

SECTION 3, CHAPTER 9

**DESIGNING IMPLEMENTATION
RESEARCH TO GUIDE THE SCALE-UP
OF EFFECTIVE EARLY CARE AND
EDUCATION ACROSS SETTINGS**

Michelle Maier, Ph.D., MDRC

JoAnn Hsueh, Ph.D., MDRC



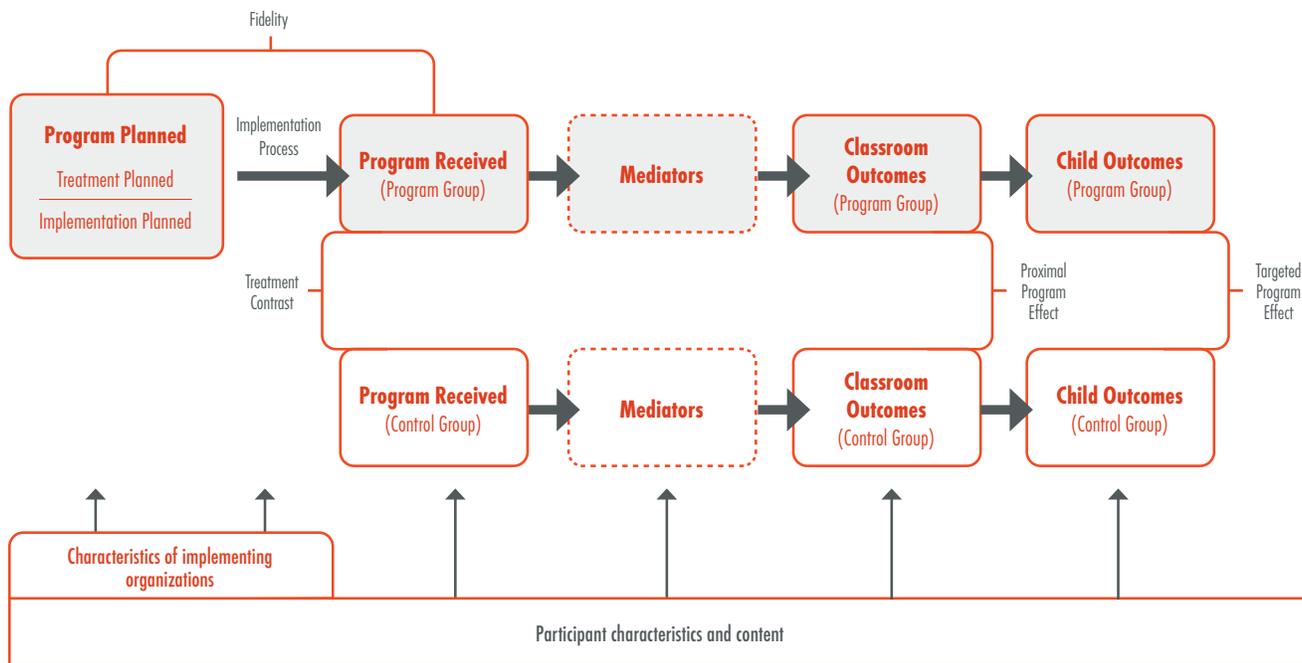
Well-designed implementation research is the key link between small-scale early care and early childhood education (ECE) programs that have been proven to work and large-scale adaptations across populations and settings. Waiting years to see whether programs work provides too little information too late. Ongoing, well-designed implementation research, however, can provide real-time feedback on necessary program adjustments, identify the supports needed to successfully put these programs into action in varied localities and contexts (Martinez-Beck, 2016), and address why and how a program works and under what circumstances. Such research gives the field the information it needs to bring promising programs to wider populations, enabling all children to have access to high-quality learning experiences (Phillips et al., 2017).

This chapter aims to help design strong implementation research to complement rigorous evaluation of ECE programming. It, therefore, has two goals: to provide a set of frameworks to help guide the empirical study of program implementation in an evidence-building context and to discuss potential methodological and measurement problems to consider when taking such an approach. It does not tell developers, researchers, and practitioners what potential areas of inquiry to prioritize in their implementation research. Instead, we aim to illuminate underexplored opportunities and methodological approaches that readers can consider and then apply in their own work. We draw on examples of innovative methodological and measurement strategies from three studies that integrate implementation research into their evidence-building efforts. In doing so, we aim to highlight research opportunities that, by going beyond describing program impacts, can further knowledge and offer a systematic guide to how policy can support at-scale ECE programs that reduce inequities in learning opportunities and disparities in children's outcomes.

A CONCEPTUAL FRAMEWORK

To empirically study program implementation in an evidence-building context, we begin with a conceptual framework for research that examines variation in program effects. Figure 1 outlines the pathway from program implementation to outcomes for ECE centers randomly assigned to receive a program (program group) and those assigned to proceed with business as usual (control group) (Weiss et al., 2014). Using an example program of a new curriculum combined with teacher professional development, researchers often hypothesize the following theory of change: the new program leads to improvements in classroom outcomes (such as more and better instruction) and ultimately to improvements in children's outcomes. Researchers may also propose a set of hypothesized mediators, such as increased teacher knowledge, more positive attitudes and beliefs, or improved teacher practices. Figure 1 illustrates this causal pathway of change as well as other critical aspects of implementation.

Figure 1. Conceptual framework for research examining variation in program effects.



Note: Adapted from Weiss, Bloom, & Brock (2014).

The far left of the framework shows that the program that is planned by developers influences the program received by classrooms with and without access to the program (program group vs. control group). The planned program includes the core components and practices for the new curriculum plus the implementation plan needed to put the program in place (e.g., staff professional development such as training and coaching, technical assistance, and other administrative supports). The procedures, methods, or activities necessary to foster implementation of core components and enact the implementation plan is referred to as the “implementation process.” The relationship between the planned program and what is received by teachers and children is described as “fidelity of implementation.” The line between the program received by the program group and the program received by the control group is termed the “treatment contrast,” which is the difference between the average treatment received with and without access to the program.

Along the bottom of Figure 1 are two boxes representing factors that influence or moderate the specified causal relationships. The top box represents staff and organizational characteristics, which are typically hypothesized to moderate many aspects of program implementation. The bottom box represents characteristics of children within the implementing organization and the organization’s social, physical, economic, financial, and political context. These characteristics are typically thought to moderate the whole chain of events from the implementation process to its effects on outcomes and, in particular, the extent to which income, immigrant, racial, ethnic, linguistic, and cultural backgrounds might affect outcomes.

This framework highlights where sources of variation may be likely to influence program effects and, therefore, underscores where research can focus. This includes operationalizing and measuring:

- fidelity of implementation of the program and implementation plan;
- proximal sources of variation in program effects such as treatment contrast, participant characteristics, and program context;
- distal sources of variation such as characteristics of the implementing organization and of the larger system; and
- potential moderators of these relationships.

In the next section, we further describe what may constitute these sources of variation and how they may be studied.

PROGRAM DEVELOPMENT, IMPLEMENTATION, AND EFFECTIVENESS IN AN EVIDENCE-BUILDING CYCLE

Evidence of program and policy effectiveness arises within an iterative cycle of program implementation, adaptation, and evidence-building activities. The process is often conceptualized as beginning with a program model in an early stage of development (pre-scale-up) that is piloted on a small scale and/or in a relatively controlled setting (for example, under the direct supervision of its developers and with eager volunteer participants). The goal at this stage is to clarify and, if necessary, refine the program goals, target population, and key activities and components as they are being implemented. At this stage, accompanying evidence-building activities designed to evaluate programs commonly entail feasibility studies, demonstrations, pilot assessments, and early efficacy tests.

