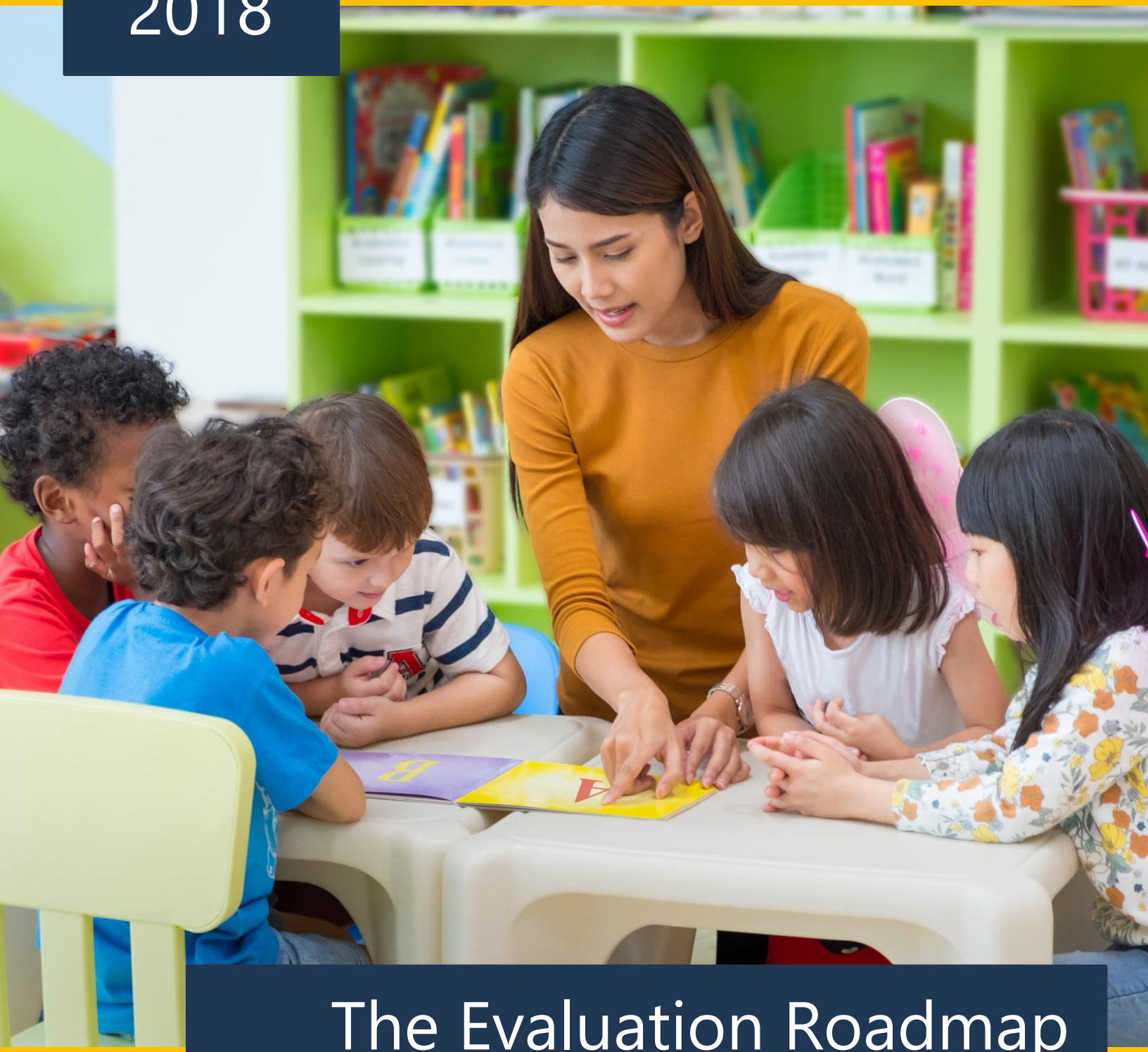
2018



The Evaluation Roadmap for Optimizing Pre-K Programs

The Evaluation Roadmap for Optimizing Pre-K Programs: *Overview*

Anna D. Johnson, Deborah A. Phillips and Owen Schochet

Introduction

In recent years, it has become increasingly clear that one of the best ways to build a productive and prosperous society is to start early—that is, before children enter kindergarten—in building children's foundation for learning, health, and positive behavior. From the U.S. Chambers of Commerce to the National Academy of Sciences, those planning our country's workforce insist we will need more people, with more diverse skills, to meet the challenges of the future. In response, educators have focused on supporting learning earlier, recognizing that early learning establishes the foundation upon which all future skill development is constructed. Identifying and replicating the most important features of successful pre-K programs in order to optimize this potential is now a national imperative.

A wealth of evidence supports continued efforts to improve and scale up pre-kindergarten (pre-K) programs. This evidence is summarized in a companion report to this evaluation roadmap: "Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects".[1] Designing programs in a way that ensures meaningful, short- and long-term effects requires evaluation of programs over time. This goal was the focus of a series of meetings and discussions among a high-level group of practitioners and researchers with responsibility for and experience with designing, implementing and evaluating pre-k programs across the country. This report reflects the best thinking of this practitioner-research engagement effort.

As you prepare to evaluate a pre-K program, we invite you to draw upon this practice- and research-informed expertise to design early education settings that better support early learning and development. Your careful attention to evaluation will help early education systems from across the country identify the factors that distinguish effective programs from less effective ones and take constructive action to better meet our country's educational and workforce goals.

We view this work as the equivalent of building a national highway. We must survey and compare local conditions, adapt designs to suit, map and share progress, and identify and resolve impediments so our country can get where it needs to go. This document is a guide – or roadmap – for those who are building this educational highway system; we hope it will ensure that we optimize our resources and learn from innovations along the way.

There is much good work to build upon. State-funded pre-K programs have been the focus of nearly two decades of evaluation research. This research has produced a large body of evidence on the immediate impacts of pre-K programs on children's school achievement and pointed to some good bets about the inputs that produce these impacts.

But there is more you can do to improve existing programs and ensure that the next generation of programs builds upon this evidence. A central finding from the initial phase of pre-K evaluations is that state and local conditions vary widely, which makes it difficult to draw firm conclusions about the effectiveness of pre-K programs across locations. As the "Puzzling it Out" authors concluded, "We lack the kind of specific, reliable, consistent evidence we need to move from early models to refinements and redesigns".[1] We don't have the evaluation evidence we need to apply lessons learned from first- to second-generation pre-K programs or from one district or state pre-K program to another. In short, we don't have the information we need to inform the continuous improvement efforts called for in "Puzzling it Out" that are so essential to fulfilling the promise of pre-k for our nation's children. It is this challenge that we take on as we attempt to build the next phase of evaluation science on firm ground so that states and school districts can continue to expand and improve their pre-K systems for the benefit of our society.

This roadmap offers direction to states and school districts at varying stages of designing, developing, implementing, and overseeing pre-K programs. It is organized around seven key questions, briefly summarized in this introduction and discussed in more detail in the full report. These questions are best addressed as an integrated series of considerations when designing and launching an evaluation so that it produces the most useful information for you and your colleagues across the country. We summarize these key questions, below.

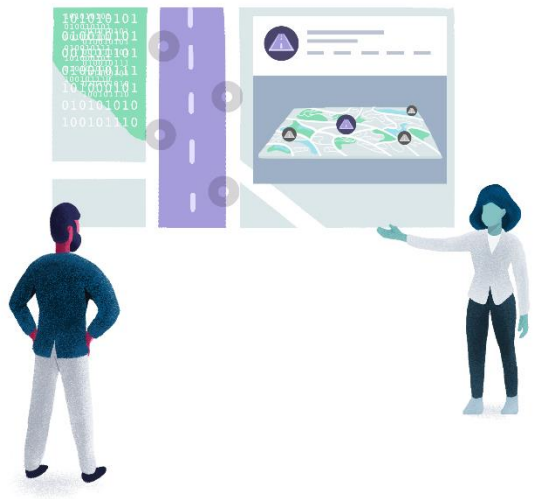
I. What do you want to learn from an evaluation?

— Choosing Your Focus —

The departure point for any evaluation is clarity in the question(s) you want the evaluation to answer. The questions you want to answer will shape the specific information you seek and other decisions you make. Consider these three broad questions: (a) Are we doing what we planned to do (implementation studies)? (b) Are we doing it well (quality monitoring)? (c) Are we doing it well enough to achieve desired impacts (impact evaluations)? (d) What elements of program design account for the impacts (design research)? There is a logical sequence to these questions: if a program has not yet been implemented fully and with fidelity, there is little value in assessing its quality. And if a program has not yet reached an acceptable level of quality, there is little value in assessing its impacts. Once impacts are documented, replicating or

strengthening them requires identifying the active ingredients or "effectiveness factors" that produced them. To wit: transportation officials don't road-test a highway before it has been graded and paved, and work is constantly underway to improve the materials and methods for building better highways. Similarly, don't test the impacts of a pre-K program before it has been fully implemented and, when evaluating impacts, be sure to include assessments of program design features that might explain the impacts. The data you collect while assessing program implementation, quality, and impacts will help you interpret and improve the program's capacity to contribute to children's learning in both the short- and longer-term.

II. What kind of program are you evaluating?



— Gathering Descriptive Data —

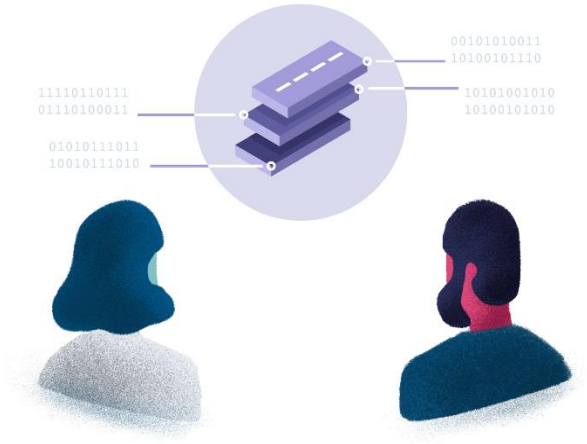
Specificity is key to all that comes after. One of the biggest challenges we face in securing comparable data from pre-K evaluations conducted across districts and states is the fact that there is no single approach to providing pre-K education. Different states have adopted different models and implemented different systems, and districts within states often adapt models and strategies to meet local needs. Many target pre-K systems to children at risk of poor school performance (usually those in poverty), while others offer pre-K to all 4-year-olds and even 3-year-olds, regardless of their socioeconomic status. Programs also differ by length and location. Some provide full-day programs, others provide half-day programs, and still others provide both. Virtually all states provide pre-K in school-based classrooms, but most also provide programs in Head Start and/or community-based child care settings—often with differing teacher qualifications and reimbursement rates. Funding for pre-K programs is often braided into federal and state child care subsidies as well as funding for other programs, such as those affiliated with Head Start, the Individuals with Disabilities Education Act, and the Every Student Succeeds Act.

Importantly, given the wide variation in pre-K programs across and within states, the first step in designing an evaluation must be to map the landscape of pre-K education in your area. Be sure to answer the following questions: How is it funded? Where is it provided? Which children and families participate in pre-K, for how much time during the school day and year, and with what attendance rates?

Moreover, because of this variation in how the provision of pre-K education is approached in different locales, as well as in program design features such as teacher qualifications and support, and reliance on specific curricula or instructional strategies, approaching pre-K as a monolithic program to be evaluated by a single set of broad questions (e.g., Is it well implemented? Did it work?) will not yield particularly

actionable data. The more informative task is to understand the *conditions* under which pre-K is well implemented, provides quality services, and produces impacts. Thus, understanding the key elements of variation in your pre-K program, as well as "best bet" candidates for design features that may explain your findings, is foundational to designing useful evaluations. Research-practice partnerships can be especially valuable in this context.

