova spent years creating the *Tools* curriculum based on a specific theoretical perspective; they then had to develop the infrastructure to market the curriculum and to provide the trainers needed to teach teachers how to use it. *Curiosity Corner,* developed by Chambers and Slavin, is similar. They developed it and now market it through the Success for All Foundation, not through a standard publisher (http://www.successforall.org/our-approach/schoolwide-programs/curiosity-corner/). Curriculum developers are not in the best position to conduct objective evaluations of their work. Similarly, because publishers are marketing their products to school districts and programs, they cannot afford to take the time to conduct systematic research or, worse, to discover null or negative results despite having received input from knowledgeable experts in the field.

The lack of empirical support for curricula is not unique to early childhood education. The same issues affect the K-12 system. Notably, in elementary mathematics, Bhatt and Koedel (2012) report that in 2011 the What Works Clearinghouse listed over 70 different elementary mathematics curricula, but as Bhatt and Koedel note: "there are few rigorous, empirical evaluations of curricular effectiveness; the research literature is surprisingly thin" (p. 391). Also, in the area of teaching mathematics at the middle school level, Jackson and Makarin (2016) found that teachers are increasingly relying on activities they can draw down from the Internet. The site Teachers Pay Teachers had an active membership of 4 million users in 2016; Jackson and Makarin point out that as of 2016 there were only 3.5 million primary and secondary teachers in all of the United States.

In other words, the education field as a whole has very little to guide it in terms of relying on curricula, and many teachers are seeking other closer-to-home sources for what they consider valid instructional activities. Before we assert the importance of scripted, intentional curricula for improving early childhood outcomes, we must be responsible for doing rigorous, relevant research and in being more inventive investigating learning activities teachers will actually use and that will matter for children.

Lesson Learned 2

Teacher Ownership of Curricula

> *Providing teachers a voice, and therefore ownership, of the curriculum a program or district is adopting, may be the most important way of ensuring the development of an effective program.*

Teacher ownership often results in curriculum adaptation. Three examples of teacher adaptation of a curriculum are illuminating. The first involved the group of developers for the peer-mediated instructional program PALS (Peer-Assisted Learning Strategies) and the adaptations developers allowed teachers to make to their approach (Kim, 2019; McMaster et al., 2014). An effort to scale-up PALS beyond the original developers' control had not yielded the expected positive outcomes for children. In response, the developers identified the core components of PALS they thought to be critical to the curriculum. Beyond implementing those few core components, they told the teachers that they were free to do one of two things. They could either implement the entire PALS approach as they had been taught, or they could use the core components as the basis for adapting and changing the curriculum to fit their individual circumstances. The developers called the latter "customized PALS." The children of those teachers who customized

PALS significantly outperformed the children whose teachers used PALS in a by-the-book manner (Lemons et al., 2014).

A different approach involved published curricula that teachers then overhauled and adapted completely. The Boston Public School System took this approach with its prekindergarten program. An article in *The Atlantic* (Mongeau, 2016) extensively detailed the development of the curriculum the program uses. There are three important components to the Boston approach. First, the form of the curriculum took several years to evolve with extensive teacher input all along the way. Second, coaching is ongoing, provided by teachers who had multiple years of experience with the curriculum before moving into the coaching position. And third, the curriculum is dynamic; teachers must stick to the district-wide curriculum, but they also have the freedom to work with the coaches and to make adaptations to fit their individual classrooms. Weiland et al. (2018) describe this procedure as giving the teachers an ongoing voice in what is chosen and how it is to be implemented.

Another notable example of a successful partnership between researchers and practitioners is the Kamehameha Early Education Program (KEEP) in Hawaii (Tharp & Gallimore, 1982). KEEP operated for years and is still recognized as one of the few genuine collaborations that produced important changes for children (Jacobs et al., 2012). Tharp and Gallimore (1982) argued that in constructing a program, the goal must be viability, involving immersion in the actual environments where the program must ultimately survive. KEEP collaborators made use of laboratory classrooms, places where the partners could tinker, where they had the opportunity to try things, fail, persist, self-correct, and gradually improve. They evaluated components of the program as they developed them but did not do a full-scale evaluation until the program elements had stabilized. This process took them 5 years before any success emerged. Sadly, most school systems and programs do not feel they have the time for this kind of iterative process.

The KEEP approach is a clear example of the community-based participatory action research many seek (James et al., 2008). This is research that involves a full range of stakeholders (including school representatives and researchers) who contribute their expertise and experience as collaborative and respected partners, and share equitably in decision-making and ownership. The KEEP model illustrates the iterative process of action, reflection, and experiential learning.

The McREL Institute issued a comprehensive report to guide states in the creation of their public prekindergarten programs (Best & Cohen, 2013) with premises similar to those underlying the Boston and KEEP process of curriculum development. Its first and most prominent suggestion was that programs should collaborate with experts from different disciplines and develop strategies driven by research and data, but also ones that fit their specific circumstances. Larman and Basili (2003) describe this process as one developed by Bell Labs in the 1930s for quality improvement. They contrast this model to the single-pass sequential cycle where a problem is outlined, an intervention is developed, implemented, and then evaluated. The latter system is the one used in most randomized control trials evaluating curricula, including the PCER project (Preschool Curriculum Evaluation Research Consortium, 2008). Someone external to the school system chose the curricula evaluated in that project with no input from the teachers and no procedures for developing ownership among them. None of the curricula assessed proved to be stronger than the processes teachers were enacting in the comparison classrooms, but once the evaluation project was over no one revisited the curricula to see how they could be improved.

In part, because they were participating in an experimental evaluation of the curriculum, the *Tools of the Mind* teachers in this study were not afforded the opportunity to alter the curriculum to fit their circumstances, although all teachers devoted at least some portion of the day to non-*Tools* activities. If developers had asked, teachers could have told them the difficulties of carrying out the *Tools* curriculum. As we have shown, the expectations for the number and complexity of activities for *Tools* teachers to implement in a day and across a week were not well calibrated to the length of the prekindergarten school day. If indeed teachers need to implement a structured curriculum, developers of such approaches must at a minimum be connected to the ecology of the classrooms where the curriculum will be implemented (Doyle & Ponder, 1977), or even better, work closely with teachers as valued partners.