If early efficacy trials establish evidence of effectiveness when the program is delivered on a relatively small scale, the program may move from the promising to the effective stage. At this stage, efforts typically focus on replicating prior results and/or expanding the program so that it can be tested in more diverse populations and contexts. This undertaking, termed “horizontal scaling,” aims to extend services to a small number of sites (Dunst et al., 2006; Hartmann & Linn, 2008). Accompanying evidence-building commonly entails random assignment efficacy trials through which the program is compared to a business-as-usual comparison/control group. Researchers may therefore adjust the goals for program development, moving on to probing under what conditions, across what contexts, and with what populations the program can be expanded, while also seeking to further test the program’s effectiveness.

As the program continues to mature, it is often scaled more extensively, with the explicit goal of building the level of effectiveness evidence for incorporating the program into an existing system to ensure longer-term sustainability, termed “vertical scaling” (Dunst et al., 2006; Hartmann & Linn, 2008). Evidence-building at this point can thus turn

toward testing and mapping systems, infrastructure, and the institutional supports needed to sustain the model across contexts, populations, and conditions.

Embedded in each of these stages of program development are three aspects of evidence-building research (Knox, Hill, & Berlin, 2018; Metz et al., 2016):

- **implementation** of the program model, which is continually in flux and evolving at each stage of program development;
- **adaptation** of and adjustment and improvement to the defined program model, organizational and system supports, and infrastructure; and
- **building** impact evidence by testing the program model.

In essence, these evidence-building activities have a cyclical relationship; iterative feedback loops aim to strengthen the model as the circumstances, context, and environment in which the program is being delivered evolve, which in turn can help the program operate successfully at each new stage of program development (Knox et al., 2018).

ECE can benefit by aligning implementation research designs and measurement to this evidence-building cycle and stages of program development. As Manno and Miller Gaubert (2016) argue, (a) many implementation research topics and questions are relevant across stages, but depending on whether a program is undertaking horizontal or vertical scale-up, the specific research questions and their emphasis will be slightly different; and (b) even in early stages of program development, implementation research can lay important groundwork for informing future scale-up.

For instance, applicable evidence-building activities in later stages of program development include large-scale studies of evidence-based programs or practices that have expanded widely. Such studies allow researchers and policymakers to examine the effectiveness of these programs across a broader set of contexts, populations, and locations. This type of study has become more prevalent; examples include the Early Head Start Research and Evaluation Study (Administration for Children and Families, 2002), the Head Start Impact Study (Puma et al., 2010), and the Mother and Infant Home Visiting Program Evaluation (Michalopoulos et al., 2015). They provide unique opportunities for researchers to rigorously ascertain sources of variation in program impacts by taking advantage of the multisite designs of such studies, which would not have been feasible in earlier stages (e.g., Weiss et al., 2017). For example, in a secondary analysis of the Head Start Impact Study (Puma et al., 2010), Bloom and Weiland (2015) found substantial variation in impacts generated across sites—variation that suggested Head Start may be more effective when fewer ECE options are available across locations and for dual language learners and Spanish-speaking children. But these kinds of multisite and national evaluations are relatively rare, even though they create unique opportunities to explore variation in the way organizations adapt components of the model and in

intervention fidelity across providers, contexts, or populations. As a result, we have relatively little information about how program models can maintain or even increase their effects as they are widely implemented.

Additionally, even at earlier stages, researchers are presented with opportunities to examine drivers of implementation that can directly or indirectly influence program effectiveness, and the results of such examinations can be useful in addressing scale-up questions of interest (Fixsen et al., 2005). Often, in early stages of program development, less systematic data are captured about more indirect drivers of implementation. However, it is nevertheless helpful to situate the program from this perspective, because these factors become influential sources of variation in implementation and impacts as programs are tested further and scaled. Thus, these topics serve as organizing tools that help researchers explore areas of inquiry for implementation research. The helpfulness of the information yielded from such studies also makes the case for more systematic data collection on these factors and for broadening the conceptualization of measures and research designs that aim to address questions at different stages of development. In undertaking this research, we may be able to build a more systematic body of evidence that can be used to ensure effective, high-quality ECE at scale that improves learning and developmental outcomes for a diverse population of young children.

ADVANCING ECE IMPLEMENTATION RESEARCH: MEASUREMENT AND METHODOLOGICAL CONSIDERATIONS

► Potential methodological approaches in implementation research

Incorporating a strong implementation study in ECE evaluations is necessary for understanding the why behind the effectiveness (or lack thereof) of a program and how best to bring a program to scale. But implementation studies can take multiple forms, using quantitative, qualitative, or mixed-method approaches. Quantitative efforts are more objective, closed-ended, and numerical in nature; use statistical analysis; and commonly rely on methods like surveys, direct assessments, structured observations, and administrative data. Qualitative efforts are more exploratory, subjective, and open ended in nature and typically rely on one-on-one interviews or focus groups (conducted at a single time point or multiple time points), ethnographies, document reviews, unstructured or semi-structured observation, and case studies, among others. Quantitative approaches in implementation research try to quantify constructs of interest—such as the level of fidelity achieved; participants’ attitudes, competencies, and behaviors; and the degree of service contrast observed. In contrast, qualitative approaches may try to explore what underlies participants’ attitudes, competencies, and behaviors as well as their perspectives on how and why fidelity or a service contrast was achieved. Mixed-method approaches combine these two types of methods.

Each approach has notable strengths and weaknesses. The quantitative approach allows us not only to assess the direction and magnitude of relationships among constructs of interest but also to compare the magnitude of

such relationships for different subgroups and to compare the results with those of prior studies that used the same measures with similar populations. Quantitative data also can be captured with larger samples at lower costs than qualitative data, therefore, making such data potentially more generalizable to the population of interest. But the downside to quantitative data is that the constructs of interest need to be prespecified, operationalized, and measured and that measures of these constructs must have been validated for or deemed reliable with the population of interest.

The qualitative approach has the potential to capture rich, descriptive information about people's behaviors, attitudes, perceptions, and experiences as they unfold in contexts that are changing as a function of new policy and programmatic efforts. Further, the often exploratory, inductive, and open-ended coding process of most qualitative studies allows researchers to begin to delineate a series of transactional and dynamic processes in settings that are often difficult to capture with more standard quantitative measurement approaches and thereby develop a theory. At the same time, qualitative approaches do have limitations. Most qualitative implementation studies are fairly small in scale due to the costs of collecting and analyzing qualitative data. Most rely on samples of convenience, developed through snowball strategies. Findings and emergent theories developed with narrow samples require replication and further investigation if researchers are to understand the extent to which the processes identified might be relevant to broader populations and other contexts.

Balancing the strengths and limitations of different methodological approaches in the context of large-scale ECE implementation research can be challenging. We often see focused qualitative endeavors added on to larger-scale implementation and evaluation studies that rely primarily on quantitative data sources. Focusing on a narrow question with qualitative data collection within the scope of a broader implementation or evaluation study provides a unique perspective through which to assess the experiences and perceptions of staff or participants involved with the program or policy initiative and can shed light on and contextualize the findings of the broader study.