III. Is the evaluation design strong enough to produce reliable evidence?



— Weighting the Strength of Your Design —

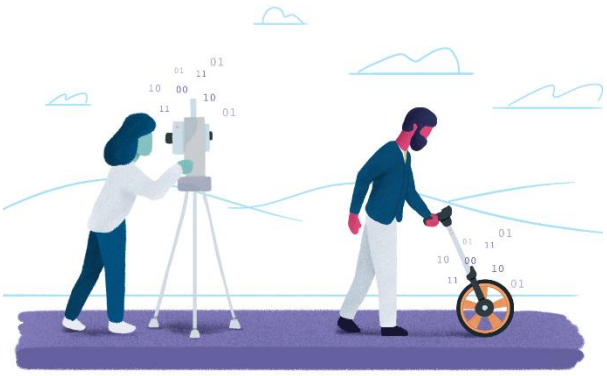
Different, though overlapping, research strategies are needed for different evaluation questions, namely those addressing (a) implementation, (b) quality monitoring, (c) impacts, and (d) program design. For questions about implementation and quality, the core design challenges relate to representation and validity. The representation challenge is to obtain data from a sufficiently representative and large sample of settings (and classrooms within settings), while the validity challenge is to ensure the use of assessment tools that capture variation in the key constructs of interest. For questions about impacts and the design elements that produce them, the core challenges relate to causality and counterfactual evidence (i.e., effects that would have arisen anyway in the absence of the pre-K program or model under review). The causality challenge is to provide the most compelling evidence that impacts can be ascribed to the pre-K program or model under study rather than to other factors. The counterfactual challenge is to be as precise as possible in identifying a non-pre-K (or "different" pre-K) comparison group from which sufficient information can be gathered about children's non-pre-K or other-pre-K experiences. Select a design that best meets these challenges and, at the same time, be sure to collect data that are not subject to bias. Importantly, different participant enrollment strategies (e.g., by lottery, with an age or income cut-off) yield different possibilities for enhancing design strength. Efforts to identify the program elements that account for impacts entails a thorough understanding of key features along which programs vary, as well as current knowledge of elements that are surfacing in other pre-K evaluations as strong candidates for effectiveness factors. Also note: Longitudinal evaluations that address long-term impacts entail additional design considerations (e.g., selection of measures that are appropriate for a span of grades, sample attrition, and how to manage) that should be considered before launching a study of longer-term impacts.

IV. Which children and how many should you include in your evaluation?

— Weighting the Strength of Your Design —

This question is about sampling strategy. The first step is to consider whom your program serves and whether you want to document its effects on specific subgroups. If so, the next step is to determine whether to identify subgroups by participant characteristics (e.g., home language, special needs status, race and gender, degree of economic, or other hardship) or program features (e.g., part-time or full-time schedule; school-based classroom or other setting; number of years children spend in the program). You may want to know, for example, if all children in the program have equal access to well-implemented programs in high-quality settings or if access varies across participants. Subgroup studies require samples of sufficient size and representation as well as measurement tools that are suitable for all participants. Another critical task is identifying the right comparison group. Ideally, the evaluation will compare "apples to apples." That is, it will compare children who do participate in the pre-K program with similar children who do not – or children who attend pre-K programs that do one thing or have certain features to those who attend programs that do another thing or have different features (e.g., school- vs. community-based programs; programs using one instructional model or curriculum versus another) – so that program participation or model is the main difference between the two groups. Random assignment designs are the best way to ensure apples-to-apples comparisons, but there are other commonly used and well-respected approaches to use when random assignment is not possible. Even with alternative approaches, the collection of pre-test information about children who do and do not participate in pre-K or who participate in different pre-K models *prior to enrollment* will strengthen your capacity to produce reliable conclusions.

V. What are the most important data to collect?



— Fitting Tools to Task —

Choosing measures for an evaluation study can be time-consuming and expensive. A good

starting place is to familiarize yourself with data that has already been collected (e.g., administration data, school records, testing data, enrollment and financial forms, etc.) and assess its completeness and quality. Then, determine what is missing, keeping your key questions in mind. If your questions are about ensuring access to high-quality pre-K classrooms for all children, you will collect different data than if your questions are about designing classrooms to promote inclusive peer interactions or increasing the odds of third-grade reading proficiency. Draft tightly focused questions to avoid the temptation to collect a little data on a lot of things; instead, do the reverse: collect a lot of data on a few things. It is helpful to think about four buckets of data to collect: (a) child and family characteristics that may affect children's responses to pre-K programs, (b) characteristics of teachers and other adults in the program who support implementation and program quality, (c) pre-K program design features and dosage, and (d) children's outcomes tied to pre-K goals and theories of change. Questions that address pre-K implementation and quality monitoring will necessarily focus on pre-K program characteristics and dosage, but information on child and family characteristics will be helpful in interpreting the findings. Questions that address pre-K impacts require coordinating measurement from all three buckets. Impact questions that extend beyond outcomes at the end of the pre-K year entail additional data-related considerations, such as pre-K-to-elementary system data linkage and how best to measure your constructs of interest at different ages.

VI. How will you get the data?

— Collecting Reliable Data —

There are many more and various data sources and strategies for obtaining data than you might imagine. Existing administrative and program data (e.g., state, district, and school records, Quality Rating and Improvement System data, data from other systems such as child welfare services or income support offices) are a good place to start, although it is critical to assess the completeness and quality of these data. Prior and ongoing pre-k evaluation efforts in other locales offer fertile ground for data collection approaches to consider (see also the benefits of research-practice partnerships). There are cost, time, training, and intrusion trade-offs to consider when deciding whether to ask parents, teachers, coaches, or principals to provide data; to conduct direct assessments of children; to independently observe classrooms; and so on. In the end, you want to ensure that, having decided *what* to measure, you next decide *how and from whom* to collect those data to ensure maximum data quality.

VII. What else should you consider before you start? *Infrastructure, partnerships, and stakeholders*

— Seeing Your Work in Context —

Planning, launching, and seeing an evaluation effort to completion (and then considering its implications for policy, practice, and next stage research) are all essential parts of effective pre-K programming. The feasibility and quality of an evaluation depends on setting up the necessary infrastructure (e.g., implementing ethical research practices and procedures, ensuring adequate staff and clarifying roles, storing and archiving data, setting up advisory committees and review processes, producing reports, etc.). Forging partnerships with local universities and colleges can be helpful in this regard. Policy-practice-research partnerships can also create an evaluation team with a broader collective skillset and lend important external credibility to the findings your efforts produce. Pre-K programs and their evaluations affect many stakeholders, including families, teachers and support staff, principals, superintendents, and other education policymakers. Informing these stakeholder groups about your evaluation at the beginning of the process, and keeping them in the loop as the evaluation proceeds and begins to produce evidence, is not only best practice for strong community relations but will greatly enhance the chances that your evidence will be used for program improvement efforts. And that is, after all, the goal of providing a strong roadmap for your evaluation effort.

VIII. Concluding thought

We have designed this roadmap for optimizing pre-K programs across the country so that children have a better chance of succeeding in school and beyond. This depends on building a stronger pre-K infrastructure that is based on sound evaluation science. We aim to provide sufficient detail and advice to ensure that future pre-K evaluations will get our country where it needs to go. We view this work as the equivalent of building a national highway. As social scientists who have engaged over many years with local and state policymakers and practitioners to conduct research about state and district pre-K programs, we have struggled with many of the questions addressed here. By sharing a roadmap and the knowledge gained from our experiences, we hope to contribute to construction of a strong, reliable "highway" infrastructure of pre-K programs that better meets our country's educational and workforce goals.

[1] D.A. Phillips, M.W. Lipsey, K.A. Dodge, R. Haskins, D. Bassok, M.R. Burchinal, G.J. Duncan, M. Dynarski, K.A. Magnuson, and C. Weiland, "Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects," in *Current State of Knowledge on Pre-Kindergarten Effects,* (Washington, DC: The Brookings Institution, 2017).

The Evaluation Roadmap for Optimizing Pre-K Programs

Anna D. Johnson, Deborah A. Phillips and Owen Schochet

Evaluations of pre-K programs in districts and states across the nation have produced strikingly uniform evidence of short-term success. Children who attend pre-K are better prepared for school than children who do not attend pre-K. This is was the conclusion of a panel of pre-K experts in a companion report to this evaluation roadmap. "Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects" summarizes what current evaluation evidence tells us about the impacts of pre-K programs and concludes with a call to accompany ongoing implementation and expansion with rigorous evaluation of pre-K impacts and the factors that produce and sustain impacts.[1] But, designing and evaluating programs in a way that contributes to continuous improvement over time takes proactive, intentional, and sustained planning. This roadmap is designed to contribute to such efforts – to build the next phase of pre-K evaluation science on firm ground so that states and school districts can continue to expand and improve their pre-K systems for the benefit of our society. As with any good roadmap, we guide you through the steps of knowing where you want to go, what you have to work with and how to prevent problems that would derail the effort. This guide offers direction to states and school districts at varying stages of designing, developing, implementing, and overseeing pre-K programs. It is organized around seven key questions that must be addressed when designing and launching an evaluation so that it produces the most useful information. Those questions are:

What do you want to learn from the evaluation?

What kind of program are you are evaluating?

Is the evaluation design strong enough to produce reliable evidence?

Which children and how many do you want to include in your evaluation?

What are the most important data to collect?

How will you get the data?

What else should you consider before you start?

I. What do you want to learn from an evaluation?

— Choosing Your Focus —

An essential first step in planning an evaluation is to identify the question the program designers seek to answer. Most people, when they think about evaluations, think about impacts on the program participants -- the "did it work?" question. But, there are other equally important evaluation questions that may need to precede and/or accompany the impact question. Sometimes you need to know whether a program was implemented properly (implementation studies). This entails understanding your program goals, your theory of change, and how you have operationalized these goals and ideas about how to achieve them. If, for example, your goal is to improve third-grade reading scores, you may have initiated a new pre-K reading curriculum. Of course, you want to know if the curriculum is producing the desired outcome, but first you need to know if the curriculum has been implemented with fidelity. Are teachers using the curriculum guide, are they spending the required time on reading instruction, and are they using the correct assessment instruments?

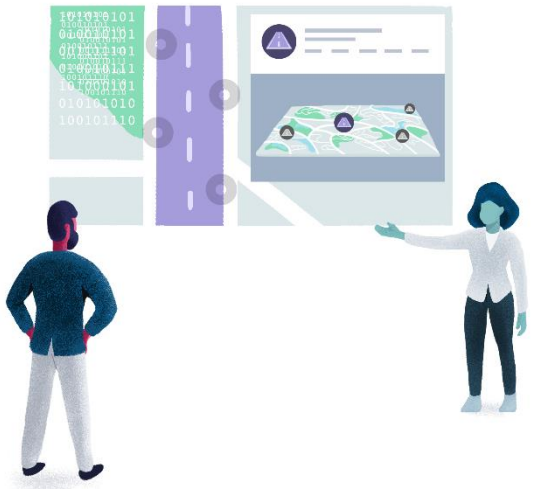
Sometimes you need to know if the program is being done well, namely meeting your quality standards or benchmarks (quality monitoring). To stay with the reading example, even if you find that the curriculum is being implemented with fidelity, there is likely variation across classrooms and schools with regard to *how well* it is being implemented. To capture program quality or other elements (e.g., bilingual or English-only instruction, extent of inclusion of children with special needs) requires observations and assessment instruments that can tap meaningful variation in dimensions of classroom processes that matter for children, e.g., how well do teachers ensure that the children are engaged with the reading lessons and materials, to what extent do they make sure all children are progressing through the curriculum?

Turning to impact evaluations, you might be interested in average impacts across all study participants or in impacts on particular subgroups of students—or both. As evaluators, you must also identify the outcomes on which you expect to find impacts (see Section V for a deep discussion of outcomes to measure). Are you evaluating a narrow set of outcomes or a broad set? Are you evaluating a program's direct effects on outcomes (reading test scores) *and* indirect effects (self-regulation skills that may affect the children's capacity to focus on reading lessons and thus improve their test scores), or only the former? At the end of the pre-K year, evaluators will likely want to assess outcomes that best align with program goals, such as academic achievement, social skills, health, and so on. As children progress through school, you may want to expand the evaluation to include outcomes that are not as closely aligned with the program, such as placement in special education programs, grade repetition, or attendance rates.

Given the strong evidence base on pre-K impacts, at least in the short-term, many in the field are now turning their attention to understanding the program design elements, or active ingredients, that distinguish effective programs (or classrooms). In other words, rather than (just) asking "did it work?" they are asking "why", "under what conditions", and "for whom did it work?" Some of the most promising design elements currently under investigation are (i) curricula that are known to build foundational and increasingly complex skills and knowledge, coupled with (ii) professional development and coaching that enable teachers (iii) to create organized and engaging classrooms.[1,2]

In sum, before you embark on an evaluation of your pre-K program, ask yourself which question you want and need to answer. This will then direct you to (a) an implementation study, (b) a quality monitoring study, (c) an impact evaluation, or (d) a study of design elements. Importantly, these are not necessarily mutually exclusive endeavors. Just as building a highway entails starting with the right materials, grading and paving the road, and then testing whether the road performs as expected, it makes sense to assess whether the important elements of your program are in place (implementation) and reaching an adequate level of quality before assessing impacts and attempting to account for them. The good news is that at least some of the data that you collect to address one question will be informative in addressing the next question as you move along your roadmap in this progression of evaluation studies.