On the other hand, *Tools of the Mind* may not be a curriculum that can be sold and adopted like any other. The inclusion of teachers' voices will then actually be more important for obtaining teacher buy-in than with any other early childhood approach, except possibly Montessori. Montessori and *Tools* are similar in that both involve a complete mindset change for teachers as well as a deep understanding of the purpose of a great many activities that most teachers have never seen. The Montessori approach is preserved by isolating and protecting the certification of teachers and being careful to validate programs claiming to enact Montessori (Lillard, 2007). Perhaps similarly, only teachers who wish to buy into the *Tools* philosophy should learn the approach and carry it out in their classrooms. In other words, teachers should have the power to choose the approach on their own.

Lesson Learned 3

Creating Better Research Designs

> *Although randomized control trials have been an important addition to educational research, we need additional, more iterative designs to help us understand which kinds of approaches in classrooms are likely to be most effective.*

When curricula are evaluated rigorously, the research almost always involves randomly assigning teachers and then assessing classrooms where teachers are engaged in new practices, sometimes ones quite different from the ways they usually teach (as was the case with *Tools*). Even with intense coaching and professional development, it is almost impossible for teachers to implement a curriculum exactly as prescribed.

There are consequences of this problem. One is that we may not learn how effective a curriculum could be if teachers had become more familiar with it, moving into a stage of true implementation instead of the learning phase. As Doyle and Ponder (1977) asserted long ago, teachers for the most part do not *adopt* a new curriculum or approach; rather they *adapt* the new innovative practices. Moreover, the tendency to adapt the new practices and alter them is more often evident in teachers who have higher levels of education and experience (Baker et al., 2010). In our study, the most experienced teachers were the least likely to change their current practices and implement *Tools*. A randomized control trial is often implemented for only a limited amount of time, and the conclusions drawn may thus also be limited. Knowing how teachers adapt a curriculum can be as informative as trying to force them to carry out the program exactly as designed.

A second consequence with randomized control trials is that teachers may view the demands to implement the new approach as temporary. Thus, even if the curriculum appears to be effective, teachers may not continue to carry it out once the study is over. Longer-term sustainability is rarely assessed in a curriculum evaluation. Even within the period of training on a new curriculum, Baker et al. (2010) found that teachers gradually decreased the implementation of the activities across a given week, something attributed to teachers running out of time to do all that was expected of them by the curriculum. From one year to the next, however, studies have found that teachers almost immediately stop implementing the new approach altogether as soon as training and support for it ceases (e.g., Lieber et al., 2010).

An alternative design could involve randomly assigning children rather than teachers. In this scenario, teachers could be offered a choice of two curricula that a school district is thinking of adopting. Teachers could be free to choose the one that fits their teaching practices better. Within a single school, children could be randomly assigned across conditions. This assures some buy-in on the part of the teachers but also preserves the value of randomization for assessing effects. Granted that this approach will work best in districts that are large enough to have several classrooms at the same grade level in the same school where randomization can take place or in districts where the prekindergarten program is in a separate building. Even with those limitations, we could learn much from this design.

Lessons Learned 4

Classroom Practices Beyond Curriculum

> *Independent of an early childhood curriculum are important classroom practices and interactions that should characterize early childhood classrooms and be the basis for coaching no matter what curriculum approach is adopted.*

The final issue confronting early childhood education involves improving classroom practices that may be orthogonal to the curriculum adopted. As Jenkins et al. (2019) demonstrated, teachers who are supposed to be implementing the same curriculum look very different in their practices, often as much variation among teachers who are supposed to be implementing the identical curriculum as among teachers who are implementing different curricula. Some of the variability may be due to differences in implementation, but much of it may be due to differences in teachers' interaction patterns. We found large variations in teacher practices and interactions within the classrooms in this study. Those patterns appeared to be independent of the specific curriculum being used. General interaction patterns tend not to be included in curriculum manuals whose focus is on academic content and specific activities—that is, whose focus is on what to teach rather than on how to teach it. As others assert, intentional curricular practices are "inseparable from issues of climate and child engagement" (Graue et al., 2004, p. 6).

The data we presented identified important classroom practices associated with children's growth and development, some with associations that were still evident when children reached kindergarten. The three important aspects of classroom interactions linked to children's developmental outcomes were (1) the emotional climate of the classroom including teacher warmth, less disapproving behavior, and more approving behavior, (2) how much children are given opportunities to engage in social-learning interactions, and (3) the level of children's involvement in classroom activities. In prior studies, as well, the quality of teachers' instruction has also been related to children's outcomes and this finding suggests a fourth crucial dimension (e.g., Durden & Dangel, 2008; Farran et al., 2017).

The classroom interactions we found are not unique to this study or our observational measures. The kinds of interactions we assessed also form the basis for the *CLASS* measure (Pianta et al., 2008), an observational rating system with the same goal of improving the general classroom atmosphere and teacher–child interactions. In addition, Hirsh-Pasek and Golinkoff focus their recommendations for quality classrooms on more positive teacher–child interactions in the context of a learning environment (Hirsh-Pasek et al., 2009; Weisberg et al., 2013, 2016).

Little research has been done, however, on how to combine these more general classroom processes with various curricula or types of curricula (i.e., global or skill-specific). It is possible to conceive of a curriculum that addresses these critical aspects of classroom functioning as part of its approach. But these aspects are almost never included in a scripted and intentional pedagogy. The focus of such curricula tends to be on teachers implementing the activities as prescribed, and there is relatively little if any attention to the quality of teachers' interactions with students or on how children respond to experiences and engage with the learning materials. Nevertheless, there is no reason to suppose that curriculum materials, training, or coaching need be silent about these goals.