► **Topics of inquiry in implementation research**

Drawing on the conceptual framework put forth by Weiss et al. (2014), in this section we highlight six main topics of inquiry for the study of program implementation:

1. Treatment planned, offered, and received
2. Implementation plan and system supports
3. Characteristics of participants
4. Characteristics of the organization/provider implementing the program
5. Institutional and contextual factors external to the organization/provider implementing the program

6. Strength of the service contrast resulting from the program (i.e., the services available to program participants versus those available to control group members)

For each topic of inquiry, we provide a definition and example research questions. We also identify opportunities and underexplored areas, as well as methodological and measurement considerations, to help advance the field. Throughout the section, we draw heavily on three empirical examples that in different ways illustrate how implementation research is critical for the evidence-building cycle:

- ***Making Pre-K Count (MPC) project***, a randomized controlled trial of an evidence-based preschool math program—Building Blocks (Clements & Sarama, 2007)—for which lead and assistant teachers receive two years of training and coaching. Sixty-nine preschools in public schools and community-based organizations with over 170 classrooms and over 2,500 children throughout low-income neighborhoods in New York City form the basis of the longitudinal study, which builds on a relatively extensive body of efficacy evidence conducted by the program developers (e.g., Clements & Sarama, 2007, 2008; Clements et al., 2011; Hofer et al., 2013) and has sought to build an infrastructure that would make its services a longer-term component of the New York City pre-K and educational system. The study features an in-depth implementation research design and measurement approach using both quantitative and qualitative measures. It aims: (a) to shed light on the results of the study’s impact analysis by describing the fidelity of implementation of the curriculum and professional development models, (b) to explore how the math program was experienced by teachers and children, and (c) to guide potential scale-up and replication of Building Blocks across the city.
- ***Researcher-Practitioner Partnerships (RPPs)*** between researchers and the Boston Public Schools’ (BPS) Department of Early Learning that undergird BPS’s data-driven decision making and help build and strengthen its programming. In a long-standing series of collaborations, the RPPs have produced seminal studies about the effects of the BPS prekindergarten program (Weiland & Yoshikawa, 2013), informed the expansion of the BPS prekindergarten model via a delivery system involving community-based prekindergarten and Head Start centers under the purview of the BPS Department of Early Learning (Yudron, Weiland, & Sachs, 2016), and informed more recent efforts via the Institute for Educational Sciences Early Learning Network—a collaboration among BPS, MDRC, the University of Michigan, and the Harvard Graduate School of Education—to extend curricular and professional development reform outward from prekindergarten to second grade.
- ***New York City Early Childhood Research Network***, a hybrid, collaborative early care and education research consortium of eight mixed-methods implementation studies that cut across public school and community-based prekindergarten programs. The studies are part of New York City’s Pre-K for All

(PKA) initiative, an expansion of full-day prekindergarten across the city’s five boroughs. Each one is led by a different research team and is guided by study-specific aims and questions while being tied together by a shared research agenda and a coordinated, place-based sampling approach. Collectively, the studies aim to unpack the complexity of the PKA initiative’s implementation and scale-up efforts. These studies are grounded in the perspectives of the ECE workforce and illuminate overlooked aspects of implementation, such as how administrators, teachers and other support staff, such as coaches, make use of essential elements of the implementation supports prescribed by the PKA initiative, as well as how the system has allocated supports and resources to better address variation in teachers’ and children’s experiences in the classroom. The consortium is a collaborative among academic researchers with the New York City Department of Education, the Mayor’s Office of Economic Opportunity, the Department of Health and Mental Hygiene, and the Administration for Children’s Services; it has funding from the Foundation for Child Development.

► Treatment as planned, offered, and received

The focus of inquiry in this area is *intervention fidelity*, or the degree to which critical components of the program are delivered as expected, in line with the intended program model. Investigation begins with defining the program model, as well as assessing differences between the intended program model and the program model as delivered and received by participants. Fidelity has a number of dimensions (Dane & Schneider, 1998; Durlak & Dupre, 2008), including:

- **dosage**: an index of the quantity of delivery, also referred to as “exposure” (e.g., how many sessions were implemented? How long did they last? How frequently did they occur?)
- **adherence**: the extent to which the specified program content was delivered as described in program materials and manuals
- **quality of delivery**: a measure of qualitative aspects of the manner in which the program components are delivered
- **program differentiation**: the extent to which a program’s theory and practices are distinguishable from other programs, which is gauged to ensure that participants receive only the planned intervention to which they are assigned
- **participant responsiveness**: a measure of participants’ response to the program (e.g., engagement levels, enthusiasm)
- **program reach**: rate of involvement and representativeness of program participants within the intended/eligible population
- **adaptation**: changes or modifications made to the original program during implementation

For programs in an early development stage, this topic of inquiry often focuses primarily on developing and refining the program model and theory of change. In contrast, later stages of program development tend to focus more on the degree to which intervention fidelity along various fidelity dimensions is achieved. Common research questions include “What program was planned and offered?,” “What program components did children receive?,” and “To what degree was there fidelity to the planned program model?”

For example, the MPC study of the Building Blocks program—a 30-week pre-K math curriculum that targets numeric, geometric, and spatial topics and skills—uses online coach logs to capture how often components (whole group and hands-on math centers that are set up daily and small group and computer activities that children participate in weekly) are delivered, the quality of teachers’ delivery of the components, and the overall quality of implementation for lead teachers. Input from the curriculum developers is used to devise benchmarks to monitor the level of intervention fidelity achieved (Mattera et al., 2017). Collection of such information across the school year allows the researchers to describe intervention fidelity in terms of dosage as the extent to which teachers are able to implement most of the main curricular components successfully at levels prespecified by the research team (Morris et al., 2016). It also highlights which curricular components may be more challenging to deliver (computer activities, in this particular case) and how implementation of those components may have changed over time. Further, qualitative findings show that, overall, teachers report engaging in formative assessment activities and differentiation practices that are highly aligned with the training they received (Leacock et al., 2016).

Answering these types of questions in the early stages of program development can help researchers produce meaningful metrics for assessing fidelity to the original model in future scale-up efforts and can help identify which elements of the program model are most essential, reveal which adaptations are appropriate and effective, and make clear what are reasonable expectations for fidelity—all of which are areas of concern once expansion efforts are underway due to cost and operational considerations. In later stages of program development, opportunities arise to describe the degree of variation or consistency in implementation of the program model across populations, locations, and contexts, as well as to link variation in implementation to variation in program impacts. Furthermore, as we underscore later in our discussion, collecting information on intervention fidelity also becomes critically important across all stages of development, as it helps show how fidelity changes as the program is replicated or scaled and makes it possible to examine the strength of the treatment contrast (Cordray & Pion, 2006; Hulleman & Cordray, 2009), even if adaptations to the original program model are made.

Methodological and Measurement Considerations. Most implementation research in this line of inquiry takes a single point-in-time approach to measurement. For example, commonly used methods for measuring intervention fidelity include checklists, surveys, observations, and interviews that typically capture a hypothesized steady state of operation (often thought to be in late winter or early spring in the context of a school year) (e.g., Preschool Curriculum Evaluation Research Consortium, 2008). Such measurement approaches inherently characterize implementation as a static set of processes.

