Does Pre-K Work?

The Research on Ten Early Childhood Programs—And What It Tells Us

KATHARINE B. STEVENS AND ELIZABETH ENGLISH

APRIL 2016

AMERICAN ENTERPRISE INSTITUTE

Does Pre-K Work?

The Research on Ten Early Childhood Programs—And What It Tells Us

KATHARINE B. STEVENS
AND ELIZABETH ENGLISH

APRIL 2016

AMERICAN ENTERPRISE INSTITUTE

Contents

Executive Summary

With growing public and political support, the early childhood field is advancing quickly, now focused primarily on expanding school-based pre-K. Yet pre-K is just one part of a broad landscape of programs for children from birth through age four, and the emphasis on pre-K often overshadows other valuable approaches, such as child care and two-generation initiatives that work with children and parents together. Neither the public nor policymakers have a clear picture of the range of early childhood programs, the varied evidence on their effectiveness, and how that evidence can guide us going forward.

This report aims to provide a starting point for a more comprehensive, nuanced dialogue around core policy goals in early childhood and the best strategies to accomplish those goals. It examines 10 of the best-known, widely cited programs of the last half century—Abbott Preschool, Abecedarian, Boston Pre-K, Chicago Child-Parent Centers, Georgia Pre-K, Head Start, Nurse-Family Partnership, Oklahoma Pre-K, Perry Preschool, and Tennessee Voluntary Pre-K—and the research on those programs. The report has two parts.

Part I is a short guide to the four research methods most commonly used to evaluate early childhood programs. While research findings are often presented in policy debates as black and white, they have a lot more gray than is often acknowledged. A basic understanding of how studies are conducted is essential to correctly interpreting their results. This brief overview aims to help nonexperts understand the methods used in early childhood research, how the choice of methods can influence study results, and the limitations of each method.

Part II describes the 10 programs, answering several broad questions about each: What is the specific nature of the program? Whom does it serve, and how is it designed? What kind of research has been conducted on it? What methods were used, and what results were found? What are the key takeaways?

A close look at these 10 programs reveals that they are as different as they are similar. Some focused on four-year-olds, some on three-year-olds, and some solely on infants and toddlers. Some programs ran for just one year, others for two, and one served children from infancy to kindergarten. Some were school-based while others were home-based. Some targeted children alone while some targeted their families too. Some programs increased the number of alphabet letters children knew when they were five; others led to large increases in social, economic, and health outcomes decades later.

The research conducted on the 10 programs also varied greatly. Researchers used different methods to investigate a range of questions: some evaluated basic academic skills in kindergarten, some examined children's performance in elementary school, and still others tracked a range of long-term social and economic effects into adulthood. Some studies were more rigorous than others.

The research shows neither that "pre-K works" nor that it does not; rather, it shows that some early childhood programs yield particular outcomes, sometimes, for some children. Overall, our report finds that this body of research provides less useful information than is commonly assumed. It shows that early childhood programs *can* have a significant, sustained impact on the lives of children born into disadvantaged circumstances, but falls far short of showing that all programs have that impact. The most rigorous research shows that the most meaningful, far-reaching effects occurred with intensive, carefully designed,

well-implemented programs—specifically Abecedarian, Nurse-Family Partnership, and Perry—that target very young children, engage parents, and teach a broad range of skills.

Two important policy implications emerge. To move the early childhood field forward, we must:

- **Strengthen and accelerate rigorous research in early childhood.** The early childhood research base is often characterized as rigorous and extensive, and it indeed includes hundreds of studies published over the last several years. Yet both the relevance and rigor of this research is considerably weaker than many realize. A stronger knowledge base is urgently needed to guide policy.

  While current research focuses overwhelmingly on the short-term impact of conventional pre-K on children's basic academic skills, the core policy question remains unanswered: what are the most effective early interventions for improving disadvantaged children's lives? To guide policy effectively, research must be improved by focusing on the most important questions instead of the most fashionable or convenient ones; increasing research transparency and replication; and pursuing new approaches to rigorous, policy-relevant research.