We do not currently have strong evidence that these classroom practices and interactions can be improved in a sustained way. The early childhood field needs a system that provides coaches, principals, and teachers a specific and actionable way to capture these dynamic aspects of classroom functioning and that offers clear guidance and training to teachers in ways that are effective for implementing best practices along these dimensions.

Conclusions

This monograph focused on a single intentional, scripted curriculum of the sort advocated in the United States by those hoping to improve long-term outcomes of children who participate in preschool and prekindergarten programs, especially children from low-income families. *Tools of the Mind* is a comprehensive, full-day curriculum that is highly appealing to those who want to incorporate the development of executive function and self-regulation skills into a program that also focuses on more traditional academic, school-readiness competencies. Even though our results were not encouraging about the benefits of using *Tools,* we believe that the issues we have identified should be useful to the field of early childhood education. As school administrators are increasingly encouraged to adopt intentional, scripted, and academically focused early childhood curricula, they need to be provided with better information on the likely effectiveness of the approach chosen. In addition, we offer reassurance to teachers that other long-standing and supportive early childhood practices need not be abandoned as they focus more on academic knowledge and skills. Involving teachers as partners in the quest to find ways to combine the two should be the focus of the next generation of research on curriculum development and early childhood education.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090533 awarded to Dale C. Farran and Mark W. Lipsey. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Special thanks to Sandra Jo Wilson, Jennifer Norvell, Mary Wagner Fuhs, Deanna Meador, and Karen Anthony for their assistance with data analysis and the Peabody Research Institute research team for their very capable management of the data collection for this project. The greatest appreciation is also extended to the administrators, teachers, families, and children who opened their schools and classrooms to us. Lastly, we would like to convey our gratitude to Elena Bodrova and Deborah Leong, the developers of the *Tools of the Mind* curriculum, for their partnership, openness, and willingness to engage with us throughout the project.

References

Abreu-Lima, I. M., Leal, T. B., Cadima, J., & Gamelas, A. M. (2013). Predicting child outcomes from preschool quality in Portugal. *European Journal of Psychology of Education*, *28*, 399–420. https://doi.org/10.1007/s10212-012-0120-y

Alcock, S., & Haggerty, M. (2013). Recent policy developments and the "schoolfication" of childhood care and education in Aotearoa New Zealand. *Early Childhood Folio*, *17*, 21–26.

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, *73*(1), 81–94. https://doi.org/10.1037/amp0000146

Baker, C., Kupersmidt, J., Voegler-Lee, M., Arnold, D., & Willoughby, M. (2010). Predicting teacher participation in a classroom-based, integrated preventive intervention for preschoolers. *Early Childhood Research Quarterly*, *25*, 270–283. https://doi.org/10.1016/j.ecresq.2009.09.005

Barnett, W. S., Kwanghee, J., 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*, 299–313. https://doi.org/10.1016/j.ecresq.2008.03.001

Behne, T., Carpenter, M., Gräfenhain, M., Liebal, K., Liszkowski, U., Moll, H., Rakoczy, H., Tomasello, M., Warneken, F., & Wyman, E. (2008). Cultural learning and cultural creation. In N. Budwig, B. Sokos, U. Muller, & J. Carpendale (Eds.), *Social life and social knowledge: Toward a process account of development* (pp. 65–101). Taylor & Francis Group/Lawrence Erlbaum Associates.

Bennett, J. (2005). Curriculum issues in national policy-making. *European Early Childhood Education Research Journal*, *13*(2), 5–23. https://doi.org/10.1080/13502930585209641

Bernstein, B. (1996). *Pedagogy, symbolic control and identity: Theory, research, critique*. Taylor & Francis.

Best, J., & Cohen, C. (2013). *Early care and education: Policy considerations for ensuring high-quality pre-k programs*. Mid-Continent Research for Education and Learning.

Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: The case of elementary mathematics. *Educational Evaluation and Policy Analysis*, *34*, 391–412. https://doi.org/10.3102/0162373712440040

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., Blair, C., Nelson, K. E., & Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI Program. *Child Development*, *79*, 1802–1817. https://doi.org/10.1111/j.1467-8624.2008.01227.x

Bilbrey, C., Vorhaus, E., & Farran, D. C. (2011). *Teacher Observation in Preschools (With adaptations for use in Tools of the Mind curriculum evaluation)*. Vanderbilt University, Peabody Research Institute. https://my.vanderbilt.edu/toolsofthemindevaluation/resources/classroomobservationmeasures/

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, *20*, 899–911. https://doi.org/10.1017/S0954579408000436

Blair, C., McKinnon, R., & Daneri, P. (2018). Effects of the tools of the mind kindergarten program on children's social and emotional development. *Early Childhood Research Quarterly*, *43*, 52–61. https://doi.org/10.1016/j.ecresq.2018.01.002

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*, 647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Bodrova, E., & Leong, D. (2015). Vygotskian and post-Vygotskian views on children's play. *American Journal of Play*, *7*, 371–388.

Bodrova, E., & Leong, D. (2017). Play and early literacy: A Vygotskian approach. In K. Roskos (Ed.), *Play and literacy in early childhood: Research from multiple perspectives* (pp. 185–200). Routledge.