II. What kind of program are you evaluating?



— Gathering Descriptive Data —

The first step when conducting a pre-K evaluation, whether it is focused on implementation, program quality, or program impact, is to understand how pre-K is delivered in your district and/or state. It is critical to know, in detail, *what* you are studying so you can interpret your findings accurately and consider their implications in a real-world context. Key metrics to capture are: (a) funding mechanisms (how is it funded?), (b) where the program is provided (e.g., in a school or community center?), (c) eligibility (whom does it serve and how are they selected?), and (d) program hours and length. Keep these metrics in mind when considering what data to collect (Section V).

Funding & Where the Program is Provided

Increasingly, states are coordinating and consolidating funds across the fragmented early education and care sector to better meet children's needs. Pre-K programs can blend or braid funds from federal, state, and/or local sources, each

of which carries its own stipulations and guidelines, to create the highest-quality and most comprehensive program possible. These "mixed delivery" programs increasingly populate the landscape of pre-K programs in the United States. Together, they seek to meet the diverse needs of children and families. At the same time, this can pose a challenge to evaluation efforts: as funding streams merge together, comparisons between children's experiences in different settings become more difficult to make. Comparing children in a Head Start program with those in a pre-K program, for example, is more complex if the Head Start program also accepts pre-K funds. Comparisons across states that deliver and fund pre-K in very different ways are also difficult. Some states, like Oklahoma, meet demand by providing pre-K primarily through the public school system, while others, like Florida and Georgia, take a broader approach, providing programs in settings including public schools, Head Start, and various types of child care centers.

Eligibility

Evaluators must consider myriad eligibility issues when defining the pre-K program's target population, each of which has implications for the research design. Participants' age is a key consideration. Most statewide pre-K programs focus primarily on children who turn 4 the year before they enter kindergarten, though some allow 3-year-olds to attend.[3] In these cases, some 4-year olds will have experienced two years versus one year of the program, thus creating a pre-existing association between age and program duration.

Another key consideration relates to availability. Is the program universal (offered to all children) or targeted (limited to a particular group)? Each approach has benefits and drawbacks, and decisions are often driven by cost.[4,5] This difference, though, has important implications for pre-K evaluation studies. Evaluators must identify the *right* children to serve as a comparison group. This is relatively straightforward when studying targeted programs; researchers can simply match or otherwise compare children who enroll in pre-K with similar children who do not. Comparisons are more challenging when evaluating universal programs because children who do not participate may differ in key ways from those who do (see discussion of the Selection Challenge, below).

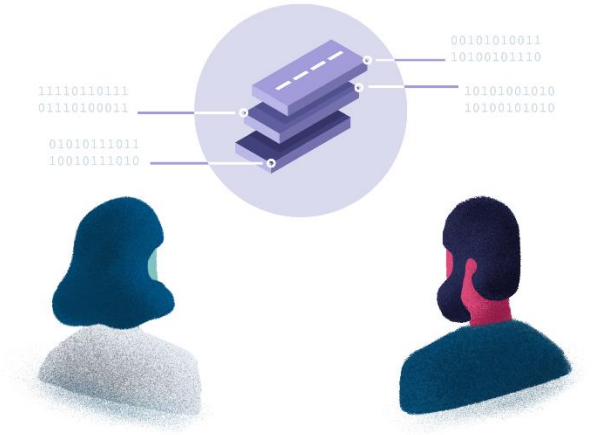
A common eligibility criterion in targeted programs is family income. This is the primary eligibility factor used by Head Start, which determines income eligibility based on the federal poverty level (FPL).[6] Children in families whose incomes fall below a certain percentage of the FPL, such as 100 or 130 percent, may qualify to attend. (In some places, state median incomes are used instead.) These income-targeted programs focus primarily on children from economically disadvantaged families, based on the theory that they reap especially strong benefits from participation. Other types of targeting criteria identify children who are vulnerable or at-risk for reasons other than family income, such as those who have special needs, have teen parents, live in households with family members who do not speak English, have experienced abuse or neglect, or are in foster care.[3] Whether a program is universal or targeted, eligibility criteria must be clearly specified by the funding agency, as these details will influence your subsequent decisions about research design and comparison groups.

Program Hours and Length

Pre-K programs operate for different hours and months of the year. A program might offer part-year (e.g., 9-months or a school-year calendar) or full-year programming, and/or may provide full-day and half-day options or offer extended care before and after the standard 6.5-hour K-12 school day. A program's available hours of care, and whether those hours align with caregivers' work schedules, affect which children can attend.

We anticipate larger effects in skill development as children's exposure to pre-K education increases. Therefore, an evaluation that successfully accounts for variation in programs should require documentation and inclusion of information about program hours and length duration and average dosage (e.g., hours per day, attendance) measures. Carefully documenting intensity and duration is an important step in ensuring that program evaluators consider these measures when characterizing program effects, identifying comparison groups for rigorous study designs (see Section III), and, when appropriate, controlling for dosage or duration when assessing the link between program participation and outcomes.

III. Is the evaluation design strong enough to produce reliable evidence?



— Weighting the Strength of Your Design —

The evaluation design you use depends, first and foremost, on the questions you seek to answer. Regardless of whether your questions are about implementation, program quality, or program impacts, detailed information about the pre-K program will enable you to select settings (and classrooms within settings) that are representative of the pre-K program or population of interest (**the representation challenge**). This can be accomplished through a random sampling of settings and/or classrooms across the whole pre-K program. Or, it can be accomplished by sampling that first sorts settings/classrooms into certain buckets (e.g., those with more or less experienced teachers or those serving larger or smaller numbers of students who are learning more than one language at a time) and then randomly sample within each bucket. This latter approach is called stratified random sampling. Carefully selecting the programs to study, as well as a large enough sample of them, will ensure that your findings accurately reflect the pre-K program as a whole.

A strong evaluation uses assessment tools that capture variation in the key constructs of interest (**the validity challenge**). A valid measure is one that measures what it purports to measure, e.g., does a measure of vocabulary knowledge show a strong association with language achievement tests? Evidence of validity can often be found in reports on the development of the measure or in large-scale studies using the measure.

The biggest design challenges occur in efforts to assess and interpret the impacts of pre-K programs. When asking whether a pre-K program works, the answer must be ascribed to the pre-K program itself and not to other factors, such as differing family circumstances of children who participate and those who do not (**the causality challenge**). Ideally, we would compare the effects of a pre-K program on a group of children to what would have happened if those same children had not attended the program. But, this is impossible. The next best option is finding some kind of comparison group of children who did not participate in pre-K to assess program impacts. The similarity between the group of children who are or will be attending the program and the comparison group of children will greatly influence the validity and credibility of your findings. Finding virtually identical pre-K participants and comparison children is difficult because children who attend pre-K are often different from those who do not. In targeted programs, for example, the families of children who attend pre-K may have lower incomes or more risk factors than those who do not, especially if program administrators seek out the "neediest" families.

In the context of thinking about children who received pre-K and those who did not – the comparison that generates the answer to "does pre-K work?" – you must also think about what design elements of the pre-K program give rise to the effects, if positive effects are found. An understanding of the "why" pre-K might work (not just "if" pre-K works) requires data on the specific program features that current research suggests are particularly powerful predictors of program impacts (e.g., quality of instruction in specific learning domains; time spent on instruction in those specific domains; see "Puzzling It Out").

Other criteria can affect which eligible families choose to enroll their children in pre-K (**the selection challenge**) – who selects pre-K and who does not? For instance, mothers with higher education levels are both more likely to enroll their child in pre-K and also more likely to engage in cognitively stimulating activities at home, thereby muddying the association between pre-K exposure and children's learning outcomes.

Finally, you will be best able to interpret your results if you have information on the experiences of the comparison children during the pre-K year(s) (**the counterfactual challenge**). Counterfactual refers to alternative realities or options for the children who attended pre-K. For example, were the children who were not enrolled in pre-K instead at home with their parents? Or were they in a different early care and education arrangement? There is some evidence that when comparisons are made between children who attended pre-K and those who stayed at home, the positive impacts on the pre-K children are larger (by comparison) than when comparisons are made between pre-K children and children who attended other early education programs.[7,8]

Representation Challenge:	Does the evaluation sub-sample represent the population of interest?
Validity Challenge:	Do the assessments actually capture the outcomes of interest?
[FOR IMPACT STUDIES]: Causality Challenge:	Can you confidently ascribe impacts to the pre-K program?
Selection Challenge:	Can you be sure that impacts are not due to characteristics of the enrolled children or their families that affect their odds of participating in pre-K?
Counterfactual Challenge:	Do the impacts vary with the circumstances of those not attending pre-K?

Randomized Control Trial (and Close Approximations Thereof)

When choosing programs for impact evaluations directed at questions about whether pre-K participation boosts children's learning relative to some alternative, opt for programs that randomly select participants from among eligible children. Random assignment of slots in a pre-K program ensures that the children who are assigned a space are, on average, the same as those who are not on all other important dimensions. This protects against factors that can contaminate your findings, notably Selection Challenges. Consider a situation where parents can freely enroll their children in a voluntary universal pre-K program. In this case, the children who attend the program might be, on average, very different from the children who do not attend. In addition to different family income levels, these children may also differ according to family background, parental involvement, or other dimensions. These dimensions may influence standardized test scores in later grades, socio-emotional adjustment, or other outcomes of interest. Comparing the outcomes of attendees and non-attendees would then confound the true impacts of pre-K with the impacts of these other factors and characteristics. Randomly assigned programs can inoculate your study against these differences.

In contemporary scaled-up pre-K programs, where random assignment is not always possible, the next best option is a lottery, which (ideally) randomly assigns slots to children when demand for the program exceeds available

space. If interested students are randomly selected to participate in a pre-K program and data can be collected about both groups of students (those who won and those who did not win the lottery), then a fairly straightforward impact analysis can be conducted by comparing the average outcomes of those selected to participate by the lottery (i.e. the treatment group) with the average outcomes of those who were not selected (i.e., the control group).

Under random assignment, the effect of the treatment – in this case, pre-K participation – can be estimated by subtracting the control group's mean outcomes (such as test scores, special education assignments, grade retention, etc.) from the treatment group's mean outcomes. Comparing mean outcomes between those randomly assigned to treatment (pre-K) and control (no pre-K or pre-K alternative) groups yields an estimate of the impact of being offered a pre-K slot. In the parlance of program evaluation, this is known as the "intention to treat" (ITT) estimate. Applied to a lottery design, in some lotteries not all winners accept offers to attend. To account for this, and to estimate the impact of *attending* a pre-K program (the so-called "treatment on treated" effect [TOT]), one must also estimate the difference in the shares of children attending pre-K between the treatment and control groups. Say, for example, that 80 percent of the treatment group enrolled in the program, and 20 percent of the control group found a way to enroll in a different pre-K program (e.g., in another school or school district). In this case, the treatment-control difference in the shares attending pre-K would be 60 percent (0.8 – 0.2 = 0.6). To estimate the impact of *attending* pre-K, one could then divide the "intention to treat" estimate by the treatment-control difference in enrollment shares (or, as in the case above, 0.6). The "treatment on treated" effect will thus be larger than the "intention to treat" estimate if there is anything other than perfect compliance with random assignment.

Regression Discontinuity Design

A randomized evaluation design may not always be feasible. For example, a pre-K program that focuses on children who are in greatest need of pre-K services (with "need" defined on the basis of family income, test scores, or some combination of criteria) cannot be evaluated using a randomized design. By intent, the children who attend this program will be lower-income and/or have less academic preparation than those who do not. There is no random assignment; comparing the outcomes of attendees with non-attendees would confound the impact of the program with the impact of the differing "needs" represented in the treatment and control children. However, alternative and strong evaluation designs are possible.