- **Advance high-quality child care and voluntary home visiting for disadvantaged children.** Our current knowledge base does not justify a large expansion of pre-K as the best path forward. Instead, the leading science and strongest research indicate that advancing high-quality, educational child care and supporting parents in better fulfilling their role as their children's "first teachers" are the most practical and promising avenues to help the children and families most in need. The growing pre-K push may well do more harm than good by diverting attention and scarce resources from other, more effective approaches.

Early childhood is gathering public and political momentum as one of the most important domestic policy areas of our time. But what America's most disadvantaged children are facing is not an achievement gap; it's a life gap. To close that gap, we need to move beyond a narrow focus on improving academic skills as the aim and expanding pre-K as the solution. Researchers, policymakers, and the public alike must remain focused on the core goal: to give all children, no matter the circumstances of their birth, a fair start in life.

Introduction

At the heart of science is an essential balance between two seemingly contradictory attitudes—an openness to new ideas, no matter how bizarre or counterintuitive they may be, and the most ruthless skeptical scrutiny of all ideas, old and new.

— Carl Sagan

Early childhood education has become increasingly prominent in the American policy landscape over the last several years. Between 1980 and 2016, the number of states with publicly funded pre-K programs increased more than fourfold, from 10 to 45. Since 2002, state spending on pre-K rose by nearly 300 percent, growing from $2.4 billion to almost $7 billion in 2016. In 2015 alone, 11 states boosted their pre-K funding by more than 25 percent. And the proportion of three- and four-year-olds attending preschool has almost tripled since 1970, up from 21 percent to 55 percent in 2013.[1]

It makes sense that early childhood is an emerging priority for policymakers. A rapidly growing body of brain research underscores the crucial impact of children's experiences from birth through age four. Other research has shown that high-quality early childhood programs hold great promise in helping disadvantaged young children succeed in school and life.

Recent polls show that the public also widely considers early childhood to be an important priority. In one 2014 poll, for example, 78 percent of Republicans, 83 percent of independents, and 93 percent of Democrats said they favor building better and more accessible preschool services.[2]

So far, the early childhood field has largely focused on expanding school-based pre-K programs for four-year-olds. But pre-K is just one part of a broad landscape of programs for children from birth through age four, and the emphasis on pre-K often overshadows other important approaches, such as child care and two-generation initiatives that work with children and parents together. Neither the public nor policymakers have a clear picture of the range of early childhood programs, the varied evidence on their effectiveness, and how that evidence can guide policy going forward.

This report aims to provide a starting point for a more comprehensive, nuanced dialogue around core policy goals in early childhood and the best strategies to accomplish those goals. It examines 10 of the best-known programs and highlights of the research on their impact, answering several broad questions about each: What is the specific nature of the program? Whom does it serve, and how is it designed? What kind of research has been conducted on it? What methods were used, and what results were found?[3]

A close look at the 10 programs reveals that they are as different from one another as they are similar. Some targeted four-year-olds; others focused on infants and toddlers. Some operated for 50 hours per week; others for just 15. Some ran for a single year; others for up to five. Some were entirely school based; others include intensive work with parents. In fact, much of the most-cited early childhood research is on programs that are not pre-K at all—and narrow debates over the pros and cons of pre-K exclude a great deal of knowledge about how to best serve children and families.

This report has two parts. Part I provides a short guide to the four research methods most commonly

used to evaluate early childhood programs. Research methods are usually ignored as esoteric and boring. But it is important to understand how those methods work, how the choice of methods can influence what results are found, and the particular limitations of each one. Part II describes 10 of the most widely cited early childhood programs, including details on program design and research findings on their impact.

The following pages are not intended as a scholarly examination or definitive review of the 10 programs, but rather aim to provide accessible information to a nonexpert audience. Our hope is to broaden participation in a crucial public debate—toward the widely shared goal of creating policy that will advance the well-being of America's most vulnerable young children.