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

Bradbury, B., Waldfogel, J., & Washbrook, E. (2018). Income-related gaps in early child cognitive development: Why are they larger in the United States than in the United Kingdom, Australia, and Canada? *Demography*, *56*, 367–390. https://doi.org/10.1007/s13524-018-0738-8

Bratsch-Hines, M., Burchinal, M., Peisner-Fineberg, & Franco. (2019). Frequency of instructional practices in rural prekindergarten classrooms and associations with child language and literacy skills. *Early Childhood Research Quarterly*, *47*, 74–88. https://doi.org/10.1016/j.ecresq.2018.10.001

Brodin, J., & Renblad, K. (2014). Reflections on the revised national curriculum for preschool in Sweden—Interviews with the heads. *Early Child Development and Care*, *184*(2), 306–321. https://doi.org/10.1080/03004430.2013.788500

Bull, R., Espy, K. A., Wiebe, S. A., Sheffield, T. D., & Nelson, J. M. (2011). Using confirmatory factor analysis to understand executive control in preschool children: Sources of variation in emergent mathematic achievement. *Developmental Science*, *14*, 679–692. https://doi.org/10.1111/j.1467-7687.2010.01012.x

Burchinal, M. (2018). Measuring early care and education quality. *Child Development Perspectives*, *12*, 3–9. https://doi.org/10.1111/cdep.12260

Burchinal, M., Zaslow, M., & Tarullo, L. (2016). Quality thresholds, features and dosage in early care and education: Secondary data analyses of child outcomes. *Monographs of the Society for Research in Child Development, 81*, 1–128. https://doi.org/10.1111/mono.12236

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, *28*, 595–616. https://doi.org/10.1207/s15326942dn2802_3

Caspi, A., Wright, B., Moffitt, T. E., & Silva, P. A. (1998). Early failure in the labor market: Childhood and adolescent predictors of unemployment in the transition to adulthood. *American Sociological Review*, *63*, 424–451. https://doi.org/10.2307/2657557

Century, J., & Cassata, A. (2016). Implementation research: Finding common ground on what, how, why, where, and who. *Review of Research in Education*, *40*, 169–215. https://doi.org/10.3102/0091732X16665332

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*, 199–218. https://doi.org/10.1177/1098214010366173

Chambers, B., Cheung, A., & Slavin, R. (2016). Literacy and language outcomes of comprehensive and developmental-constructivist approaches to early childhood education: A systematic review. *Educational Research Review*, *18*, 88–111. https://doi.org/10.1016/j.edurev.2016.03.003

Chmielewski, A., & Reardon, S. (2016). Patterns of cross-national variation in the association between income and academic achievement. *AERA Open*, *2*, 1–27. https://doi.org/10.1177/2332858416649593

Christopher, C., & Farran, D. C. (2020). Academic gains in kindergarten related to eight classroom practices. *Early Childhood Research Quarterly*, *53*, 638–649. https://doi.org/10.1016/j.ecresq.2020.07.001

Clements, D. H., Sarama, J., Layzer, C., Unlu, F., & Fesler, L. (2020). Effects on mathematics and executive function of a mathematics and play intervention versus mathematics alone. *Journal for Research in Mathematics Education*, *51*, 301–333. https://doi.org/10.5951/jresemtheduc-2019-0069

Cochran, M. (2011). International perspectives on early childhood education. *Educational Policy*, *25*, 65–91. https://doi.org/10.1177/0895904810387789

Committee for Children. (2011). *Second Step Learning Program*. Committee for Children.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomized controlled trials in education research 1980–2016. *Educational Research*, *60*, 276–291. https://doi.org/10.1080/00131881.2018.1493353

Cooper, D. H., & Farran, D. C. (1988). Behavioral risk factors in kindergarten. *Early Childhood Research Quarterly*, *3*, 1–19. https://doi.org/10.1016/0885-2006(88)90026-9

Cooper, D. H., & Farran, D. C. (1991). *The Cooper-Farran Behavioral Rating Scales*. Clinical Psychology Publishing.

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain* (Doctoral dissertation). McGill University. http://digitool.Library.McGill.CA:80/R/-?func=dbin-jumpfull&object_id=93903&silo_library=GEN01

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

Diamond, A., Lee, C., Senften, P., Lam, A., & Abbott, D. (2019). Randomized control trial of *Tools of the Mind*: Marked benefits to kindergarten children and their teachers. *PLOS One*, *14*(9), e0222447. https://doi.org/10.1371/journal.pone.0222447

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science*, *333*, 959–964. https://doi.org/10.1126/science.1204529

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". *Developmental Psychobiology*, *29*, 315–334. https://doi.org/10.1002/(sici)1098-2302(199605)29:4%3C315::aid-dev2%3E3.0.co;2-t

Dickinson, D. K. (2011). Teachers' language practices and academic outcomes of preschool children. *Science*, *333*(6045), 964–967. https://doi.org/10.1126/science.1204526

Dodge, D. T., Colker, L. J., Heroman, C., & Bickart, T. S. (2002). *The creative curriculum for preschool*. Teaching Strategies, Inc.

Domitrovich, C. E., Greenberg, M. T., Cortes, R., & Kusche, C. (1999). *Manual for the preschool PATHS curriculum*. The Pennsylvania State University.

Doyle, W., & Ponder, G. (1977). The practicality ethic in teacher decision making. *Interchange*, *8*, 1–12.

Duncan, G., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*, 109–132. https://doi.org/10.1257/jeb27.2.109

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

Durden, T., & Dangel, J. (2008). Teacher-involved conversations with young children during small group activity. *Early Years*, *28*, 251–266. https://doi.org/10.1080/09575140802393793

Einarsdottir, J., & Wagner, J. T. (2006). *Nordic childhoods and early education: Philosophy, research, policy and practice in Denmark, Finland, Iceland, Norway, and Sweden*. Information Age Publishing.

Fantuzzo, J., Gadsden, V., & McDermott, P. (2011). Integrated curriculum to improve mathematics, language, and literacy for Head Start children. *American Educational Research Journal*, *48*, 763–793. https://doi.org/10.3102/0002831210385446

Farran, D. C. (2011). *Child observation in preschool (With adaptations for use in the Tools of the Mind curriculum evaluation)*. Peabody Research Institute, Vanderbilt University. https://my.vanderbilt.edu/toolsofthemindevaluation/resources/classroomobservationmeasures/

Farran, D. C., Bilbrey, C., & Vorhaus, E. (2010). *Procedural handbook for the Narrative Record of early childhood classroom observations. With adaptations for use in the Tools of the Mind curriculum evaluation*. Peabody Research Institute, Vanderbilt University. https://my.vanderbilt.edu/toolsofthemindevaluation/resources/classroomobservationmeasures/

Farran, D. C., Meador, D., Christopher, C., Nesbitt, K. T., & Bilbrey, L. E. (2017). Data-driven improvement in prekindergarten classrooms: Report From a partnership in an urban district. *Child Development*, *88*(5), 1466–1479. https://doi.org/10.1111/cdev.12906

Farran, D. C., & Nesbitt, K. T. (2019). New information on evaluating the quality of early childhood education programs. In O. Saracho & B. Spodek (Eds.), *Handbook of research on the education of young children* (4th ed., pp. 333–347). Routledge/Taylor & Francis.