Regression discontinuity designs (RDD) are one of these alternatives. This design can be used whenever children are assigned to pre-K (or not) based on some arbitrary cutpoint. Here we provide an overview of a generic and relatively simple RDD. For a fuller treatment of the appli-cation of RDD to pre-K studies, see Lipsey, Weiland, et al., 2015.[9]

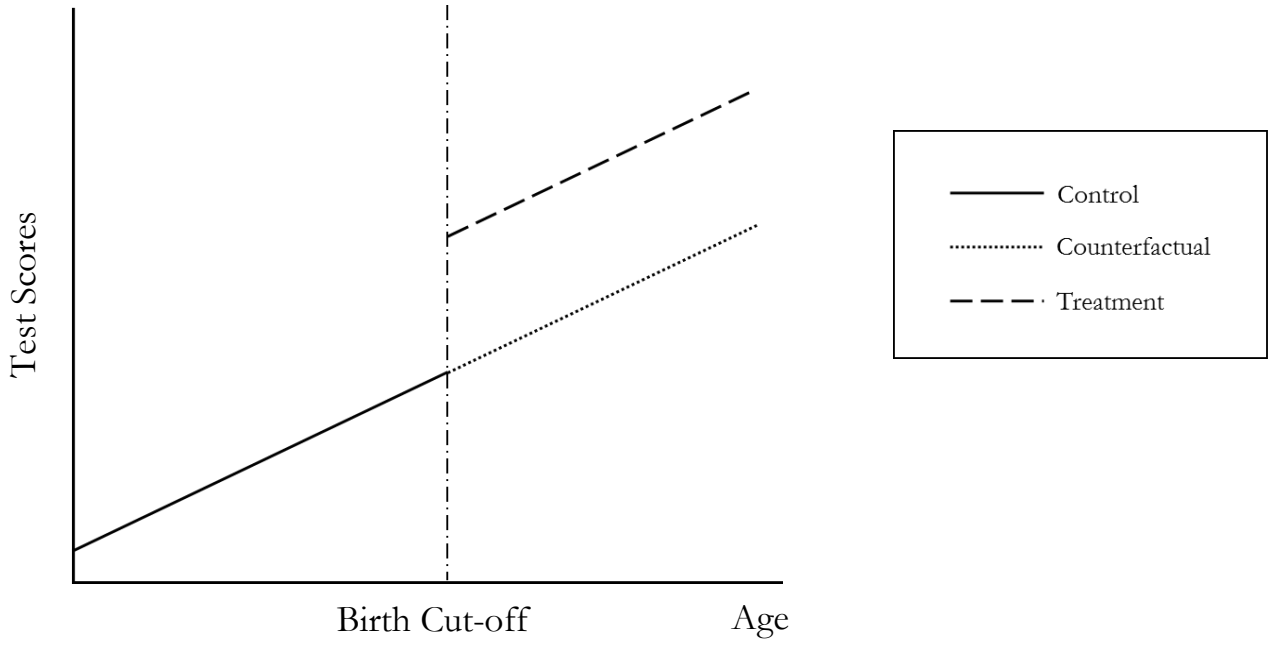
Let's consider a school district with a universal pre-K program that assigns children to pre-K using date of birth (as is often the case with entry into Kindergarten). If eligibility for that program is determined by date of birth, such as September 30th (e.g., children must be 4 years old by September 30th to be eligible to attend), then children who turn 4 in the days and weeks before September 30th are eligible, while those

who turn 4 soon after are not. There is no reason to believe that children born on September 30th and those born on October 1st are systemically different from each other; this is a key assumption of this approach. But, one will receive pre-K and the other will have to wait another year to enroll. As a result, pre-K eligibility by age is almost "as good as" random assignment among a subset of children whose birthdays are close to the Sept. 30th cut-off. By September 30th of the following year, the first group of children will have attended pre-K and the second group of children will just be entering pre-K, allowing for strong estimates of the impact of the pre-K program.

Figure 1 provides a hypothetical illustration of this regression discontinuity design. The broken line to the right of the birthday cut-off date shows the hypothetical test scores of the pre-K group, and the solid line to the left of the cut-off date shows the hypothetical test scores of the children who are not enrolled in pre-K because they are too young as determined by the cut-off. The dotted line to the right of the cut-off date depicts the counterfactual, or what the alternative reality would have been for the pre-K participants had they not attended pre-K. If the real data show a significant difference (a gap) between the test scores representing the pre-K group and those representing the counterfactual (the children who have to wait a year), it is reasonable to ascribe it to the impacts of the pre-K program.

Another example of a regression discontinuity design uses an income cut-off as the dividing line between those children who attend pre-K and those who do not. Consider a school district that has decided to fund two pre-K classrooms of 20 children each, for the lowest-income children in a community. Far more children are eligible than the 40 that can be served by the program. School administrators rank the

Figure 1



children by income, from the lowest to the highest, and offer a slot to students ranked #1 through #40. Comparing the average outcomes of the children who attend the pre-K program (those ranked #1 through #40) with those who do not (those ranked #41 and above) can be problematic, for the reasons described above. Nevertheless, this approach can yield valid results. The first few students denied access (e.g., those ranked #41 through #45) come from families with incomes that are, on average, likely to be very similar to those who are barely eligible (e.g., students ranked #36 through #40). So, even though pre-K attendance is not randomly assigned in this case, as it would be in a lottery, the data are almost "as good as" randomly assigned data—at least among the subset of children from families with incomes that place them at or near the threshold for program eligibility.

In both the birthday and income cut-off examples, there are several important issues to note:

1. A key aspect of this design is to measure the same outcomes at the end of the pre-K year for both the children in the program and those not (yet) in the program. It is these outcomes that are then compared.
2. Because the children close to the cut-offs best fit the assumptions of random assignment (i.e., there are no differences between pre-K participants and non-participants other than the experience of pre-K), it is important to first restrict the analytic sample to this smaller group (e.g., children with birthdays between August and November in the case of a birthday cut-off) and then expand the sample to include children at greater distances from the cut-off. The smaller sample will be more defensible as meeting the assumptions of random assignment but will also be less representative of the broader population of young children.
3. Sometimes programs allow a few families to "break the rule" and thus enroll in the program when they are not really eligible and some eligible families will not participate even though they can. If these children are removed from the analytic sample, you will generate results that are considered to provide effects if the "treatment on the treated". If you include them, the results are considered to represent an "intention to treat" effect. There are pros and cons to both approaches and it is often a good idea to do both and consider the "true" effect somewhere in between.
4. This approach has an important limitation in that it is only capable of yielding estimates of impacts at the start of kindergarten: children who just completed pre-K are compared to children just entering pre-K, which means that one year later both groups will have received pre-K thus obviating pre-K vs. not-pre-K comparisons.

Alternative Control Group Approach

Thus far, we have considered design procedures at the local (e.g., school or school district) level. From an impact evaluation perspective, research designs with strong experimental or quasi-experimental components are ideal but not essential. At times, for example, school districts will not have or provide sufficient data to support an evaluation that includes a control group, as in the designs discussed above. In this case, the best course is to gather data across school districts (but this is only possible when districts use similar implementation approaches and collect similar data).

If a pre-K program is fully funded and does not have to turn children away, researchers can still tease out the program's impact by comparing pre-k participants in one district to non-recipients in another district. For instance, if a pre-K program is adopted in a subset of districts, then evaluators can compare the difference in outcomes between treated districts and other similar districts. Or, if a pre-K program is adopted in a few elementary schools in a large district, evaluators might compare outcomes to other, non-treated elementary schools in the same district. This approach – comparing the *differences* in outcomes between pre-K participants and non-participants in *different* districts (or cohorts) is referred to as a difference-in-difference analysis. To implement a successful difference-in-differences study, evaluators need comparable data on outcomes in the district or schools that offered pre-K before *and* after the program was introduced as well as in other, demographically similar areas that do not offer the program.

Difference-in-difference approaches can also be implemented after the fact by comparing the aggregate experience of students in one state to that in another comparable state or set of states. Several high-quality studies of universal pre-K programs in Oklahoma and Georgia have taken this approach. In Oklahoma, which adopted a statewide, universal pre-K program in 1998, evaluators used a variety of federally collected datasets to track participation rates and student outcomes. To account for other factors (such as general trends and changing demographics), researchers compared "before" and "after" outcomes in Oklahoma to those in other states. The challenge to this kind of approach is finding a control group that closely resembles the treatment group before the program is introduced. In the Oklahoma study, researchers were able to find these kinds of groups in southern states (other than Georgia, which adopted its own universal program in 1995).

For a good example of a random assignment study, see Tennessee pre-k study (Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013)

For a good example of a regression discontinuity design (RDD) study, see Tulsa pre-k study (Gormley, Gayer, Phillips, & Dawson, 2005)

For a good example of an alternative control group approach, see North Carolina pre-k study (Dodge, Bai, Ladd, and Muschkin, 2017) and Boston pre-k study (Weiland & Yoshikawa, 2013)

Sources: M.W. Lipsey, K.G. Hofer, N. Dong, D.C. Farran, and C. Bilbrey, "Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow-Up Results from the Randomized Control Design," (Nashville, TN: Vanderbilt University, Peabody Research Institute, 2013).

W.T. Gormley, T. Gayer, D.A. Phillips, and B. Dawson, "The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology* 41, no. 6 (2005): 872-884.

C. Weiland and H. Yoshikawa, "Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills," *Child Development* 84, no. 6 (2013): 2112-2130.

K.A. Dodge, Y. Bai, H.F. Ladd, and C.G. Muschkin, "Impact of North Carolina's Early Childhood Programs and Policies on Educational Outcomes in Elementary School," *Child Development* 88, no. 3 (2017): 996-1014.

IV. Which children and how many do you want to include in your evaluation?

— Weighting the Strength of Your Design —

After determining the research question and evaluation design, the next step is to develop a plan (a.k.a., a sampling strategy) to select the number and characteristics of children who will be studied. For example, if your question is about program implementation, you are probably only sampling children who are enrolled in the pre-K program. If you are asking a question about impacts, you will need to sample a mix of children who attended pre-K and children who did not. If you are trying to understand if one program model is more effective than another, you will want to compare children who attend pre-K programs that do one thing or have certain features to those who attend programs that do another thing or have different features. If you want to know whether impacts are stronger according to a specific participant characteristic or set of characteristics (e.g., household income. home language, special needs status, race and gender) or program features (e.g., instructional model or time spent on specific instructional content; part-time or full-time schedule; school-based classroom or other setting; number of years children spend in the program), you will need to recruit samples of sufficient size and representation, as well as select measurement tools that are suitable for all participants. For all of these analyses, you will need to identify a comparison group or counterfactual condition.

Power Analyses and Sampling

The first important consideration is sample size; the study needs a sample of participants that has the statistical power to detect effects. If the sample is too small, evaluators risk finding no effect—even when one is present.[10] There are several important drivers of statistical power to consider when determining sample sizes. For instance, the anticipated size of the effect can shape the sample size needed: identifying a smaller effect requires a larger sample size. Research suggests that one year of preschool generally has larger effects on early reading skills than on social-emotional skills. An evaluation of a pre-K's impact on early reading would thus require a smaller sample of children than an evaluation of impacts on social-emotional skills. Fortunately, several free and easy-to-use calculators can walk evaluators through sample size calculations.[11,12] Sample size aside, including a pre-test measure of the outcome (e.g., social-emotional ratings from right before pre-K program entry) can increase power and precision to detect significant effects.

Subgroup Sampling

If you do not have the resources to include all study children but are interested in studying the impact of the program on subgroups of students, evaluators can use an over-sampling procedure to draw a larger sample size. Say, for example, you want to study the effects of the

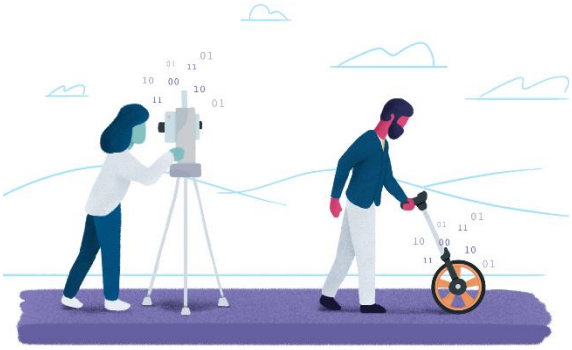
program on students with special needs, but these students comprise only 10 percent of the program. In this case, evaluators could sample *all* children in the program with special needs, while drawing a subsample of typically developing children. If resources are available to test for differences in outcomes between treatment and control groups in the full sample, sampling strategies that permit subgroup analyses are important to consider. Even if no impacts are detected in the full sample, impacts for certain subgroups may mask or offset overall effects. If impacts *are* observed in the full sample, they may vary in intensity, or disappear entirely, when broken down by subgroup. Thus, sampling strategies should consider the statistical power that subgroup analyses have to detect differences in experimental groups—both overall and by subdivisions of interest.

Sample Selection and the Counterfactual

As discussed in Section II, a pre-K program's impact depends on differences between children's experiences in the program and in other child care or early education programs they would have experienced if they were not enrolled in the pre-K program (i.e., the counterfactual condition). Sample selection plays an important role in allowing researchers to make valid comparisons to counterfactual conditions: to effectively measure whether program impacts differ depending on program characteristics (such as program location or program intensity), evaluators must determine whether significant differences exist between children who experience one type of pre-K program versus another. If you want to know whether children who attend pre-K in a public school have different outcomes than those who attend it in child care centers, evaluators must account for the possibility that the subset of kids who attend programs in school-based settings might differ in important ways from the subset who attend programs in other settings. Of course, under random assignment, this is not an issue as characteristics of children are evenly distributed across those in pre-K and those in comparison settings.