Table 1. Overview of Program Characteristics

	Program Scale		Program Access		Target Group		Earliest Enrollment Age			Minimum Duration		
	Single-Site Model	Scaled-Up	Universal	Targeted	Children	Children and Families	Infancy	Age Three	Age Four	One Year	Two Years	Three+ Years
Abbott Preschool Program		●		●	●			●		●		
Abecedarian Project	●			●		●	●					●
Boston Pre-K		●	●		●				●	●		
Chicago Child-Parent Centers		●		●		●		●		●		
Georgia Pre-K		●	●		●				●	●		
Head Start		●		●		●		●		●		
Nurse-Family Partnership		●		●		●	●				●	
Oklahoma Pre-K		●	●		●				●	●		
Perry Preschool Program	●			●		●		●			●	
Tennessee Pre-K		●		●	●				●	●		

Source: Authors.

Part I

Early Childhood Research 101

Making sound policy decisions about whether to establish new early childhood programs or expand existing ones requires sufficient evidence about those programs' impact and effectiveness. However, both the nature and quality of the research used to assess program impact vary a great deal.

For example, some researchers study pre-K programs for four-year-olds, some study programs that serve infants and toddlers, and still others study programs that include both families and children. Some studies compare children whose parents sent them to pre-K with children whose parents did not, while others focus on children from only the first group. Some research methods give us information on long-term effects of programs while others solely provide data on kindergarten readiness. Some studies are done well; some studies are done poorly; some methods are more rigorous than others.

Understanding what questions are answered by particular studies and how those studies have been conducted is crucial to determining the relevance and value of their results. Early childhood experts know a lot about this. But it is often difficult for those outside the field to understand the various methods researchers have used to study programs, the implications of using one method over another, and how to interpret the findings studies have generated.

This section aims to help nonexperts navigate the often-confusing landscape of early childhood research. It provides a brief primer on the four methods that researchers most commonly use to study the impact of early childhood programs: Randomized Control Trial, Regression Discontinuity Design, Propensity Score Matching, and Difference-in-Differences.

We begin by explaining the Randomized Control Trial, which is widely considered the most rigorous method in experimental research. We then discuss the remaining three methods in order of most to least rigorous. For each, we first provide an overview describing the basics of the approach. For those interested in the nuances of these methods, we have also presented some more specific details on important aspects of research procedures.

It is important to note that all research methods have flaws, most studies have flaws, and flawed studies using flawed methods can still yield valuable information. This discussion is not meant to condemn or defend any particular method, nor to argue that any specific study's results are invalid or useless.

Yet research results are often reported as though they are universal truths, rather than findings from a particular study of a particular program in a particular context. Nonexperts—including parents, policymakers, and the general public—often fail to realize the extent to which the reported results are uncertain, shaped by the specific methods that generate them, and speak only to narrowly tailored questions. In other words, while research findings may be presented in black-and-white terms, especially in policy conversations around early childhood, those findings have more gray and less relevance than is often acknowledged.

Early childhood research plays an essential role in informing the focus and direction of the early childhood field. Our aim in this section is to put an intimidating body of knowledge into plain English to help nonexperts better evaluate the results and implications of that research for themselves.

Randomized Controlled Trial

Researchers have long considered Randomized Controlled Trials (RCTs) to be the most rigorous research method for determining a program's true impact.[4] In

fact, RCTs are the method used in high-stakes research, such as testing the efficacy of new drugs. The goal of an RCT is to maximize confidence that any change observed after the implementation of a program or policy was *caused* by the intervention, not by some other factor. As with any method, RCTs can be well or poorly executed, but if done well, an RCT provides the strongest, most convincing evidence for a program's impact.