Fonsén, E., Varpanen, J., Strehmel, P., Kawakita, M., Inoue, C., Marchant, S., Modise, M., Szecsi, T., & Halpern, C. (2019). International review of ECE leadership research—Finland, Germany, Japan, Singapore, South Africa, and the United States. In P. Strehmel, J. Heikka, E. Hujala, J. Rodd, & M. Waniganayake (Eds.), *Leadership in early education in times of change: Research from five continents* (pp. 253–276). Verlag Barbara Budrich. https://doi.org/10.2307/j.ctvmd84fc.21

Friedman-Krauss, A. H., Barnett, W. S., Weisenfeld, G. G., Kasmin, R., DiCrecchio, N., & Horowitz, M. (2018). *The state of preschool 2017: State preschool yearbook*. The National Institute for Early Education Research, Rutgers University. http://nieer.org/state-preschool-yearbooks/yearbook2017

Fuhs, M., Nesbitt, K., Farran, D., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology*, *50*, 1698–1709. https://doi.org/10.1037/a0036633

Fuhs, M. W., Farran, D. C., & Nesbitt, K. T. (2013). Preschool classroom processes as predictors of children's cognitive self-regulation skills development. *School Psychology Quarterly*, *28*, 347–359. https://doi.org/10.1037/spq0000031

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*, 31–60. https://doi.org/10.1037/0033-2909.134.1.31

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.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology*, *41*, 872–884. https://doi.org/10.1037/0012-1649.41.6.872

Graue, E., Clements, M., Reynolds, A., & Niles, M. (2004). More than teacher directed or child initiated: Preschool curriculum type, parent involvement, and children's outcomes in the child-parent centers. *Education Policy Analysis Archives*, *12*, 1–36.

Halpern, R. (2013). Tying early childhood education more closely to schooling: Promise, perils and practical problems. *Teachers College Record*, *115*, 1–28.

Hammer, C. S., Blair, C., Lopez, L., Leong, D., & Bedrova, E. (2012, March). *Tools of the Mind: Promoting the school readiness of ELLs*. Paper presented at the spring meeting for the Society for Research on Educational Effectiveness, Washington, DC.

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*, 1257–1274. https://doi.org/10.1111/cdev.12184

Harms, T., Clifford, R., & Cryer, D. (1998). *Early Childhood Environmental Rating Scale-Revised*. Teacher's College Press.

Hill, H., & Erickson, A. (2019). Using implementation fidelity to aid in interpreting program impacts: A brief review. *Educational Researcher*, *48*, 590–598. https://doi.org/10.3102/0013189X19891436

Hirsh-Pasek, K., & Golinkoff, R. (2003). *Einstein never used flash cards*. Rodale Press.

Hirsh-Pasek, K., Golinkoff, R., Berk, L., & Singer, D. (2009). *A mandate for playful learning in preschool: Applying the scientific evidence*. Oxford University Press.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*, 275. https://doi.org/10.1016/j.ecresq.2007.05.002

Howse, R. B., Lange, G., Farran, D. C., & Boyles, C. D. (2003). Motivation and self-regulation as predictors of achievement in economically disadvantaged young children. *Journal of Experimental Education*, *71*, 151–174. https://doi.org/10.1080/00220970309602061

Hsueh, J., Lowenstein, A. E., Morris, P., Mattera, S. K., & Bangser, M. (2014). *Impacts of social-emotional curricula on three-year-olds: Exploratory findings from the Head Start CARES demonstration* (OPRE Report 2014-78). U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/exploratory-impacts-of-three-social-emotional-curricula-on-three-year-olds-in-the-head-start-cares-demonstration

Hughes, C. (2011). Changes and challenges in 20 years of research into the development of executive functions. *Infant and Child Development*, *20*, 251–271. https://doi.org/10.1002/icd.736

Institute of Education Sciences. (2019, February). *The condition of education: Preschool and kindergarten enrollment*. https://nces.ed.gov/programs/coe/indicator_cfa.asp

Jackson, C. K., & Makarin, A. (2016). *Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment* (NBER Working Paper No. 22398). https://www.nber.org/system/files/working_papers/w22398/w22398.pdf

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

Jacobs, J., Sisco, M., Hill, D., Malter, F., & Figueredo, A. (2012). Evaluating theory-based evaluation: Information, norms, and adherence. *Evaluation and Program Planning*, *35*, 354–369. https://doi.org/10.1016/j.evalprogplan.2011.12.002

James, E. A., Milenkiewicz, M. T., & Bucknam, A. (2008). *Participatory action research for educational leadership: Using data-driven decision making to improve schools*. Sage Publications.