Repeated measurement strategies and designs, in contrast, allow for exploration of dynamic processes and changes in intervention fidelity over time. Measurement approaches, like time use, daily diaries, or surveys collected on an ongoing basis can illuminate consistency in dimensions of fidelity such as dosage, adherence, and quality, allowing researchers to (a) map the arc of changes in implementation as teachers progress toward achieving fidelity to the intended model, (b) predict the variation in implementation that can be expected at different points in time, (c) show how this pattern might differ across multiple years of implementation as the program model matures, and (d) glean insights into the challenges faced by or adaptations made to the program model (see Odom et al., 2010, and Zvoch, 2009 for examples). Findings from MPC, for example, underscore that it's important to understand the arc of implementation within a given school year and across multiple years. Here, with repeated measures of dosage and quality of curriculum implementation collected across two years, the findings suggest that that dosage of all MPC components dips slightly during the winter holiday season (November–December) and toward the end of the year (May–June), a typically more chaotic time (e.g., field trips, moving-up ceremonies). Yet it appears that two years of professional development help teachers start a second school year strong, both in terms of the amount and quality of curriculum implementation, which has implications for the dosage of the curriculum that children receive over a single year. Notably, the quality trends suggest that the overall level of quality achieved each year does not appear to be very different. This kind of information not only can help set expectations when scaling up Building Blocks and when thinking about how curriculum implementation may change across multiple years of implementation but also can suggest potential hypotheses that can be tested in later research.

Processes that feed into the adaptation and evolution of a program model are also important to measure and describe, as they could be relevant to strengthening program effects (e.g., Cannata & Rutledge, 2017; Center on the Developing Child at Harvard University, 2016; Chambers, Russell, & Stange, 2013). For example, the experiences of those implementing the model arguably can best be captured by the qualitative or ethnographic work of staff that links their experiences of transitions and changes brought about by the program model with changes in their delivery of the model. This could help answer interesting implementation questions such as “What are the staff’s perceptions of the model as it is being rolled out?,” “What personal narratives do teachers supply about the purpose of the model and how its components affect their interactions with children?,” and “What difficulties and successes have teachers had in implementing these components, and how do they intersect with their daily experiences working with other staff and with children?” Research that is taking up these issues includes studies being conducted as part of the New York City Early Childhood Research Network that mix qualitative and quantitative methods to better understand the relationships among characteristics of ECE professionals, program components and supports, and classroom instruction in the midst of scaling up universal pre-K.

Another area of potential study in implementation research is analyzing the transactional processes involved in implementing a new model with fidelity, the results of which can then be used for continuous quality improvement efforts. The evolution of BPS’s prekindergarten programming offers a striking example of how such research is

important. In 2013, BPS began rolling out Focus, a system-wide language, literacy, and STEM curriculum that aligns content and instruction from kindergarten through second grade, with the aim of ensuring that kindergarten teachers build effectively on what children are taught in prekindergarten, that first-grade teachers build on what children learn in kindergarten, and so on. Drawing on extant literature and research, the district hypothesized four key ways in which instruction in kindergarten and beyond could be aligned to build off of an already well-developed prekindergarten model: through the content of instruction, the format of instruction, opportunities to tailor instruction to children's skill levels, and professional development support.

The BPS reform effort used a stepwise rollout across the district, an implementation model where the new curriculum for a given grade level is first piloted and then scaled across the district. Yet while the aligned curriculum was being developed and brought to scale across the district, it was unclear whether teachers were implementing Focus as designed or intended, whether BPS should allocate resources and professional development to support teachers in their implementation of Focus and if so, how, and how to ensure that BPS's decision-making around adaptations to the Focus model supported children's gains in the ways intended. In 2016, a collaborative effort was launched to build a data infrastructure that addresses BPS's desire to support children's growth from prekindergarten through third grade by continuously assessing and improving the curricular model. At the core of this work is the development of fidelity tools, co-constructed by researchers and BPS staff. After various iterations and pilot testing of the program, researchers trained BPS coaches and staff to collect fidelity data using the tools. BPS coaches collected prekindergarten data across 40 schools in 2017, kindergarten data across 53 schools in 2018, and first-grade data across 28 schools in 2019; they are planning to collect second-grade data in 2020.

The fidelity tools are designed to capture not only dosage, adherence, and quality of implementation for a given grade but also a set of intentional teaching practices and classroom interactions that are supported by the curricular model and cut across curricular components. These practices and interactions help to extend children's learning and development of unconstrained, higher-order skills—such as receptive and expressive vocabulary, critical thinking, and problem solving—that are thought to contribute to sustained academic achievement and success over time. The research team and the BPS plan to continue their deep and meaningful engagement and collaboration with the aim of advancing the field through careful examination of practices in one district that is working hard to improve students' prekindergarten to third-grade experiences.

The fidelity tools therefore aim to build BPS's capacity to collect and use data that can help guide decision-making around the aligned Focus model. The goals are to better understand the variation in implementation of the aligned model, beginning with prekindergarten and extending through second grade; to identify elements of the curricular model, including components, format of instruction, and intentional teacher practices that are crucial for supporting children's within-year gains and sustained growth over time; and to identify which elements and constructs of fidelity are clear predictors of children's gains and to share that data with teachers in easy-to-understand reports that can

help them strengthen their practices. The fidelity tools will allow the coaches and staff to develop fidelity reports and accessible data they can use to guide BPS decision making.

Last, a generally overlooked aspect of research in this area is children’s classroom experiences as related to the program model. Most commonly, studies capture intervention fidelity as delivered by the provider and less so variation in children’s exposure within classrooms to the program model. Using a propensity-score approach to predict subgroups of children based on levels of absenteeism, Arbour et al. (2016) found that children in Chile who demonstrated a higher likelihood of being absent benefited less from a pre-K program than those who had a lower likelihood of being absent. These findings suggest that measuring and exploiting this source of variation can help illuminate how dilution of the strength of intervention fidelity might undermine program impacts in future scale-up efforts.

MPC has also examined children’s experiences more deeply via a qualitative study (a field visit and teacher interviews), looking closely at how teachers differentiated instruction. Findings show that teachers vary in their beliefs about children and teaching and that these beliefs appear to be related to the ways they modify lessons for children, particularly those who struggle in math. The most prominent differentiation strategy for children struggling in math, the MPC study shows, is changing the difficulty level of an activity. One teacher describes planning the difficulty level for children in the following way:

We played X-Ray Vision One a few weeks ago, so I always have my notes, and I write down my notes on my sheet, so before I do the game for the week, always on a Sunday, I go and I look and I plan and I see what they did the game before, and I write little notes by their name, like, “Start from six,” because the last time, I saw that they did one to ten. They knew it. They counted on from any number, so I said, “They can move up a little.”

Teachers report giving math tasks that go beyond the skills the children currently demonstrated to children they consider to be excelling in math; however, many teachers express hesitation about challenging children they perceived as struggling. These qualitative findings, which would have been difficult to tease out via quantitative methods, have several implications for the project of scaling up the Building Blocks program and for the field’s understanding of differentiated instruction more generally.

► **Implementation plan and system supports**

The implementation plan outlines how the implementing organizations or providers plan to operate the program. The plan includes procedures and activities necessary for fostering implementation of the program model’s core components and practices, such as changes in staffing, professional development (i.e., training and coaching), and other supports like the purchasing of materials or the building of partnerships with other organizations that will

enable the implementing organization to deliver the program model as intended. Related implementation research questions include describing the prescribed implementation supports that are in place; implementation fidelity, or the extent to which the implementation plan is delivered as intended; plans for reaching targeted participants (such as teachers, directors, coaches, etc.); and plans for outreach to and recruitment of children who are currently participating in ECE programming.