The recent literature on this process offers several different sample selection examples from which to draw. Studies of programs in Tulsa and Boston sampled all treatment and control group children whose families consented to participate.[13,14] Evaluators of Tennessee's voluntary statewide pre-K program followed the full population of study children via administrative records and selected an intensive subsample to assess through a battery of tests that aligned with program goals.[15]

V. What are the most important data to collect?



— Fitting Tools to Task —

To evaluate the program's effect on children, families, schools, and communities, it's important to think about the type, collection, and use of data prior to the pre-K program's implementation or expansion. Advance planning increases the likelihood that needed data will be available and may also save money. The pre-K program itself will generate a lot of data, such as information about program characteristics and the number of children served and their characteristics. These data are important for running the program (e.g., implementation studies or quality monitoring) but usually are not sufficient for evaluating program impacts. This is because such data are available for children in the program but not for comparison groups; even the data collected on program children may be insufficient for research purposes (see page 11). Most impact evaluation methodologies require data on control or comparison groups and seek to answer different research questions. Data collection efforts for these evaluations should consider and include sources of data that are available for children who are not in the program as well as those who are. We underscore that most existing pre-K evaluation studies have not collected intensive "features of the program" (e.g., classroom quality observations) data on both the program and comparison groups, which makes understanding "what works and why" difficult: in an ideal world, identical data would be gathered on both groups, to permit analyses of sources of impact variation.

Ideally, evaluators will collect data on measures that distinguish pre-K programs from alternatives, to illuminate what participating children actually experienced. Evaluators might consider measures such as the number, frequency, and length of coaching visits to a classroom and the characteristics of coaching (i.e., what it looks like). Both impact studies and evaluations of program quality or effectiveness of implementation require collecting measures that effectively capture this or other information about measures such as classroom quality or teacher and child attendance data – for program classrooms as well as comparison settings or classrooms. Evaluators can use these data to communicate descriptive statistics to teachers, coaches, directors, and stakeholders; monitor the quality of the program; and inform implementation efforts.

These same factors, used either as covariates (a variable such as family income or child gender that may – along with program features – predict the outcome) or outcomes in implementation studies, are also informative in an impact evaluation. But it is equally important to consider factors that may contribute to or help explain differences in program effects, like family income. In a state that offers universal pre-K programs serving both low- and middle-income children, it may be worthwhile to examine whether the program has the same impact on the subgroup of low-income children as it does on the middle-income children. In a study of a

program that enrolls a substantial number of dual language learners (i.e., children who are learning more than one language at a time), it may be important to examine whether dual-language learners and native speakers experience similar outcomes (while also taking into account issues around the language of assessment for outcome measures mentioned earlier). In a program with substantial variability in classroom quality or teacher qualifications, it might be important to examine differences in impact on the basis of these classroom characteristics. Depending on how the pre-K program is organized or on variability across the state programs, some factors may be of greater importance to consider than others. The particular subgroups of interest depend on the questions evaluators are asking.

Ultimately, evaluators need to make decisions about which measures to collect and when to collect them based on the research question, evaluation type, and study design. We proceed by identifying a comprehensive list of measures that capture sources of variability in pre-K programs that researchers should use at their discretion. These include: (a) participant outcomes grounded in a program's conceptual framework and guided by theory and research; (b) characteristics of pre-K teachers, programs, classrooms, eligibility, and exposure that will help you understand whether and why the program was effective; and (c) child and family characteristics that may serve as covariates and that affect the impact of participation on outcomes.

Outcome Measures

Choosing appropriate outcome measures is one of the most important and challenging aspects of conducting a well-designed impact study. Many different measures exist to assess skills, and sorting through them can be daunting. Some studies have reviewed measures at length and offer guidance.[16] Rather than recommend particular measures, we advise selecting measures based on a balanced set of principles.

First, choose measures that align with the child developmental domains targeted by a given program. If a preschool program is systematically spending little or no time on mathematics (which is quite typical),[17] measuring children's numeracy skills is likely not the best use of your resources—unless there is a strong developmental rationale or hypothesis to do so (or a desire to show that the program should consider focusing on that outcome). For instance, evaluators in a study of the impacts of a pre-K program in Boston included measures assessing pre-K impacts on executive function even though the program did not target these skills.[14] They did so, however, to test a hypothesis theorizing developmental links between mathematics education (which *was* targeted) and emerging executive function capacities.

Second, evaluators should include measures of developmentally important skills, particularly those that predict later outcomes and that are fundamental, malleable, and would not be gained in the absence of pre-K.[18] For instance, most children will acquire basic early reading skills such as letter-word recognition, whether they have had pre-K or not; in contrast, more complex language skills such as vocabulary are fundamental but may not be learned in the absence of pre-K.

We note also that much program administrative data generated as part of the pre-K program typically *does not* measure developmentally important skills that predict later outcomes (i.e., simple letter recognition versus developing a

rich vocabulary; counting versus problem solving), and thus evaluators should prioritize addition of these research-informed outcome measures.

Third, measures should show good reliability and validity. That is, good measures should produce consistent data regardless of the assessor, and they should assess the skill they purport to measure. The Peabody Picture Vocabulary Test, 4th Edition (PPVT IV), which measures receptive vocabulary, has excellent reliability and validity.[19]

Fourth, evaluators may want to prioritize selecting measures that have been widely used for two reasons. First, this approach will enable you to select measures that have been demonstrated to detect the effects of high-quality preschool. Many studies, for example, have found that the Woodcock-Johnson Letter-Word Identification subscale (an early reading measure) can detect children's gains in preschool.[13-15,20] Second, using the same measures across studies facilitates cross-study comparisons of results that can inform national discussions of pre-K. The PPVT has been used across many different recent pre-K evaluations.[14,20,21] Although assessments of social-emotional skills have been used less consistently across studies, there is now some agreement about the most important social-emotional constructs to capture when measuring these outcomes. For instance, behavior problems – those expressed as acting out and aggression, as well as those expressed as withdrawal and anxiety – have been raised as powerfully predictive of later negative social and academic outcomes and therefore may be worth prioritizing in the measures 'toolbox'. Using the same measures or measuring the same constructs across contexts places study results more cleanly within the broader landscape of what is already known about the impacts of different preschool programs.

Finally, consider available languages for a given assessment, and inclusion of a short screener to determine in which language multi-lingual children should be assessed. An English-language assessment of a child who does not speak English will measure English-language abilities but not necessarily the developmental domain of interest; assessing a non-English speaker's executive function in English will not accurately measure his or her executive function. Some assessments are available in both English and Spanish, such as the Woodcock-Johnson III and the Clinical Evaluation of Language Fundamentals, Version 5 (CELF-5). Sometimes non-equivalencies in the same test make it impossible to compare scores between English and Spanish versions. A Spanish-language version, for example, may have different content or rules about when to discontinue the assessment. Nevertheless, they do provide data that *are* informative and more culturally appropriate than English-only testing batteries.

Evaluators cannot give assessments in all languages, of course. Nor can you always find, train, and hire test assessors who speak all languages in a diverse district. The aim should be to meet the needs of the largest concentration of non-English speakers. One approach would be, as in the National Head Start Impact Study where the largest concentration of non-English speakers were Spanish speakers, to assess native English speakers in English, and Spanish-speaking children in Spanish, for instruments with English and Spanish versions at baseline. In that study, English-speaking children were also assessed on mathematics and early reading skills at baseline, but non-English speaking children

were not, because the instruments were not capable of yielding valid assessments among those children. At the end of the Head Start year, all children were assessed in English.

Assessment-selection principles *can* conflict. Evaluators of a preschool program that targets health may not be able to find many widely used measures. Health outcomes are not a common focus of many pre-K programs, and health is not typically included as an outcome in preschool evaluations. Nonetheless, keeping these principles in mind and balancing decisions across them can be useful in choosing what to measure and how to measure it.

Characteristics of Pre-K Programs, Classrooms, and Teachers

Effectively evaluating the effects of early childhood education programs requires a clear understanding of programmatic features. Was the program a full or half day? Did it run for the calendar year, the school year, or a portion thereof? Did it focus on early academic skills, or did it address broader skills? What did the program cost per child? What did children experience in the program? Documenting this information will allow you to determine whether program impacts varied by feature. This information also sheds important light on what is working well, what isn't working, and needed improvements. To this end, it is worth investing in good data systems to carefully document variations in early learning experiences, even when services are provided or funded through multiple sectors. Below we highlight classroom features that may inform program implementation or quality monitoring or mediate links between pre-K participation and child outcomes for impact evaluations.

Curricula. Pre-K programs vary by type (and presence) of curricula and the degree to which they are implemented as intended across programs and within classrooms. Types of curricula include teacher-created curricula, off-the-shelf "emergent" curricula (which follow students' interests within a broader framework), and child skill-specific curricula (which offer specific activities or emphases, such as mathematics, and follow a particular scope and sequence). Some curricula are easier to implement than others,[22] and some are more effective than others in improving targeted skills (see the What Works Clearinghouse). Research increasingly points to domain-specific, empirically tested curricula as being more promoting of early learning gains than generic, whole-child curricula.[1,2] Research also points to the importance of supports – like coaching – provided alongside a proven curriculum to amplify the value of the curriculum.

Documenting curricula in use (if any) and the degree to which they are well implemented is essential to ongoing quality assurance and process evaluations. These data (a) identify program and classroom strengths and areas for growth; (b) strengthen professional development activities, such as teacher coaching and training; and (c) document the conditions that contribute to impacts on children's skills. In short, program impacts are larger when implementation fidelity is higher.

Evaluators will likely face measurement challenges, however, in documenting whether curricula are implemented as intended. Some curricula come with pre-made checklists to help you determine whether core components are

implemented as intended. In general, though, these checklists are not psychometrically valid (i.e., they have not been *proven* to measure what they intend to measure) and may not identify all the key implementation components. There are multiple ways to measure curriculum fidelity;[23,24] the three most common measures relate to curriculum dosage, teacher adherence to core practices, and quality of delivery.[25] Off-the-shelf checklists may not include these measures, even though they are important elements of curriculum implementation.

Best practices in fidelity measurement offer a way forward. In quality assurance and impact studies, researchers must identify the curriculum's core practices and make sure that implementation measures and tools align with these practices.[26,27] At minimum, the tool should measure dosage, adherence, and quality. Measuring implementation well will entail at least some on-site visit by someone who is trained to use the tool and who knows the curriculum well. Several high-quality curriculum implementation studies offer guidance.[22,28]

Other dimensions of classroom quality. Measuring classroom and program quality provides important information that can explain why positive effects might occur and why variation in impacts might exist. It can also shed light on the level of quality within pre-K systems. Understanding classroom processes, meanwhile, can help inform discussions about what types or levels of quality are needed to yield higher outcomes.

Classroom and program quality vary markedly in the structural features that are regulated by the pre-K system, such as class size or teacher qualifications. Quality also varies by the classroom features that shape day-to-day experiences and learning environments, such as teacher-child interactions.[29] Thus, researchers should measure the multiple domains of classroom and program quality that may affect children's learning and experiences—even when they are studying a single program. Evaluators should pay particular attention to program-level and individual classroom quality; indeed, many quality features vary across classrooms in the same center or program.[30]

A growing body of evidence shows that children's learning gains from pre-K programs are often larger when programs provide cognitively stimulating experiences in an emotionally supportive environment and use developmentally appropriate practices.[31] Assessing this type of "process quality" often requires classroom observations, which can be expensive to conduct, particularly if multiple classrooms are observed in a given program or center (as recommended).[32] Nevertheless, observing children's direct experiences in classrooms can help explain why certain pre-K programs are more effective than others and potentially provide possible targets for teacher improvement. When assessing children's experiences in the classroom, domain-specific observational tools are preferred over global assessments of classroom quality. This is especially important if your questions have to do with the effectiveness of specific curricula or instructional strategies.

In addition, data collection efforts ideally include assessing program efforts to promote respectful interactions and cultural competence, such as using the child's home language in the classroom and training teachers in cultural sensitivity.[33,34] Also, programs' ability to support children with special needs or disabilities, such as providing specialized staff training, incorporating screening procedures, planning for and

accommodating children with special needs, and documenting plans and activities, can provide more information about quality.

Training and professional development. Measuring teacher training, professional development and other "information drivers" is important when evaluating process.[35] These drivers typically include general child development training, specific curricular training, and/or coaching by an expert mentor who may or may not be tied to the delivery of a specific curriculum. Regardless of the model used, specifying support(s) is key; this might mean providing details on the content and dosage of the training or supports and teacher ratings of their usefulness and effectiveness. Evaluators can collect these data via document review, surveys (of teachers, coaches, and trainers), observations (of training and coaching sessions), and/or qualitative interviews with a sample of teachers and coaches. These data identify programs' strengths and weaknesses.