Jenkins, J., Duncan, G., Auger, A., Bitler, M., Domina, T., & Burchinal, M. (2018). Boosting school readiness: should preschool teachers target skills or the whole child? *Economics of Education Review*, *65*, 107–125. https://doi.org/10.1016/j.econedurev.2018.05.001

Jenkins, J., Whitaker, A., Nguyen, T., & Yu, W. (2019). Distinctions without a difference? Preschool curricula and children's development. *Journal of Research on Educational Effectiveness*, *12*, 514–549. https://doi.org/10.1080/19345747.2019.1631420

Jones, S. M., Bailey, R., Barnes, S. P., & Partee, A. (2016). *Executive function mapping project: Untangling the terms and skills related to executive function and self-regulation in early childhood* (OPRE Report # 2016-88). U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/untangling-the-terms-and-skills-related-to-executive-function-and-self-regulation-in-early-childhood

Justice, L., Mashburn, A., Hamre, B., & Pianta, R. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly*, *23*, 51–68. https://doi.org/10.1016/j.ecresq.2007.09.004

Kamerman, S., & Gatenio-Gabel, S. (2007). Early childhood education and care in the United States: An overview of the current policy picture. *International Journal of Child Care and Education Policy*, *1*(1), 23–34. https://doi.org/10.1007/2288-6729-1-1-23

Kern, M., & Friedman, H. (2009). Early educational milestones as predictors of lifelong academic achievement, midlife adjustment, and longevity. *Journal of Applied Developmental Psychology*, *30*, 419–430. https://doi.org/10.1016/j.appdev.2008.12.025

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., Ruzek, E. A., & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*, 1171–1190. https://doi.org/10.1111/cdev.12048

Kim, J. (2019). Making every study count: Learning from replication failure to improve intervention research. *Educational Researcher*, *48*, 599–607. https://doi.org/10.3102/0013189X19891428

Ladd, H. (2012). Presidential address: Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, *31*, 203–227. https://doi.org/10.1002/pam.21615

Larman, C., & Basili, V. (2003). Iterative and incremental development: A brief history. *Computer*, *36*, 47–56. https://doi.org/10.1109/MC.2003.1204375

Lemons, C., Fuchs, D., Gilbert, J., & Fuchs, L. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*, 242–252. https://doi.org/10.3102/0013189X14539189

Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E., & Rolla, A. (2015). Teacher–child interactions in Chile and their associations with prekindergarten outcomes. *Child Development*, *86*, 781–799. https://doi.org/10.1111/cdev.12342

Lieber, J., Hanson, M., Butera, G., Palmer, S., Horn, E., & Craja, C. (2010). Do preschool teachers sustain their use of a new curriculum? *NHSA Dialog*, *13*, 248–253. https://doi.org/10.1080/15240754.2010.513778

Lillard, A. (2007). *Montessori: The science behind the genius*. Oxford University Press.

Lipsey, M., Farran, D., & Durkin, K. (2018). Effects of the Tennessee prekindergarten program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly*, *45*, 155–176. https://doi.org/10.1016/j.ecresq.2018.03.005

Lipsey, M. W., Nesbitt, K. T., Farran, D. F., Dong, N., Fuhs, M. W., & Wilson, S. J. (2017). Cognitive self-regulation measures for prekindergarten children that perform well for predicting academic achievement: A comparative evaluation. *Journal of Educational Psychology*, *109*, 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 presented at the spring meeting for the Society for Research on Educational Effectiveness, Washington, DC.

Mackey, A., Finn, A., Leonard, J., Jacoby-Senghor, D., West, M., Gabrieli, C., & Gabrieli, J. (2015). Neuroanatomical correlates of the income-achievement gap. *Psychological Science*, *26*, 1–9. https://doi.org/10.1177/0956797615572233

McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental Psychology*, *43*, 947–959. https://doi.org/10.1037/0012-1649.43.4.947

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*, 1–14. https://doi.org/10.3389/fpsyg.2014.00599

McClelland, M. M., & Morrison, F. J. (2003). The emergence of learning-related social skills in preschool children. *Early Childhood Research Quarterly*, *18*, 206–224. https://doi.org/10.1016/S0885-2006(03)00026-7

McClelland, M. M., Tominey, S. L., Schmitt, S. A., Hatfield, B., Purpura, D., Gonzales, C., & Tracy, A. (2019). Red Light, Purple Light! Results of an intervention to promote school readiness for children from low-income backgrounds. *Frontiers in Psychology*, *10*, 2365. https://doi.org/10.3389/fpsyg.2019.02365

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., & Barmer, A. (2019). *The Condition of Education 2019* (NCES 2019-144). U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/pubs2019/2019144.pdf

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037//1082-989x.1.1.30

McLachlan, C., Fleer, M., & Edwards, S. (2010). *Early childhood curriculum: Planning, assessment and implementation*. Cambridge University Press. www.cambridge.org/9780521759113

McMaster, K. L., Jung, P. G., Brandes, D., Pinto, V., Fuchs, D., Kearns, D., Lemons, C., Saenz, L., & Yen, L. (2014). Customizing a research-based reading practice. *Reading Teacher*, *68*, 173–183. https://doi.org/10.1002/trtr.1301

Meador, D., Nesbitt, K., & Farran, D. (2015). *Experimental evaluation of the Tools of the Mind PreK Curriculum: Fidelity of implementation technical report* (Working Paper). Peabody Research Institute, Vanderbilt University. https://my.vanderbilt.edu/toolsofthemindevaluation/publications/

Mendive, S., Weiland, C., Yoshikawa, H., & Snow, C. (2016). Opening the black box: Intervention fidelity in a randomized trial of a preschool teacher professional development program. *Journal of Educational Psychology*, *108*, 130–145. https://doi.org/10.1037/edu0000047

Miller, E., & Almon, J. (2009). *Crisis in the kindergarten: Why children need to play in school*. Alliance for Childhood Press.