To maximize learning in later stages, implementation research should go beyond describing what the implementation plan is and look at how the plan is enacted and why supports seem to work (or not). Further, when a program is being replicated or scaled, implementation research could outline the variation in implementation plans across different providers operating the program. This could include depicting system-level mechanisms that help ensure fidelity to the implementation plan—for example, what management, staffing, funding, and structure of oversight systems are needed to help maintain the dosage, adherence, and quality of training and coaching across multiple providers or geographic locations.

Methodological and Measurement Considerations. Often when high levels of intervention fidelity are achieved, particularly in small-scale studies, details of the implementation plan and supports—and fidelity to the intended levels of these supports—are glossed over (Powell & Diamond, 2016). Commonly used measures tend to focus on structural features of the implementation supports, such as the amount, dosage, and frequency of training or coaching received by recipients; the components of professional development (e.g., in-person observation, one-on-one or small-group consultation); and mode of delivery (in-person, via technology, or through a combination) (e.g., Hamre et al., 2010; Wasik et al., 2013; Powell et al., 2010).

But it is important to capture a host of other aspects of the implementation plan and supports, including:

- *process or content* features, such as the quality of interpersonal dynamics between coaches and teachers, the mechanisms for modeling and providing feedback to participants, the content of professional development, and teacher responsiveness to supports (see Diamond & Powell, 2011; Landry et al., 2009; Neuman & Wright, 2010);
- the extent to which there are *conflicting messages* in the objectives and information being shared with teachers via the program or elsewhere, which may have unintended, countervailing implications for the successful delivery of the intended program model; and
- *factors that facilitate the quality of professional development supports* provided to teachers, such as the characteristics, credentialing, experience, and/or qualifications that make a coach or trainer effective and the supervisory and support systems, caseload specifications, and trainings that can inhibit or facilitate a coach's or trainer's ability to support the delivery of a program model.

Because no two program models are exactly the same, development of measures and unique observational coding schemes are needed in this area of inquiry. Initial implementation research that takes a qualitative approach to understanding the implementation plan and supports could help guide the development and design of appropriate quantitative measures and coding schemes. Further, this kind of information can help explain variation in implementation and program impacts.

The consortium studies being conducted under the auspices of the New York City Early Childhood Research Network employ a variety of strategies to unpack experiences with formal and informal sources of implementation supports for teachers' instructional practice during the PKA initiative. For example, two studies in the consortium (Bank Street College of Education and the National Center for Children in Poverty at Columbia University) take a focused look at how administrators, as leaders of community-based and public school PKA programs, explain adherence to staff members and how they monitor whether staff members are following regulations and standards. The studies examine issues like teacher engagement in training and coaching, use of assessments and curricula, staff qualifications, and whether administrators' teaching priorities are synchronized with teachers' perceptions and prioritization of instructional activities in the classroom. Another study, by the Institute of Human Development and Social Change at New York University, uses network analysis to describe the nature of teachers' social networks within and across PKA programs through which teachers acquire different types of information and mentoring to support their classroom practices. Yet another study, led by Rutgers University's National Institute of Early Education Research, examines how coaches working in PKA programs use their time and explores their perceptions around their roles as influencers of teachers' ECE practices.

In a separate but related vein, a group of studies by Hunter College aims to take a more focused look at how teaching staff use formative assessment tools tied to specific curricula in their planning of classroom activities and implementation of the curricular models. Another study headed up by the Institute of Human Development and Social Change explores how administrators and teachers use existing data sources, such as CLASS scores collected as part of the PKA initiative, to strengthen instructional quality in classrooms through improved professional development and related efforts.

Taken together, the New York City Early Childhood Research Network studies shed light on the processes by which information about standards and regulations are translated and internalized by teachers. Such information could be particularly informative for the design of initiatives in and outside of New York City that aim to strengthen the scale-up of high-quality practices via the existing roles of administrators, mentors, and other informal implementation support networks.

► Characteristics of participants

In implementation studies, the intended target population and the population that ultimately is recruited, enrolled, and served are both of interest. While research suggests that low-income, racial and ethnic minority, and dual language-learning children benefit more from ECE (Gormley et al., 2005; Magnuson et al., 2006; Weiland & Yoshikawa, 2013), an important question as a program is scaled continues to be whether a program is effective for all children or just subgroups of children. Accordingly, implementation studies in early and later phases of development should focus on how the sociodemographics and other risk factors of the families and children that are recruited, enrolled, and served differ from those of the intended target population for the program.

Methodological and Measurement Considerations. As a program is scaled and expands its reach, it becomes important to consider how the characteristics of the actual participants might change as a result of changes in the number of participants being served, the number of providers/organizations delivering the program, and geography. Understanding how the sample population that is successfully recruited, enrolled, and served differs from the intended target population of the program or the samples of earlier studies can help explain program impacts (or lack thereof), as well as guide adaptations to the program model made in response to these differences. Recent trials of Building Blocks in San Diego and New York City, for example, did not have the positive effect on children’s math learning at the end of preschool that prior efficacy trials of the model had suggested it would (Clements et al., 2016; Morris et al., 2016). A confluence of factors may have contributed to the unexpected results, among them, the fact that the preschools participating in these studies served more Hispanic children than those in earlier efficacy studies.

At the same time, disparities in the quality of the ECE learning opportunities of children of color, dual language learners, and those with immigrant backgrounds may also be relevant very early in children’s educational experiences (see the chapters in this volume by Iheoma Iruka and Linda Espinosa). Multiple factors are likely in play, such as unequal access to high-quality educational opportunities, implicit bias and racial stereotyping, and a lack of culturally responsive practices that may better support children of color in classroom environments. While such factors have long been acknowledged in K-12 educational systems, in ECE settings these issues and processes—and how they may build on each other in synergistic and interactive ways—remain poorly understood because we have very little theory and only a small body of empirical research that addresses these matters. The research that has been carried out so far suggests that certain practices, interaction methods, and activities are in fact either culturally responsive or at least acknowledge the diversity of children’s backgrounds, languages spoken, and cultures in classroom learning activities. This is one potential set of strategies for a strengths-based approach to enhancing the learning opportunities and achievement of young children of color, children who are learning dual languages, and children from immigrant backgrounds. Here, implementation research has the unique capacity to contribute to underexplored areas in policy and program models that may facilitate or contribute to disparities in children’s learning opportunities.

Being able to understand and detail the processes at play when thinking about disparities in children’s learning opportunities requires new measurement techniques and focused inquiry in areas like implicit biases that are less typically assessed in implementation research. Indeed, there is a need to develop measures, protocols, and observational tools that will allow us to better capture dynamic processes as they unfold in classrooms. Such information in turn would help us better understand how ECE curricular models, as well as implementation supports and systems, can abate negative influences like implicit biases in children’s ECE experiences.

The New York City Early Childhood Research Network has carried out a set of studies that focus squarely on understanding variation in the delivery and implementation of PKA programming as a way to support learning among children who are dual language learners or who come from immigrant or underrepresented cultural backgrounds. One study, led by Fordham University, examines variation in institutional practices, level of preparation, and the amount and types of support provided to teachers in PKA programs that have concentrations of children with racially and ethnically diverse backgrounds. Another study, run by the Research Foundation of the City University of New York under the auspices of the City College of New York and Teachers College, aims to describe the variability in levels of instructional quality and strategies used to engage underrepresented families across PKA programs.