Numerous studies suggest that ongoing quality assurance and teacher supports can ensure that classrooms offer children high-quality learning experiences.[14,36,37] Key elements of effective professional development and on-site technical assistance are: training staff on the key elements of classroom quality and/or training individuals in evidenced-based curriculum, and emphasizing the application of knowledge to practice.[38-40] Coaching models can also improve classroom quality and improve outcomes, particularly those in which expert coaches work with teachers to improve direct practices and provide constructive feedback using direct classroom observations.[41,42]

Professional development or coaching services differ by type, quantity, and quality. An evaluation study should carefully measure each component. This may include: the goals of training or coaching and content (e.g., curriculum, classroom environment), coaches' and trainers' qualifications, the number and frequency of visits or training sessions, and session duration and length (the number of hours per session and the number of weeks or months). Professional development and supports are offered by various sources, such as curriculum developers, individual pre-K programs, the broader pre-K system, or through state Quality Rating and Improvement Systems (QRIS). In addition, individual teachers may experience different types of supportive services within the same program. An evaluation study should attempt to capture the full set of professional services offered as well as differences in teacher utilization of these services.

Classroom composition. It is also important to document the composition of the children in the pre-K program and in classrooms within the program. For instance, as mentioned earlier some pre-K programs only serve low-income children whereas others are open to all children. This level of information yields important data about on-the-ground realities and can determine how to best allocate program resources. Take a program in which 10 percent of children are English Language Learners (ELLs): Different supports are needed if ELLs are distributed evenly across program sites and classrooms than would be if they are highly concentrated in a few sites and classrooms. In addition, some research suggests that children see higher gains in school readiness if their pre-K peers have stronger cognitive skills and/or come from higher socio-economic status backgrounds.[43-46]

Some programs have explicit mechanisms in place for integrating children with different skills and backgrounds; others don't. Either way, take care to document the profile of program participants overall as well as the classroom-level average and range of child and family characteristics. This will help you communicate who is served and guide internal decisions and quality improvement efforts.

Dosage and attendance data. On the child or family level, some children may attend pre-K programs inconsistently or for fewer days than the program offers due to family preference, instability or irregular routines, or because of constraints due to parental work schedules, the presence of other young children in the home, or transportation issues. Many elements of classroom and program quality have strong associations with children's learning, but these associations depend on how much the child actually attends the program which is sometimes referred to as dosage.[47,48] A child who participates in a pre-K program for five hours a week, for example, may not experience the intended impact, even if the program is of very high quality. As a result, evaluators should ideally track attendance data, days enrolled, chronic absenteeism, and dropout or termination rates as they all affect dosage.

Enhanced Covariates (Typically Child, Family, and Program Characteristics)

In addition to the usual covariates that can be obtained from administrative data sources, it is sometimes possible and often advisable to get rich, textured information on child and family characteristics from the most authoritative source possible: parents. There are two very good reasons for this: First, it may reduce bias when estimating program impact by controlling for child and family characteristics that might be linked to both the treatment and desired outcomes. And second, it may help evaluators discern whether program impacts are greater for some subgroups of children than for others.

Parents can add to the mix of covariates with information relating to the size, composition, and income of the household (adults and children); children's prior child care history; parents' employment and marital status, education levels, and countries of origin; the primary language spoken at home; and estimates of children's health and utilization of health care services (e.g., those provided by physicians, dentists, etc.). As with any survey, information about the respondent is important. Questions of particular interest might address the respondent's relationship to the child in the study and whether he or she is the child's primary caregiver.

Precise phrasing is also important. Remember that many people aren't familiar with details about our education system, such as differences between state-funded pre-K programs and those offered by Head Start or that take place in a day care center or a family home. As such, questions about child care history should be written in a way that is clear and easy to understand by the lay public. Some people, meanwhile, are not comfortable revealing information about income, so evaluators might consider questions that ask respondents to check a box that discloses an income range (e.g., between $20,000 and $30,000) rather than asking them to estimate or reveal precise amounts. Also, questions about use of health care services

should allow for specification of the type of service provided (e.g., wellness visit, an emergency visit, etc.).

Be sure to take proven steps to increase participation rates. Parents of young children are very busy, and their response rates, perhaps not surprisingly, can be disappointing. To boost participation rates, send survey forms home with children through "backpack mail." This increases the likelihood that parents will see the form when looking through other classroom materials. Ask them to return it either by backpack mail or "snail mail"—via a pre-addressed stamped envelope. Another strategy is to keep the survey short and offer a financial or other type of incentive to encourage participation.

Some incentives target individuals (e.g., small gift cards of $10 or $20 for completing and returning the survey within a specified period of time). Others target groups (e.g., pizza parties for classrooms or schools that meet specified response rates). Either will likely yield higher response rates. Last but not least, consider including Spanish-language versions of surveys in areas with large Hispanic populations.

When drafting surveys, consider questions about the following characteristics:

Child Level:	Parent and Family Level:
• Birthdate • Race or ethnicity • Home language • Gender • Individualized Education Program status and diagnosis type • Baseline assessments of target school readiness skills (e.g., language, literacy, numeracy, executive function, socio-emotional skills, etc.) • Prior experience in child care programs	• Level of education • Neighborhood of residence • Household income • Marital status • Biological father's place of residence (i.e., in or out of child's primary residence) • Immigration status and history (i.e., parents' and children's countries of origin and date of arrival in the United States) • Number of children in home • Number of books in the home • Primary home language

VI. How will you get the data?

— Collecting Reliable Data —

Who Will Collect Data?

Determining who will collect data, and which type of data they will collect, can be a complex decision. Different choices offer different tradeoffs, and tradeoffs vary by data type. Data collected by teachers is often less expensive than data collected by external staff. Young children may be more comfortable being assessed by a teacher than by someone they don't know. One potential concern with teacher-collected data: Teachers may not be able to provide unbiased assessments of children's progress or effectively rate the quality of instruction. Teachers are directly involved in the delivery of instruction and may have a vested interest in the results of the data. And, teachers' assessments of student skills can interrupt instructional time. A recent study found that teacher-collected data is not very useful because teachers almost universally rate themselves as strong curriculum implementers.[49] Ratings by coaches in the same study showed far more variability in teachers' level of curriculum implementation. Independent observations of classroom quality by outside observers tend to be more costly, but they can provide more objective assessments of classroom quality, teacher practices, and children's progress on learning outcomes.

In some cases, middle-ground options are possible. Teachers can collect baseline data that establish children's skill levels at the beginning of the program, and external observers can collect the end-of-program outcomes that will be used to determine impact. Alternatively, coaches can collect some data, though researchers must determine whether the coach has incentives to over- or under-score child outcomes or classroom processes. Some coaches may be less biased (intentionally or unintentionally) if the outcome of the data is not tied to their own performance review and if the data are portrayed and used as formative and if they provide data on teachers they don't coach.

Ensuring objective data collection is a priority, especially in impact evaluations. Determining whether a program is successful in affecting outcomes is a higher-stakes endeavor than a formative process study. A clear firewall between program administrators and teachers (or coaches) on the one side, and the impact study team on the other, will ensure that the study's conclusions are correct—and less subject to debate.

Another possible source of data is existing data – that is, administrative data already collected for other purposes. Some of those other sources/purposes are listed below.

State administrative data sources. States (as well as counties, cities, and other local government bodies) have vast amounts of data that may be relevant to evaluation of pre-K programs. Often referred to as administrative or management information system data, these records include information on

families who have participated in different social programs such as Temporary Assistance to Needy Families (TANF), the Supplemental Nutrition Assistance Program (SNAP, formerly called food stamps), health insurance programs such as Medicaid or the State Children's Health Insurance Program (SCHIP), and subsidies for child care, energy, and other needs. States also collect information from employers on employment and earnings, which covers most workers in the state, to inform the unemployment insurance system. Administrative data, particularly when linked across data sources, can provide substantial background information and prevent the need to collect it directly from families. As discussed at the end of this roadmap, concerns about data linking and data privacy must be addressed.

K-12 school records. School records are another important source of information: they can include child assessments as well as more ancillary data that provide important school-level contextual information such as staff orientation practices, staff retention rates, and frequency of staff meetings. Building an infrastructure to collect, store, and link early childhood data with K-12 child assessment (and other) data will provide opportunities to track children over time and evaluate longer-term outcomes. Planning how to gather, store, and link these data in advance is key; it may be difficult (or impossible) to reconstruct longitudinal data after the evaluation has begun. Many states are creating longitudinal databases for K-12, and some are linking data to records from early education, post-secondary education, and workforce systems.

Other early childhood data. Information from a Quality Rating and Improvement System (QRIS), Head Start programs, or from state child care assistance programs may be useful for assessing the types and quality of early childhood programs children participated in before or instead of the pre-K program.

Program cost data. Cost data is also important to consider. This includes information on all resources needed to implement and run the program, whether paid by the school district, parents, state or federal government funds, or other sources (Bartik, Gormley, & Adelstein, 2012 discuss this briefly as it relates to the Tulsa pre-K program).[50] Comprehensive cost information, which is needed for cost-effectiveness analysis, also includes in-kind donations such as donated space or equipment and volunteer time. (See Levin & McEwan, 2000 for technical details on cost effectiveness analysis).[51]

Reliability of Data Collection

Developing a clear plan to establish and maintain reliability for primary data collection will ensure the quality and rigor of your evaluation and will help produce stable and consistent results that can be compared to other evaluation studies. In addition, sound data collection procedures boost teachers' and administrators' confidence that they are being assessed fairly and ensure their continued buy-in.

Establishing and maintaining data reliability is expensive and should be calculated into time and budget estimates. Consider the following factors when assessing the quality of data collection efforts. First, a study should clearly define the qualifications of data collectors. This includes the collectors' level of education and

their experience working with early childhood education programs and research projects. Second, the study needs a clear plan to ensure reliability based on the measures selected. This is particularly important for classroom observations, which often require standardized, but subjective, coding of the classroom environment. Components to consider are: training schedules (i.e., how much time will elapse between the training and the beginning of data collection efforts?); duration (i.e., how long will the training last?), and trainer identity (who will lead the training?). Evaluators should also determine how to establish reliability for each measure. After data collectors are trained, evaluators should check in periodically to ensure quality assurance. This will give collectors the opportunity to practice coding and give feedback on assessments, as well as to provide checks on reliability to avoid drift over time. In addition, data collection should take place at consistent times of the day and document factors that may affect data quality, such as fire drills or the presence of substitute teachers.[52,53]

Frequency of Data Collection

Implementation data. Implementation data are often collected via classroom observations or teacher surveys, both of which can burden teachers or other staff in the classroom. As such, these data are typically not collected as often as evaluators would like. Ideally, evaluators will collect quality classroom observation data at the beginning and end of each school year. In practice, however, this type of high-quality data collection is less frequently carried out. Many Quality Rating and Improvement Systems (QRIS), for example, only collect quality data every three years. Also, note that teachers are already over-burdened with data collection and other responsibilities, so it may be helpful to embed pre-K data collection into larger, ongoing data collection efforts.

Outcome data. Evaluators must also decide how often to collect outcome data. In many pre-K evaluations, outcome data are collected at or near the end of the pre-k year; studies show it takes about this much time for many children to realize preschool's positive benefits. More broadly, decisions about timing should consider the program's theory of change, especially if the theory anticipates when particular effects will manifest, based on findings from past research and child development.

Tracking Students over Time

To understand whether and why pre-K programs have lasting impacts beyond kindergarten, you will want to track students over time: this is a longitudinal study. The hardest part of a longitudinal impact study is simply keeping track of your students. Some will continue their education in the same public school district. Others will move to another school district or state. And still others will move to private or charter schools. Some may eventually attend school on a military base and others at home. The fact is: students are highly mobile!

So, think of the search for elusive students as an iterative process. Start with administrative records: at the state level, these should tell you where a child moved if they were given a state ID for pre-K that is unique and maintained through K-12 education. Begin with the district or county in which a child attended school (or pre-K) the previous year, and then search nearby counties. Keep in touch with families to get periodic address updates and to ask parents about children's future school plans. The state

department of education has a wealth of information. The advantage of this data is its breadth; in theory, they know the whereabouts of every public school student in every public school in the state. Other state officials can also be helpful (e.g., those who administer the state child care subsidy program; officials in the state department of health and mental health). You might even seek a letter of support from the governor or another high-ranking official.

Much of your digging will likely require the consent and active cooperation of local school districts. The advantage of this data is its depth. Local districts have information on standardized test scores, grades, courses and coursework, attendance records, disciplinary records, and so forth. Keep in mind that a local school superintendent's consent is needed to access confidential data. With a higher-ranking official's support, many mid-level managers will field your requests. Don't take such cooperation for granted, though. The superintendent's official support opens the door to people who have needed data, but it does not guarantee that your needs will be prioritized. Also, you may still need to follow district norms and practices, such as the submission of lengthy research request forms. Be patient but also persistent. If foot-dragging is a problem, ask the superintendent or other officials to intervene.