Missett, T., & Foster, L. (2015). Searching for evidence-based practice: A survey of empirical studies on curricular interventions measuring and reporting fidelity of implementation published during 2004–2013. *Journal of Advanced Academics*, *26*(2), 96–111. https://doi.org/10.1177/1932202X15577206

Moffitt, T., Arseneault, L., Belsky, D., Dickson, N., Hancox, R., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomas, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 2693–2698. https://doi.org/10.1073/pnas.1010076108

Mongeau, L. (2016, August 2). What Boston's preschools get right. *The Atlantic*. https://www.theatlantic.com/education/archive/2016/08/what-bostons-preschools-get-right/493952/

Morris, P., Mattera, S., & Maier, M. (October 2016). *Making Pre-K count: Improving math instruction in New York City*. mdrc. https://eric.ed.gov/?id=ED569994

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). U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/impact-findings-from-the-head-start-cares-demonstration-national-evaluation-of-three-approaches-to-improving-preschoolers-social

Murray, J. (2015). Early childhood pedagogies: Spaces for young children to flourish. *Early Child Development and Care*, *185*, 1715–1732. https://doi.org/10.1080/03004430.2015.1029245

National Center on Quality Teaching and Learning. (2015). *Preschool curriculum consumer report*. https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/curriculum-consumer-report.pdf

National Institute of Child Health and Human Development Early Child Care Research Network, & Duncan, G. J. (2003). Modeling the impacts of childcare quality on children's preschool cognitive development. *Child Development*, *74*, 1454–1475. https://doi.org/10.1111/1467-8624.00617

Nelson, M., Cordray, D., Hulleman, C., Darrow, C., & Sommer, E. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services & Research*, *39*, 374–396. https://doi.org/10.1007/s11414-012-9295-x

Nesbitt, K., Farran, D., & Fuhs, M. (2015). Executive function skills and academic achievement gains in prekindergarten: Contributions of learning-related behaviors. *Developmental Psychology*, *51*, 865–878. https://doi.org/10.1037/dev0000021

Nguyen, T., Duncan, G., & Jenkins, J. (2019). Boosting school readiness with preschool curricula. In A. Reynolds & J. Temple (Eds.), *Sustaining early childhood learning gains: Program, school, and family influences* (pp. 74–100). Cambridge University Press.

Noble, K. G., Houston, S. M., Brito, N. H., Bartsch, H., Kan, E., Kuperman, J. M., Akshoomoff, N., Amaral, D. G., Bloss, C. S., Libiger, O., Schork, N. J., Murray, S. S., Casey, B. J., Chang, L., Ernst, T. M., Frazier, J. A., Gruen, J. R., Kennedy, D. N., Van Zijl, P., …, Sowell, E. R. (2015). Family income, parental education and brain structure in children and adolescents. *Nature Neuroscience*, *18*, 773–778. https://doi.org/10.1038/nn.3983

Noble, K. G., McCandliss, B. D., & Farah, M. (2007). Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental Science*, *10*, 464–480. https://doi.org/10.1111/j.1467-7687.2007.00600.x

Osborn, A. F., Butler, N. R., & Morris, A. C. (1984). *The social life of Britain's five-year-olds: A report of the child health and education study*. Routledge & Kegan Paul.

Outhwaite, L. A., Gulliford, A., & Pitchford, N. (2019). A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes. *International Journal of Research & Method in Education*, *3*, 225–242. https://doi.org/10.1080/1743727X.2019.1657081

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., Duncan, G. J., Dynarski, M., Magnuson, K. A., & Weiland, C. (2017). *Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects. A consensus statement*. The Brookings Institution.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science*, *9*, 144–159. https://doi.org/10.1207/s1532480xads0903_2

Pianta, R. C., La Paro, K., & Hamre, B. (2008). *Classroom Assessment Scoring System (CLASS) manual, Pre-k*. Paul Brookes.

Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology*, *45*, 605–619. https://doi.org/10.1037/a0015365

Preschool Curriculum Evaluation Research Consortium (PCER). (2008). *Effects of preschool curriculum programs on school readiness* (NCER 2008–2009). U.S. Department of Education, Institute of Education Sciences. https://ies.ed.gov/ncer/pubs/20082009/

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, D., Mashburn, A., & Downer, J. (2012). *Third grade follow-up to the Head Start impact study final report* (OPRE Report # 2012–45). U.S. Department of Health and Human Services. https://www.acf.hhs.gov/opre/resource/third-grade-follow-up-to-the-head-start-impact-study-final-report

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschooler' preacademic skills: Self-Regulation as a mediating mechanism. *Child Development*, *82*, 363–378. https://doi.org/10.1111/j.1467-8624.2010.01561.x

Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Smallwood, K., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly*, *23*, 10–26. https://doi.org/10.1016/j.ecresq.2007.09.001

Rogoff, B., Correa-Chavez, M., & Cotuc, M. (2005). A cultural/historical view of schooling in human development. In S. White & D. Pillemer (Eds.), *Developmental psychology and social change: Research, history and policy* (pp. 225–263). Cambridge University Press.

Saracho, O. (2015). Historical and contemporary evaluations of early childhood programmes. *Early Child Development and Care*, *185*, 1255–1267. https://doi.org/10.1080/03004430.2014.989675

Schmitt, S. A., Geldhof, G. J., Purpura, D. J., Duncan, R., & McClelland, M. M. (2017). Examining the relations between executive function, math, and literacy during the transition to kindergarten: A multi-analytic approach. *Journal of Educational Psychology*, *109*(8), 1120–1140. https://doi.org/10.1037/edu0000193

Schmitt, S. A., McClelland, M. M., Tominey, S. L., & Acock, A. C. (2015). Strengthening school readiness for Head Start children: Evaluation of a self-regulation intervention. *Early Childhood Research Quarterly*, *30*, 20–31. https://doi.org/10.1016/j.ecresq.2014.08.001

Sharpe, N., Davis, B., & Howard, M. (2017). *Indispensable policies and practices for high-quality pre-k: Research and pre-k standards review*. New America Foundation. https://www.newamerica.org/education-policy/policy-papers/indispensable-policies-practices-high-quality-pre-k/

Slavin, R. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, *55*, 21–31. https://doi.org/10.1080/00461520.2019.1611432

Solomon, T., Plamondon, A., O'Hara, A., Finch, H., Chaban, P., Huggins, L., Ferguson, B., & Tannock, R. (2018). A cluster randomized-controlled trial of the impact of the *Tools of the Mind* curriculum on self-regulation in Canadian preschoolers. *Frontiers in Psychology*, *8*, 2366. https://doi.org/10.3389/fpsyg.2017.02366

Spivak, A., & Farran, D. (2016). Predicting first graders' social competence from their preschool classroom interpersonal context. *Early Education and Development*, *27*, 735–750. https://doi.org/10.1080/10409289.2016.1138825

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education*, *16*, 1–11. https://doi.org/10.1187/cbe.16-03-0113

Tharp, R., & Gallimore, R. (1982). Inquiry process in program development. *Journal of Community Psychology*, *10*, 103–118.