This consortium of New York City Early Childhood Research Network studies also takes a more focused look at the diversity of the ECE workforce, exploring how this diversity influences the implementation of PKA programming and the supports that are necessary to foster implementation. One study, led by the Research Foundation of the City of University of New York through the Borough of Manhattan Community College, examines male ECE teachers’ perceptions of and experiences with supports during the implementation of PKA programming, including recruitment and retention activities, professional development, and mentoring. Another study, carried out by the Institute of Human Development and Social Change at New York University, uses administrative data to describe how teacher qualifications are distributed across PKA programs and addresses differences across community-based and public school settings. Taken together, these studies illustrate underexplored ways to illuminate how diversity across a large-scale preschool system influences implementation and children’s learning experiences and opportunities in the classroom.

► **Characteristics of organizations/providers implementing the program**

The credentials, academic qualifications, prior work experiences, attitudes, beliefs, knowledge, teaching priorities, readiness, buy-in, motivation to execute the program model, engagement, and stress and burnout of front-line staff carrying out ECE programs as well as supporting staff such as administrators, directors, trainers, and coaches are commonly captured in implementation studies. Other important constructs include information about the organizational climate and culture, the extent to which the leadership is committed, staff turnover rates, the population served, the governance and staffing structure, funding, and the resources and capacity for taking on and maintaining the program and implementation supports.

Examining staffing, management, and organizational characteristics such as these is critical to understanding implementation success and effectiveness or the lack thereof as the program enters different phases of development. Documenting these characteristics in a systematic way early on can impart operational lessons and help predict the types of adaptations to the program model and implementation plan required or the degree of change in preexisting organizational characteristics needed to successfully support the delivery of the program when scaled. As the program moves toward later stages of development and scale-up and the scope of the reach of the program increases, there will likely be more opportunities to exploit naturally occurring variation in organizational characteristics and thereby further assess the importance of these drivers in supporting or inhibiting a program's success and effectiveness.

The importance of moving toward identification of organizational characteristics, management factors, and other processes within organizations that can support or inhibit program success is underscored in a mixed-methods study conducted by Christina Weiland and her colleagues. This study describes the 2.5-year pilot scale-out of the BPS's prekindergarten model into 14 community-based preschool classrooms in high-poverty areas. Weiland and colleagues collected data on instructional quality in each classroom at baseline and at the end of each school year, conducted interviews with key stakeholders at multiple time points, and measured fidelity of implementation in the second and final year of the pilot. The findings indicate that although use of intervention components was high, by the end of the pilot, intervention fidelity of the curricula was generally low, with the community-based classrooms showing lower levels of instructional quality than their BPS-counterpart classrooms (Yudron et al., 2016). Qualitative data pointed to a number of structural factors in the community-based settings that appeared to interfere with implementing the BPS prekindergarten model with fidelity, such as the flexibility permitted in attendance, the lack of common planning time among teachers, the use of mixed-age classrooms, and higher turnover rates among teachers. All of these highlight the need to attend to structural distinctions among pre-K programming delivery models.

Methodological and Measurement Considerations. As the list of potential factors we have listed suggests, the scope of what could be examined is vast. Yet we know that none of these influences operate in isolation from each other but rather are likely linked with others in predictable ways. Tracking dynamic and interactive changes within settings and across levels of ecological analysis could help advance our understanding of contextual factors and their influence on implementation. Changes at a systems level may require intervening levels of institutional and organizational change to ultimately support implementation of the program model and bring about changes in the classroom as experienced by a child. A new curricular model and professional development supports, for example, could influence and be influenced by not only organizational characteristics but also contextual factors over time. Integrating quantitative and qualitative data can illuminate what changes—across different levels and within the implementing organization—shape how the program is being implemented. Research on these linkages and the patterns of organizational, participant, and—as we describe next—system and contextual influences could help the field identify subsets of factors that are most salient.

Implementation research should go beyond describing what happens in the classroom and also look at the broader set of contextual factors that might influence the nature of classroom interactions among teachers and children.

To the extent that the root of inequities in children’s outcomes lies in disparities in exposure to high-quality, adaptive, and responsive learning opportunities in ECE settings, implementation research should go beyond describing what happens in the classroom and also look at the broader set of contextual factors that might influence the nature of classroom interactions among teachers and children. Indeed, such processes may be embedded in institutional systems and settings—as a result of cultural norms, structural biases in ECE settings, and resource allocation—in a way that promotes inequity in children’s experiences. By investigating whether disparities in classroom experiences are evident, as well as how and why they might persist at an organizational level, implementation

research has a unique opportunity to augment our understanding of the role organizational characteristics may play in furthering inequity and how to address it.

► Contextual factors external to an organization

Investigating the contextual factors external to the implementing organization can help to situate the findings from evidence-building efforts of a program at different stages of development. Contextual factors include the funding and policy environment, rules and regulations, and local economic and population characteristics. In early stages of development, implementation studies can aim to describe the systems or structures that are in place as the program is being delivered. This information can be used to guide decisions about the feasibility of scaling the program to particular locations or to identify key funding and policy changes that would be needed for the program to be successfully scaled. When a program operates on a larger scale, systematically documenting contextual factors can provide an opportunity to learn more rigorously about how variation in contextual factors explains when, where, and how a program is more or less effective.

In the MPC project, for example, the research team is interested in describing the context in which MPC is being implemented: New York City preschool programming. The team has found that the preschool landscape in NYC has changed over the course of the study as various reform initiatives have been rolled out, including the Common Core standards, the EarlyLearn initiative (which links quality early care and education standards to child outcomes and has consolidated funds for child care, Head Start, and pre-K to support quality early care services), and Mayor de Blasio’s Pre-K for All initiative (which expanded the number of full-day pre-K slots). These changing circumstances appear to be a driving force in findings regarding the magnitude of the service contrast in MPC, which ought to be taken into account when scaling the model in other contexts.

Methodological and Measurement Considerations. In other policy domains, analysts have assessed patterns of co-occurrence of select contextual dimensions. For example, in the welfare, anti-poverty, and employment policies adopted in the 1980s and 1990s, several common policy dimensions emerged that varied in their mandatory work requirements and their provision of earnings supplements to help sustain families' incomes, time-limited benefits, and child care subsidies, which brought about differential patterns of increases in family income, child care arrangements, and children's outcomes (Morris et al., 2001; Morris et al., 2005). Taking a holistic approach to capturing a combination of potential influences across ecological levels by aggregating information or using community-level data to characterize constructs at higher levels of ecological analysis, researchers could adopt a similar idea to characterize typologies of ECE systems. They could then sample purposefully with these typologies in mind to analyze how this variation might influence program implementation and what impact it might have on children. For example, Coburn et al. (2016) characterize four policy regimes defined along dimensions of alignment with and accountability to the Common Core Standards with hypothesized differential consequences for instructional practices. Following this model, we may be better able to identify sets of processes with cumulative or countervailing influences that moderate implementation or program impacts or that capture the reciprocal nature of influences across levels of system functioning. Such research could guide when, where, and how to scale effective programs.