Figuring out who is who can be surprisingly difficult (unless there are unique student IDs used). Students may change their names as a result of divorce, adoption, or other circumstances, and typos and misspellings occur. You may need to do some detective work to determine whether Student Y and Student X are one and the same. What if they have the same name but different dates of birth? What if they have the same date of birth but different names? To answer these types of questions, develop an algorithm or use one that is already available. For example, STATA (a data analysis resource) offers a "reclink" program that matches students by three data points and produces an estimated probability that the students are the same. You may have to add additional descriptors into the mix. Inevitably, though, judgment calls are unavoidable.

Capacity to Gather Data from Participants and Comparison Children

Process and implementation data. Ideally, all process and implementation data will capture the experiences of both children in the pre-K program and those in the comparison group (assuming a research design that has a comparison group). Data on the experiences of both the treatment and comparison group are crucial to identifying the true treatment contrast—e.g., the difference in what the treatment group received versus the counterfactual.[54] Several recent studies have highlighted that preschool program impacts differ greatly depending on the experiences of the children in the comparison group. For example, effects tend to be considerably larger if the comparison group experienced care provided by a parent than by another center-based preschool program.[7,8,55]

Practically and logistically, comparable data on the control group can be difficult and expensive to obtain. Collecting data on curriculum fidelity and instructional quality experienced by control-group children, for example, entails tracking them into their alternative settings and obtaining permission to collect program- and classroom-level data. Using the same setting-level measures may not be possible in all treatment and comparison group settings, especially if comparison group children are at home or in

family child care (FCC) settings. Families of comparison group children and their program directors and teachers also likely will not feel as connected to the research project and are more likely to refuse to participate in data collection efforts.

At minimum (and for relatively little cost), evaluators can ask parents of comparison group children about their children's primary care setting. The impact evaluation of Boston's pre-K program used this type of data and was able to determine that about two-thirds of comparison group children were enrolled in other non-parental care settings.[14] If more funds are available, the same kinds of process and setting-level data should be collected in both treatment and comparison settings.

The National Head Start Impact Study, for example, tracked and documented in detail the experiences of comparison-group children whenever possible.[20] In addition to teacher surveys, the study team administered observational quality measures of settings experienced by treatment and control groups. For those in center-based preschools, the study team used the Early Childhood Environment Rating Scale-Revised (ECERS-R). For study children enrolled in family child care settings, they used the Family Day Care Rating Scale (FDCRS), which is similar but not identical to the ECERS-R. For children in parent care, no comparable quality measures were available. Despite these limitations, the available data on the comparison group's experiences were extremely helpful in unpacking the study's results. These data also subsequently made possible sophisticated analytic approaches to understanding Head Start's effects on study children.[56]

Outcome measures. Ideally, comparable evaluation data should be collected on both the treatment and control groups and on the primary care setting each child attends. Operationalizing outcome measures in the same way in both groups is crucial to drawing valid inferences.[9] Data on the experiences of both the treatment and comparison groups are crucial to identifying the true treatment contrast—e.g., the difference in what the treatment group received versus the counterfactual.[54] Treatment impacts (or lack thereof) can then be interpreted considering what the actual differences in experiences were between the treatment and control groups.

Logistically, as explained in the *Process Evaluation* section, comparable data on the control group can be difficult and costly to obtain. It can require extensive (and expensive) tracking of children into many different alternative care settings. Measurement limitations are also at play. For children in parent care, no measures of the home environment are directly equivalent to measures of preschool classroom quality. Nonetheless, to the extent possible, evaluation teams should budget for tracking the experiences of children in the control group. Again, the Head Start Impact Study offers one example of how to do so (see full example in the *Process Evaluation* section).

VII. What else should you consider before you start?

— Seeing Your Work in Context —

There are a few other things to consider when planning, launching, and seeing an evaluation effort to completion. The first bucket of additional items may be thought of as infrastructure: the feasibility and quality of an evaluation depends on setting up the necessary infrastructure, which could include systems for data linking, protecting private data, and sharing data. We discuss those items below, but urge you to also think about things like ensuring adequate staff and clarifying roles, setting up advisory committees and review processes, and planning for the production of reports and dissemination. Next, we briefly highlight the role of research-practice partnerships, which can be invaluable when designing and conducting a pre-K evaluation as well interpreting findings from the evaluation efforts. Finally, we touch on stakeholders.

Infrastructure

Data linking. As mentioned above, a unique student identifier (where each child has a unique number that is the same in each data system) is ideal as it facilitates tracking students longitudinally after they leave the pre-K program and if they leave the school district. However, one of the biggest challenges of creating longitudinal datasets is the ability to link data across different systems. To find family information, linking to government agency and workforce information systems may require unique family or parent identifiers, which must also link to the child data. Social security numbers (SSNs) are unique identifiers for most parents and children; however, the federal government encourages the use of alternative identifiers rather than SSNs to reduce the risk of identity theft. Keeping and encrypting SSNs in the database as a secondary identifier is an acceptable practice (see Social Security Administration regional guidelines for protecting SSNs).

When evaluators lack access to a unique identification number for each child, statistical methods can match children across data systems in order to link data. These approaches typically use a combination of name, gender, birthdate, and other information to match children on a probabilistic basis.

Data privacy. Educators are generally very familiar with data privacy regulations for student records (see the Family Educational Rights and Privacy Act, FERPA). Other data, if linked with student records, may be subject to additional state or federal data privacy rules. As part of the evaluation planning process, consider how the linked data will be stored, who will have access, and how confidentiality will be protected. All users of the data should be aware of and follow data privacy restrictions. Local research partners, particularly if they are new to the education research field, may not be aware of data privacy rules. A data sharing agreement, specifying roles and responsibilities with regards to

data privacy, may be especially helpful in this case.

Data sharing agreements. Government agencies, school districts and other organizations are likely to have policies in place that require signed, written data-sharing agreements when sharing data with other organizations or researchers. These agreements typically describe the purpose of the agreement, including the research and evaluation objectives, detail the data to be shared, and specify the roles and responsibilities of the parties. These agreements can also include specific privacy safeguards and any restrictions on use or sharing of the data. More broadly, a long-term strategy for the governance and support of longitudinal data systems should be considered.

Research Partnerships

When planning pre-K evaluation studies, it is helpful to form partnerships with local researchers; this will inform the overall evaluation plan and support data collection. Local researchers may be able to provide valuable additional perspectives on the work and can help you decide which children to include, what types of data can address the research goals, and when new data should be collected (beyond any existing data that might be used). Research partners can also provide guidance in the selection of appropriate measures. And, researchers can often bring capacity to apply for research grants that would bring in external funds to evaluate programs with rigor.

Universities and community colleges are also key partners, as they may have the capacity to provide local research support for pre-K evaluations. Most universities and community colleges employ faculty with expertise in training college students to teach or study children. These faculty or staff members can be found in departments administering teacher preparation programs for early childhood, elementary, or special education. Additional experts can be found in other types of departments with interests in young children, such as psychology, human or child development, and public policy. In addition to their expertise, these individuals also will be working with undergraduate and graduate students who have interests in the same areas and who can provide the basis for a team of data collectors. Such students may be interested in conducting research as part of their degree programs and may appreciate opportunities for extra income. Universities also have the capacity to manage tasks such as maintaining compliance with review boards that govern research regulations and managing payroll for data collectors. Some local nonprofit organizations with interests and expertise in early childhood may also have the capacity to manage or assist with research tasks in ways similar to universities or community colleges. For more information on research-practice partnerships, see Coburn et al., 2013.[57]

Stakeholders

Pre-K programs and their evaluations affect many stakeholders, including families, teachers and support staff, principals, superintendents, and other education policymakers. Setting up a strategy for keeping your stakeholders informed about the planning, progress, and outcome of your evaluation efforts is important: such a strategy might involve convening regular town-hall style meetings at community centers, libraries, schools, etc.; posting short summaries or briefs on websites; creating newsletters to be distributed electronically.

Informing these stakeholder groups about your evaluation at the beginning of the process, and keeping them in the loop as the evaluation proceeds and begins to produce evidence, is not only best practice for strong community relations but will greatly enhance the chances that your evidence will be used for program improvement efforts. And that is, after all, the goal of providing a strong roadmap for your evaluation effort.

VIII. Conclusion

It is our hope that after reading this roadmap, you feel more prepared to continue the critical work of building this educational highway system of which pre-K is a core element. State-funded pre-K programs have been the focus of nearly two decades of evaluation research – research that has produced a large body of evidence on the immediate impacts of pre-K programs on children's school achievement and pointed to some good bets about the inputs that produce these impacts. But there is more work to be done: with this roadmap in hand, evaluators can share and compare knowledge so that, together, we can support school systems across the country. As they are better able to identify the factors that distinguish effective programs from less effective ones and take constructive action to better meet our country's educational and workforce goals, we can truly build a better tomorrow.

[1] D.A. Phillips, M.W. Lipsey, K.A. Dodge, R. Haskins, D. Bassok, M.R. Burchinal, G.J. Duncan, M. Dynarski, K.A. Magnuson, and C. Weiland, "Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effect" (Washington, DC: The Brookings Institute, 2017).

[2] C. Weiland, "Pivoting to the 'How': Moving Preschool Policy, Practice, and Research Forward," *Early Childhood Research Quarterly,* In Press (2018).

[3] W.S. Barnett, M.E. Carolan, J.H. Squires, and K. Clarke Brown, "The State of Preschool 2013: State Preschool Yearbook" (New Brunswick, NJ: National Institute for Early Education Research, 2013).

[4] J.T. Hustedt and W.S. Barnett, "Early Childhood Education and Care Provision: Issues of Access and Program Quality," in *International Encyclopedia of Education: Vol. 2 (3rd edition)*, eds. P. Peterson, E. Baker, & B. McGaw, (Oxford, UK Elsevier, 2010), 110-119.

[5] W.T. Gormley, "Universal vs. Targeted Pre-Kindergarten: Reflections for Policymakers," in *Current State of Knowledge on Pre-Kindergarten Effects,* (Washington, DC: The Brookings Institute, 2017), 51-56.

[6] U.S. Department of Health and Human Services, Administration for Children and Families. "Head Start Program Performance Standards 45 CFR Chapter XIII," (Washington, DC: U.S. Department of Health and Human Services: Administration for Children and Families, 2009).

[7] A. Feller, T. Grindal, L. Miratrix, and L.C. Page, "Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care Type," *The Annals of Applied Statistics* 10, no. 3 (2016): 1245-1285.

[8] F. Zhai, J. Brooks-Gunn, and J. Waldfogel, "Head Start and Urban Children's School Readiness: A Birth Cohort Study in 18 Cities," *Developmental Psychology* 47, no. 1 (2011): 134-152.

[9] M.W. Lipsey, C. Weiland, H. Yoshikawa, S.J. Wilson, and K.G. Hofer, "The Prekindergarten Age-Cutoff Regression-Discontinuity Design: Methodological Issues and Implications for Application," *Educational Evaluation and Policy Analysis* 37, no. 3 (2014): 296-313.

[10] R.J. Murnane and J.B. Willett, *Methods Matter: Improving Causal Inference in Educational and Social Science Research,* (Oxford, UK: Oxford University Press, 2010).

[11] N. Dong and R.A. Maynard, "PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Sample Size Requirements for Experimental and Quasi-Experimental Designs," *Journal of Research on Educational Effectiveness* 6, no. 1 (2013): 24-67.

[12] J. Spybrook, S.W. Raudenbush, X.F. Liu, R. Congdon, and A. Martínez, "Optimal Design for Longitudinal and Multilevel Research: Documentation for the 'Optimal Design' Software," (Anne Arbor, MI: Survey Research Center of the Institute of Social Research at the University of Michigan, 2006).

[13] W.T. Gormley, T. Gayer, D.A. Phillips, and B. Dawson, "The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology* 41, no. 6 (2005): 872-884.

[14] C. Weiland and H. Yoshikawa, "Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills," *Child Development* 84, no. 6 (2013): 2112-2130.

[15] M.W. Lipsey, K.G. Hofer, N. Dong, D.C. Farran, and C. Bilbrey, "Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow-Up Results from the Randomized Control Design," (Nashville, TN: Vanderbilt University, Peabody Research Institute, 2013).

[16] T. Halle, M. Zaslow, J. Wessel, S. Moodie, and K. Darling-Churchill, "Understanding and Choosing Assessments and Developmental Screeners for Young Children: Profiles of Selected Measures," (Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2011).

[17] H.P. Ginsburg, J.S. Lee, and J.S. Boyd, "Mathematics Education for Young Children: What it is and How to Promote it," *Social Policy Report* 22, no.1 (2008): 3-22.

[18] D. Bailey, G.J. Duncan, C.L. Odgers, and W. Yu, "Persistence and Fadeout in the Impacts of Child and Adolescent Interventions," *Journal of Research on Educational Effectiveness* 10, no. 1 (2017): 7-39.