Tominey, S. L., & McClelland, M. M. (2011). Red light, purple light: Findings from a randomized trial using circle time games to improve behavioral self-regulation in preschool. *Early Education and Development*, *22*, 489–519. https://doi.org/10.1080/10409289.2011.574258

Tough, P. (2012). *How children succeed*. Houghton-Mifflin.

Tough, P., Ravitch, D., & Vander Ark, T. (2009, September). The make-believe solution. *The New York Times Magazine*, pp. 30–35.

United Nations General Assembly. (2015). *Resolution 70/1 Transforming our world: the 2030 Agenda for Sustainable Development* (A/RES/70/1). https://sustainabledevelopment.un.org/post2015/transformingourworld

Upshur, C., Heyman, M., & Wenz-Gross, M. (2017). Efficacy trial of the Second Step Early Learning (SSEL) curriculum: Preliminary outcomes. *Journal of Applied Developmental Psychology*, *50*, 15–25. https://doi.org/10.1016/j.appdev.2017.03.004

Ursache, A., Blair, C., & Raver, C. (2012). The promotion of self-regulation as a means of enhancing school readiness and early achievement in children at risk for school failure. *Child Development Perspectives*, *6*, 122–128. https://doi.org/10.1111/j.1750-8606.2011.00209.x

U.S. Department of Health and Human Services. (2005). *Head Start impact study: First year findings*. https://www.acf.hhs.gov/opre/resource/head-start-impact-study-first-year-findings

U.S. Department of Health and Human Services. (2010). *Head Start impact study*. Final Report. https://www.acf.hhs.gov/opre/resource/head-start-impact-study-final-report

Valentino, R. (2017). Will public pre-k really close achievement gaps? Gaps in prekindergarten quality between students and across states. *American Educational Research Journal*, *55*, 79–116. https://doi.org/10.3102/0002831217732000

van Dijk, W., Lane, H., & Gage, N. A. (2019, October 7). The relation between implementation fidelity and students' reading outcomes: A systematic review of the literature. https://doi.org/10.35542/osf.io/vhrp5

Vorhaus, B., & Meador, D. (2010). *Tools of the Mind fidelity*. Peabody Research Institute, Vanderbilt University. https://my.vanderbilt.edu/toolsofthemindevaluation/resources/fidelitymeasures/

Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (Eds.), *The collected works of L.S. Vygotsky, Volume 1: Problems of general psychology*. Plenum Press.

Webster-Stratton, C., Reid, J., & Hammond, M. (2001). Social skills and problem-solving training for children with early-onset conduct problems: Who benefits? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*, 943–952. https://doi.org/10.1111/1469-7610.00790

Weiland, C., McCormick, M., Mattera, S., Maier, M., & Morris, P. (2018). Preschool curricula and professional development features for getting to high quality implementation at scale:

A comparative review across five trials. *AERA Open*, *4*, 1–16. https://doi.org/10.1177/2332858418757735

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, *28*, 199–209. https://doi.org/10.1016/j.ecresq.2012.12.002

Weisberg, D., Hirsh-Pasek, K., & Golinkoff, R. (2013). Guided play: Where curricular goals meet a playful pedagogy. *Mind, Brain, and Education*, *7*, 104–112. https://doi.org/10.1111/mbe.12015

Weisberg, D., Hirsh-Pasek, K., Golinkoff, R., Kittredge, A., & Klahr, D. (2016). Guided play: Principles and practices. *Current Directions in Psychological Science*, *25*(3), 177–182. https://doi.org/10.1177/0963721416645512

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, *102*, 43–53. https://doi.org/10.1037/a0016738

Wenz-Gross, M., Yoo, Y., Upshur, C. C., & Gambino, A. J. (2018). Pathways to kindergarten readiness: The roles of Second Step Early Learning curriculum and social emotional, executive functioning, preschool academic and task behavior skills. *Frontiers in Psychology*, *9*, 1886. https://doi.org/10.3389/fpsyg.2018.01886

What Works Clearinghouse. (2020). *What Works Clearinghouse Procedures Handbook, Version 4.1*. U.S. Department of Education, Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/handbooks

Wolf, R., Morrison, J., Inns, A., Slavin, R., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness*, *13*, 428–447. https://doi.org/10.1080/19345747.2020.1726537

Wood, E., & Hedges, H. (2016). Curriculum in early childhood education: Critical questions about content, coherence, and control. *The Curriculum Journal*, *27*(3), 387–405. https://doi.org/10.1080/09585176.2015.1129981

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Achievement*. Riverside Publishing.

Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When does preschool matter? *The Futures of Children*, *26*, 21–35.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Phillips, D., & Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. Foundation for Child Development, Society for Research in Child Development.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*, 297–301. https://doi.org/10.1038/nprot.2006.46

Zigler, E., Marsland, K., & Lord, H. (2009). *The tragedy of child care in America*. Yale University Press. https://doi.org/10.12987/yale/9780300122336.001.0001

Authors

Kimberly T. Nesbitt is an Assistant Professor of Human Development and Family Studies at the University of New Hampshire. Dr. Nesbitt's research focuses on the development of cognition in early childhood as it relates to school readiness, with a primary focus on academic achievement and executive function. Her work seeks to understand how to prepare young children for school.

Dale C. Farran is Professor Emerita of Peabody College at the Vanderbilt University. Farran has been involved in research and intervention for high-risk children and youth for all of her professional career. Her recent research includes directing an evaluation of the State of Tennessee's Prekindergarten program and evaluations of alternative curricula.