A related consideration is how challenging it typically is for researchers to amass a sample in smaller-scale implementation studies that allows them to systematically assess and generalize findings with broader contextual and situational influences in mind. To address this issue, the consortium of studies in the New York City Early Learning Network is using an innovative, coordinated, and place-based sampling approach that cuts across public school and community-based prekindergarten programs. A set of community districts in New York City was stratified by the level of resources available in the community using NYC demographic data and city data. From this, researchers selected nine community districts that reflected NYC demographics and were distributed across low, moderate, and high levels of concentrated households living in poverty. Using publicly available administrative data, they identified an eligible pool of PKA programs that served 4-year-old children across the nine community districts. This pool was then used to identify study-specific samples of PKA programs that were stratified to ensure representation of each community district and setting type (public school and community-based PKA programs), as well as racial, ethnic, and linguistic diversity in student-level characteristics, among others. Thus, the coordinated sampling strategy fulfilled practical considerations by ensuring that the research teams did not overtax participating programs with research activities and that each study had a sufficient sample to fulfill its specific aims. It also furthered the learning agenda by guaranteeing some generalizability across the study-specific samples that could help identify emerging cross-cutting themes and show how community-level characteristics might shape findings across studies.

► Service contrast resulting from the program

The effectiveness of a program is a function of a culmination of two sets of influences: the strength of the critical components of the program model being tested and the degree of service contrast (Hulleman & Cordray, 2009), or the difference in experiences with active ingredients of the program model versus other services that might be available to the target population of the program model. We have thus far delineated influences that strengthen or undermine the quality of a program's services as delivered and received by participants, but strengthening the implementation of a program alone is not sufficient to guarantee positive impacts of these investments in ECE at scale. For example, Mendive et al. (2016) found that teachers in a pre-K program in Chile (Un Buen Comienzo) demonstrated fidelity to teaching practices prescribed by the intervention, which they measured by using videotapes of classrooms at three points during the year to assess dosage and adherence. Yet the levels at which teachers engaged in such practices were only modestly higher in the intervention than in the control group, which may help to explain the overall absence of intervention impacts on children's skills.

The research from MPC underscores the need to examine whether some of the primary services being put in place through the program (e.g., training and coaching; math curriculum; math software; monitoring of student progress in math) were being received in the control group. Understanding the services received by the control group, and the degree to which that differs from the program group, guides analysis of the service contrast. This has proved to be particularly important in the MPC study, which, as noted, coincided with several initiatives aiming to improve the academic quality of pre-K instruction in New York City. The team has found that in control sites, a substantial amount of teacher-led math instruction—about 35 minutes in a 3-hour observation—is being delivered at the end of the second year. That is much higher than reported in control group sites in prior Building Blocks studies (Clements & Sarama, 2008; Clements et al., 2011). Such a high level of math instruction in typical New York City pre-K sites may make it harder to detect the effects of Building Blocks (Morris et al., 2016), highlighting the need to interpret impacts (or lack thereof) while considering the service contrast and larger context of the study.

With that said, the amount of math-related professional development and the use of math curricula do yield a distinct service contrast in MPC between program and control conditions (Morris et al., 2016). Quantitative survey data on math-related services, collected at the end of the second year of implementation from school administrators, showed that teachers in control sites received less coaching in math: 66 percent of control sites reported that their pre-K teachers received no coaching in math, and those that did report some coaching described teachers as receiving far less than the program group did. Lead teachers in control sites were offered about 13.8 total hours of training on math, less than half the 30 total hours of training on math that lead teachers in program sites were offered in the same year. Although many control sites reported using some aspects of a math curriculum, there still appeared to be a service contrast: 42 percent of control sites reported using a published math curriculum compared with 100 percent using Building Blocks in program sites, and about half of the control sites reported having computer software with math activities compared with 100 percent of program classrooms that used Building Blocks math computer software.

Thus, a systematic understanding of the service contrast, over the course of different program stages of development, should be a key goal of any implementation study aimed at optimizing the extent to which programs reliably produce positive impacts for young children. This is particularly important given that prior evidence suggests the magnitude of the service contrast is diluted as programs that begin as hot-house, small-scale studies in controlled settings are replicated and scaled (Hulleman & Cordray, 2009). It is thus critical to reassess the strength of the service contrast as the program is delivered in new contexts and environments and with different populations, especially given the changing landscape of ECE programming. Such information can reveal which aspects of the program model add the most value relative to the mix of services that are already available and help to identify strategies for expanding effective programming to reach a broader number of children across localities and contexts.

Methodological and Measurement Considerations. Researchers can bring service contrasts to light in many ways. For example, they can collect descriptive information about other services in the community. Or they can explicitly measure the services received by teachers or children who are in a control or comparison group and then compare the information to the services received by teachers or children in the program group, as in the MPC study. However, capturing the differential in experiences with critical components and practices of the program model requires innovation in measurement and the creation of intervention and implementation fidelity measures that are not only closely tied to the program model and implementation plan but also broad enough that they can be used to capture activities and practices in the control/comparison condition (for examples, see Hulleman & Cordray, 2009; Preschool Curriculum Evaluation Research Consortium, 2008; Bierman et al., 2008; Mattera et al., 2013). When measuring the service contrast, it is also important to assess not only dosage (the amount of services being received or how often they are received) but also the quality of those services.

CONCLUSION

This chapter aims to guide the design of strong implementation research to complement rigorous evaluation research of ECE programming. It suggests three key considerations developers, researchers, and practitioners should bear in mind when designing an implementation study. First, implementation frameworks can guide implementation study design. Second, these frameworks can help determine which critical areas of inquiry to prioritize so that a better understanding of the full story of a program, regardless of where it lies in terms of program development stages, can be developed. Third, the degree of breadth—and in some areas, depth—of measurement for each area of inquiry prioritized is important. Some topics lend themselves to quantitative approaches via data sources like surveys, observational tools, and direct assessments, while others lend themselves to qualitative approaches that make use of interviews, focus groups, time-use reports, or document reviews. A combination of approaches, or an intentional mixed-method approach, may prove best depending on what is prioritized given the program's development stage.

We do not state how to prioritize the various areas of inquiry. Instead, we conclude with several questions to help developers, researchers, and practitioners reflect on and address these considerations, so that their unique implementation study can be poised not only to strengthen the particular program under investigation but also to generate insights as to how policy can support ECE programs at scale that address inequities in learning opportunities and disparities in children's outcomes:

- At what stage of development is the program under study? What level of evidence has already been gathered?
- Where in the evidence-building cycle is the program under study?
- What areas of inquiry are most critical to examine given the program's current stage of development and evidence base?
- Which areas of inquiry may provide information most useful for developing the program and design and measurement strategies?

In sum, we call for concerted efforts to design and enhance implementation research that aims to better understand variation in implementation and program impacts from multiple and holistic perspectives. Such research could guide the development of policy and practice to support and sustain effective programming that reaches a broad number of children in scale-up efforts.