[19] L.M. Dunn and D.M. Dunn, "The Peabody Picture Vocabulty Test, Fourth Edition," (Bloomington, MN: NCS Pearson, Inc., 2007).

[20] M. Puma, S. Bell, R. Cook, and C. Heid, "Head Start Impact Study: Final report," (Washington, DC: U.S. Department of Health and Human Services, 2010).

[21] V.C. Wong, T.D. Cook, W.S. Barnett, and K. Jung, "An Effectiveness-Based Evaluation of Five State Pre-Kindergarten Programs," *Journal of Policy Analysis and Management* 27, no. 1 (2008): 122-154.

[22] P. Morris, S.K. Mattera, N. Castells, M. Bangser, K. Bierman, and C. Raver, "Impact Findings from the Head Start CARES Demonstration: National Evaluation of Three Approaches to Improving Preschoolers' Social and Emotional Competence. Executive Summary," (Washington, DC: Office of Planning, Research and Evaluation, 2014).

[23] A.V. Dane and B.H. Schneider, "Program Integrity in Primary and Early Secondary Prevention: Are Implementation Effects Out of Control?" *Clinical Psychology Review* 18, no. 1 (1998): 23-45.

[24] J.A. Durlak, "The Importance of Doing Well in Whatever You Do: A Commentary on the Special Section 'Implementation Research in Early Childhood Education'," *Early Childhood Research Quarterly* 25 (2010): 348-357.

[25] C.L. Darrow, "Measuring Fidelity in Preschool Interventions: A Microanalysis of Fidelity Instruments Used in Curriculum Interventions" (Washington, DC: Paper presented at the Conference of the *Society for Research on Educational Effectiveness*, 2010).

[26] C.S. Hulleman and D.S. Cordray, "Moving from the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength," *Journal of Research on Educational Effectiveness* 2, no. 1 (2009): 88-110.

[27] M.C. Nelson, D.S. Cordray, C.S. Hulleman, C.L. Darrow, E.C. Sommer, "A Procedure for Assessing Intervention Fidelity in Experiments Testing Educational and Behavioral Interventions," *The Journal of Behavioral Health Services and Research* 39, no.4 (2012): 374-396.

[28] S.J. Wilson and D.C. Farran, "Experimental Evaluation of the Tools of the Mind Curriculum" (Washington, DC: Paper presented at the Conference of the *Society for Research on Educational Effectiveness*, 2012).

[29] Pianta, R.C., W.S. Barnett, M. Burchinal, and K.R. Thornburg, "The Effects of Preschool Education: What We Know, How Public Policy Is or Is Not Aligned with the Evidence Base, and What We Need to Know," *Psychological Science in the Public Interest* 10, no. 2 (2009): 49-88.

[30] L. Karoly, G.L. Zellman, and M. Perlman, "Understanding Variation in Classroom Quality Within Early Childhood Centers: Evidence from Colorado's Quality Rating and Improvement System," *Early Childhood Research Quarterly* 28, no. 4 (2013): 645-657.

[31] G. Camilli, S. Vargas, S. Ryan, and W.S. Barnett, "Meta-Analysis of the Effects of Early Education Interventions on Cognitive and Social Development," *The Teachers College Record* 122, no. 3 (2010): Article 15440.

[32] S.W. Raudenbush and X.F. Liu, "Effects of Study Duration, Frequency of Observation, and Sample Size on Power in Studies of Group Differences in Polynomial Change," *Psychological Methods* 6, no.4 (2001): 387-401.

[33] K. Magnuson, C. Lahaie, and J. Waldfogel, "Preschool and School Readiness of Children of Immigrants," *Social Science Quarterly* 87, no. 5 (2006): 1241-1262.

[34] K. Tout, R. Starr, M. Soli, S. Moodie, G. Kirby, and K. Boller, "The Child Care Quality Ratin System (QRS) Assessment: Compendium of Quality Rating Systems and Evaluations" (Washington, DC: Child Trends and Mathematica Policy Research, 2010).

[35] D.L. Fixsen, K.A. Blasé, S. Naoom, and F. Wallace, "Core Implementation Components," *Research on Social Work Practice* 19, no. 5 (2009): 531-540.

[36] K.L. Bierman, C.E. Domitrovich, R.L. Nix, S.D. Gest, J.A. Welsh, M.T. Greenberg, C. Blair, K.E. Nelson, and S. Gill, "Promoting Academic and Social-Emotional School Readiness: The Head Start REDI Program," *Child Development* 79, no. 6 (2008): 1802-1817.

[37] D.H. Clements and J. Samara, "Experimental Evaluation of the Effects of a Research-Based Preschool Mathematics Curriculum," *American Educational Research Journal* 45, no. 2 (2008): 443-494.

[38] V. Buysse, P.J. Winton, B. Rous, "Reaching Consensus on a Definition of Professional Develpoment for the Early Childhood Field," *Topics in Early Childhood, Special Education* 28, no. 4 (2009): 235-243.

[39] B. Hamre, "Teachers' Daily Interactions with Children: An Essential Ingredient in Effective Early Childhood Programs," *Child Development Perspectives* 8, no. 4 (2014): 223-230.

[40] S.M. Sheridan, C.P Edwards, C.A. Marvin, and L.L. Knoche, "Professional Development in Early Childhood Programs: Process Issues and Research Needs," *Early Education and Development* 20, no. 3 (2009): 377-401.

[41] R.C. Pianta, M. Burchinal, F.M. Jamil, T. Sabol, K. Grimm, B. Hamre, J. Downer, J. LoCasale-Crouch, and C. Howes, "A Cross-Lag Analysis of Longitudinal Associations Between Preschool Teachers' Instructional Support Identification Skills and Observed Behavior," *Early Childhood Research Quarterly* 29, no. 2 (2014): 144-154.

[42] C.C. Raver, S.T. Jones, C. Li-Grining, F. Zhai, M. Metzger, and B. Solomon, "Targeting Children's Behavior Problems in Preschool Classrooms: A Cluster-Randomized Controlled Trial," *Journal of Consulting and Clinical Psychology* 77, no. 2 (2009): 302-316.

[43] G.T. Henry and D.K. Rickman, "Do Peers Influence Children's Skill Development in Preschool," *Economics of Education Review* 26, no. 1 (2007): 100-112.

[44] A.J. Mashburn, L.M. Justice, J.T. Downer, and R.C. Pianta, "Peer Effects on Children's Language Achievement During Pre-Kindergarten," *Child Development* 80, no. 3 (2009): 686-702.

[45] J.L. Reid and D.D. Ready, "High-Quality Preschool: The Socioeconomic Composition of Preschool Classrooms and Children's Learning," *Early Education and Development* 24, no. 8 (2013): 1082-1111.

[46] C. Weiland and H. Yoshikawa, "Does Higher Peer Socio-Economic Status Predict Children's Language and Executive Function Skills Gains in Prekindergarten?" *Journal of Applied Developmental Psychology* 35, no. 5 (2014): 422-432.

[47] J.L. Hill, J. Brooks-Gunn, and J. Waldfogel, "Sustained Effects of High Participation in an Early Intervention for Low-Birth-Weight Premature Infants," *Developmental Psychology* 39, no. 4 (2003): 730-744.

[48] M. Zaslow, R. Anderson, Z. Redd, J. Wessel, L. Tarullo, and M. Burchinal, "Quality Dosage, Thresholds, and Features in Early Childhood Settings: A Review of the Literature, OPRE 2011-5" (Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation, 2010).

[49] C.E. Domitrovich, S.D. Gest, D. Jones, S. Gill, and R.M. Sanford Derousie, "Implementation Quality: Lessons Learned in the Context of the Head Start REDI Trial," *Early Childhood Research Quarterly* 25, no. 3 (2010): 284-298.

[50] T. Bartik, W.T. Gormley, and S. Adelstein, "Earnings Benefits of Tulsa's Pre-K Program for Different Income Groups," *Economics of Education Review* 31, no. 6 (2012): 1143-1161.

[51] H.M. Levin and P.J. McEwan, "Cost-Effectiveness Analysis: Methods and Applications (2nd ed.)" (Thousand Oaks, CA: Sage, 2001).

[52] T.W. Curby, M. Stuhlman, K. Grimm, A.J. Mashburn, L. Chomat-Mooney, J. Downer, … and R.C. Pianta, "Within-dat variability in the quality of classrooms interacrtions during third and fifth grade," *The Elementary School Journal* 112, no. 1 (2011): 16-37.

[53] J.P. Meyer, A.E Henry, and A.J. Mashburn, "Occasions and the Reliability of Teaching Observations: Alternative Conceptualizations and Methods of Analysis," *Educaitonal Assessment Journal* 16, no. 4 (2011): 227-243.

[54] M.J. Weiss, H.S. Bloom, and T. Brock, "A Conceptual Framework for Studying the Sources of Variation in Program Effects," *Journal of Policy Analysis and Management* 33, no. 3 (2014): 778-808.

[55] F. Zhai, J. Brooks-Gunn, and J. Waldfogel, "Head Start's Impact is Contingent on Alternative Type of Care in Comparison Group," *Developmental Psychology* 50, no. 12 (2014): 2572-2586.

[56] A.H. Friedman-Krauss, M.C. Connors, and P.A. Morris, "Unpacking the Treatment Contrast in the Head Start Impact Study: To What Extent Does Assignment to Treatment Affect Quality of Care?" *Journal of Research on Educational Effectiveness* 10, no.1 (2017): 68-95.

[57] C.E. Colburn, W.R. Penuel, and K.E. Geil, "Research-Practice Partnerships: A Strategy for Leveraging Research for Educational Improvement in School Districts" (New York, NY: William T. Grant Foundation, 2013).

Appendix

Names and Affiliations of Contributors to the Pre-K Roadmap

Daphna Bassok, Associate Professor of Education and Public Policy; Associate Director of EdPolicyWorks, University of Virginia

Janet Bock-Hager, Coordinator, Office of Early Learning, West Virginia Department of Education

Kathleen Bruck, CEO, Pre-K 4 SA (Retired), San Antonio, Texas

Kimberly Burgess, Early Childhood Policy Team Lead, Child and Youth Policy, U.S. Department of Health and Human Services

Tim Burgess, Mayor and Chair of City Council (Retired), City of Seattle, Washington

Elizabeth Cascio, Associate Professor of Economics, Dartmouth College

Ajay Chaudry, Senior Fellow & Visiting Scholar, New York University

Liz Davis, Professor of Applied Economics, University of Minnesota

Cindy Decker, Director of Research and Innovation, CAP Tulsa

Lauren Decker-Woodrow, Senior Study Director, Westat

David Deming, Professor of Public Policy and Professor of Education and Economics, Harvard University

Libby Doggett, Early Learning Expert and Consultant, Libby Doggett Consulting

Janis Dubno, Director, The Sorenson Impact Center

Chloe Gibbs, Assistant Professor of Economics, University of Notre Dame

William Gormley, University Professor of Public Policy and Co-Director of the Center for Research on Children in the United States, Georgetown University

Carolyn Hill, Senior Fellow, MDRC

Jason Hustedt, Associate Professor of Human Development and Research Director, Delaware Institute for Excellence in Early Childhood, University of Delaware

Erica Johnson, Manager, Early Learning Policy & Innovation, City of Seattle Department of Education and Early Learning

Danielle Kassow, Independent Consultant, Danielle Z. Kassow Consulting

Mark Lipsey, Research Professor of Human and Organizational Development and Director of the Peabody Institute, Vanderbilt University

Joan McLaughlin, Commissioner, National Center for Special Education Research, IES

Pamela Morris, Professor of Applied Psychology and Social Intervention, New York University

Ellen Peisner-Feinberg, Senior Research Scientist at the Frank Porter Graham Child Development Institute, University of North Carolina at Chapel Hill

Terri Sabol, Assistant Professor of Human Development and Social Policy, Northwestern University

Jason Sachs, Executive Director of Early Childhood Education, Boston Public Schools

Diane Schanzenbach, Professor of Human Development and Social Policy; Director of the Institute for Policy Research, Northwestern University

Gerard 'Sid' Sidorowicz, Deputy Director, City of Seattle Department of Education and Early Learning

Aaron Sojourner, Associate Professor, Department of Work and Organizations, University of Minnesota

Albert Wat, Senior Policy Director, Alliance for Early Success

Christina Weiland, Assistant Professor of Education, University of Michigan

This report was made possible by the generous financial support provided to Brookings by the Heising-Simons Foundation and the David and Lucile Packard Foundation. Further support came from the Reflective Engagement in the Public Interest grant gifted by Georgetown University.

The authors would like to extend a special thanks to Christina Weiland for her thoughtful review of this report.

BROOKINGS