References

- Administration for Children and Families (2002). *Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start: Vol. 1. Final Technical Report*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- Arbour, M. C., Yoshikawa, H., Willett, J., Weiland, C., Snow, C., Mendive, S., ... , & Treviño, E. (2016). Experimental impacts of a preschool intervention in Chile on children's language outcomes: Moderation by student absenteeism. *Journal of Research of Educational Effectiveness*, 9(51), 117-149.
- Bierman, K. L., Domitrovich, C.E., Nix, R. L., Gest, S. D., Welsh J. A., Greenberg, M. T., ... & Gill, S. (2008), Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child Development*, 79(6), 1802-1817.
- Bloom, H., & Weiland, C. (2015). *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study*. New York: MDRC.
- Cannata, M., & Rutledge, S. (2017). Introduction to New frontiers in scaling up research [Special issue]. *Peabody Journal of Education*, 92(5), 559-568.
- Center on the Developing Child at Harvard University (2016). *From best practices to breakthrough impacts: A science-based approach to building a more promising future for young children and families*. <http://www.developingchild.harvard.edu>
- Chambers, D. A., Glasgow, R. E., & Stange, K. C. (2013). The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. *Implementation Science*, 8(117), doi: 10.1186/1748-5908-8-117.
- Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal of Research in Mathematics Education*, 38(2), 136-163.
- Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, 45(2), 443-493.
- Clements, D. H., Sarama, J., Layzer, C., Unlu, F., Germeroth, C., & Fesler, L. (2016). *Effects on mathematics and executive function learning of an early mathematics curriculum synthesized with scaffolded play designed to promote self-regulation versus the mathematics curriculum alone*. Manuscript submitted for publication.
- Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal of Research in Mathematics Education*, 42(2), 127-166.
- Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher*, 45(4), 243-251.
- Cordray, D. S., & Pion, G. M. (2006). "Treatment strength and integrity: Models and methods." In Bootzin, R. R., & McKnight, P. E. (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (103-124). Washington, DC: American Psychological Association.
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, 18(1), 23-45.
- Diamond, K. E., & Powell, D. R. (2011). An iterative approach to the development of a professional development intervention for Head Start teachers. *Journal of Early Intervention*, 33(1), 75-93.
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3-4), 327-350.

CHAPTER 9 DESIGNING IMPLEMENTATION RESEARCH TO GUIDE THE SCALE-UP OF EFFECTIVE EARLY CARE AND EDUCATION ACROSS SETTINGS

- Dunst, C. J., Trivette, C. M., & McInerney, M. (2006). Scaling up early childhood intervention literacy learning practice. *Center for Early Literacy Learning: Cell Papers*, 1(2), 1-10.
- Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* [FMHI Publication #231]. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute; National Implementation Research Network.
- Gormley, W., Gayer, T., Phillips, D. A., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology*, 41(6), 872-884.
- Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downer, J. T., et al. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Associations with preschoolers' language and literacy growth. *Early Childhood Research Quarterly*, 25(3), 329-347.
- Hartmann, A., & Linn, J. (2008). *Scaling up: A framework and lessons for development effectiveness from literature and practice* [Wolfensohn Center Working Paper No. 5]. Washington DC: Brookings Institute.
- Hofer, K., Lipsey, M., Dong, N., & Farran, D. (2013). *Results of the early math project: Scale-up cross-site results*. Unpublished manuscript. Peabody Research Institute, Vanderbilt University.
- Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, 2(1), 88-110.
- Knox, V., Hill, C. J., and Berlin, G. (2018). Can evidence-based policy ameliorate the nation's social problems? *Annals of the American Academy of Political and Social Science*, 678(1), 166-179.
- Landry, S. H., Anthony, J. L., Swank, P.R., & Monseque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology*, 101(2), 448-465.
- Leacock, N., Kou, A., Strassberger, M., Maier, M. F., & Morris, P. (2016, July). *Set goals to meet the needs of your students: Unpacking differentiated instruction on pre-K classrooms*. Paper presented at Administration for Children and Families National Research Conference on Early Childhood, Washington, DC.
- Magnuson, K., Lahale, C., & Waldfogel, J. (2006). Preschool and school readiness of children immigrants. *Social Science Quarterly*, 87(5), 1241-1262.
- Manno, M., & Miller Gaubert, J. (2016). *A conceptual framework for supporting scale-up in implementation research*. Unpublished manuscript. MDRC, New York, NY.
- Martinez-Beck, I. (2016). Where is the new frontier of implementation science in early care and education research and practice? In T. Halle, A. Metz, & I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. xix-xxx). Baltimore, MD: Paul H. Brookes.
- Mattera, S., Lloyd, C. M., Fishman, M., & Bangser, M. (2013). *A first look at the Head Start CARES Demonstration: Large-scale implementation of programs to improve children's social-emotional competence: OPRE Report 2013-47*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- Mattera, S. K., Morris, P. A., Jacob, R., Maier, M., & Rojas, N. (2017). Designing studies to test causal questions about early math: The development of Making Pre-K Count. *Advances in Child Development and Behavior*, 53, 227-253.
- Mendive, S., Weiland, C., Yoshikawa, H., & Snow, C. (2016). Opening the black box: Intervention fidelity in a randomized preschool teacher professional development program. *Journal of Educational Psychology*, 108(1), 130-145.
- Metz, A., Halle, T., Bartley, L., & Blasberg, A. (2016). The key components of successful implementation. In T. Halle, A. Metz, and I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 21-42). Baltimore, MD: Paul H. Brookes.

Michalopoulos, C., Lee, H., Duggan, A., Lundquist, E., Tso, A., Crowne, S., ... , & Knox, V. (2015). *The Mother and Infant Home Visiting Program Evaluation: Early findings on the Maternal, Infant, and Early Childhood Home Visiting program: OPRE Report 2015-11*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Morris, P. A., Gennetian, L. A., Duncan, G. J. (2005). Effects of welfare and employment policies on young children: New findings on policy experiments conducted in the early 1990s. *Social Policy Report* 19(2), 1-20.

Morris, P. A., Huston A. C., Duncan, G. J., Cosby, D., & Bos, J. (2001). *How welfare and work policies affect children: A synthesis of research*. New York: MDRC.

Morris, P. A., Mattera, S. K., & Maier, M. F. (2016). *Making Pre-K Count: Improving math instruction in New York City*. New York: MDRC.

Neuman, S. B., & Wright, T. S. (2010). Promoting language and literacy development for early educators: A mixed-methods study of coursework and coaching. *The Elementary School Journal*, 111(1), 63-86.

Odom, S. L., Fleming, K., Diamond, K., Lieber, J., Hanson, M., Butera, G., ... & Children's Success Project. (2010). Examining different forms of implementation and in early childhood curriculum research. *Early Childhood Research Quarterly*, 25(3), 314-328.

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

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. L. (2010). Effects of an early literacy professional development intervention on Head Start teachers and children. *Journal of Educational Psychology*, 102(2), 299-312.

Preschool Curriculum Evaluation Research Consortium (2008). *Effects of preschool curriculum programs on school readiness: Report from the Preschool Curriculum Evaluation Research initiative*. Washington DC: National Center for Education Research, Institute of Educational Sciences.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., ... , & Spier, E. (2010). *Head Start Impact Study: Final report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Wasik, B. A., Mattera, S. K., Lloyd, C. M., & Boller, K. (2013). *Intervention dosage in early childhood care and education: It's complicated* [Research Brief OPRE 2013-15]. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. http://www.acf.hhs.gov/sites/default/files/opre/dosage_brief_final_001_0.pdf

Weiland, C., & Yoshikawa, H. (2013). The impacts of an urban public prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills: Evidence from Boston. *Child Development*, 84(6), 2112-2130.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, 33(3), 778-808.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, 10(4), 843-876.

Wilson, S. J., Lipsey, M. W., & Derzon, J. H. (2003). The effects of school-based intervention programs on aggressive behavior: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 71(1), 136-149.

Yudron, M., Weiland, C., & Sachs, J. (2016). *BPS K1DS: Piloting the Boston Public Schools' prekindergarten model in community-based organizations*. Boston, MA: Boston Public Schools.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multiple longitudinal examination of provider adherence, status, and change. *American Journal of Evaluation*, 30(1), 44-61.