In early childhood research, an RCT usually starts with a group of children, all of whom are applying to a particular program. The study randomly accepts some of the children into the program while the rest are assigned to a group that does not participate in the program. If randomization is done correctly, any differences between children are also randomly distributed between the two groups—meaning that the two groups will effectively be the same, apart from their participation in the program.

For example, imagine a pre-K program that has only 300 available spaces, but 600 children who want to attend. In a study conducted using an RCT, applications are accepted for all 600 children. Researchers then use a lottery-like process to select 300 children who are given a space in the program. Those children are called the treatment group. The other 300 children are excluded from the program and form the non-program control group for the study.

Using this method, researchers can assess the program's long-term impact by following both groups of children and comparing them on whatever measures are selected. Any difference in outcomes found between the two groups can be attributed to the program, because the study started with children who were alike in every other way, in both observable and unobservable characteristics. In some studies, researchers have followed participants for years to determine both shorter-term outcomes, such as kindergarten readiness and elementary school performance, and longer-term outcomes, such as high school completion, adult criminal activity, and employment.

Discussion. While RCTs are the most rigorous research method available in the social sciences, they still have several drawbacks. First, the approach is typically dependent on having more demand for a program than available spaces. That means that only a limited universe of programs can be subject to RCTs. Second, they are the most expensive and complex kind of study to conduct and require allocating significant resources to research rather than potentially serving more children. Understandably, the long-term benefit of assessing a program's impact is often outweighed by a short-term desire to serve as many children as possible.

Third, many RCTs can only tell us a program's average effects. While those averages are useful for assessing a program's overall impact, they obscure varied and even contradictory results and leave unanswered the important question of *what* works best for *whom*.

For example, an RCT showing that a 50-site program has minimal impact tells us only that the collective impact is small, not that the program is ineffective across all sites and for all children. Some sites might have large impacts while others have none. Some children may benefit much more than others. An RCT can be designed to answer some of those questions, but it is more difficult to carry out, and many studies focus on only overall effects.

Finally, conducting social science research well is always a challenge, and this is not less true in early childhood. When an RCT study randomizes children out of a particular program, excluded families will actively seek other options. Many will place their children in alternative programs, which may be either better or worse than the program being studied. Because preventing families from doing this is neither feasible nor ethical, establishing "no program" control groups is often not possible. So these kinds of studies usually tell us the program's impact compared to families' other options, not compared to having no program at all.

RCTs, like all methods, are subject to flaws when implemented in the real world. Researchers cannot aim to conduct perfect research—only the best possible research. And all things being equal, RCTs are the most rigorous among several research methods in early childhood. Ultimately, the important question to ask about any one study is not whether it is perfect, but whether it provides the best, most rigorously obtained information available.

Regression Discontinuity Design

Regression Discontinuity Design (RDD) studies are often used to examine the short-term impact of pre-K on children's early kindergarten skills. While a Propensity Score Matching study, which we discuss next, explicitly compares children whose parents send them to pre-K with children whose parents do not, an RDD study focuses exclusively on children from the first group. In other words, children who are eligible for pre-K but whose parents choose not to enroll them are excluded from RDD studies entirely.

RDD studies are not randomized experiments like RCTs, but are considered "quasi-experimental." Researchers attempt to approximate random assignment by assigning children to two groups (program and no-program) based on a "forcing variable." In early childhood research, that forcing variable is often age: the RDD method is frequently used to study public pre-K programs in cities and states that have a strict age cutoff for pre-K enrollment. Commonly referenced RDD studies include those on pre-K programs in New Jersey (the Abbott program), Boston, Oklahoma, and Georgia, which are discussed in the next section.

In an RDD study, researchers identify two groups of preschool-age children: one that makes the age cutoff and enrolls in a pre-K program (the treatment group) and one that misses it and enters pre-K the following year instead (the control group). A year later, the treatment group has completed a year of pre-K and is starting kindergarten while the control group is just entering pre-K. The researchers test both groups of children and compare their scores. The premise of these studies is that children who were born just before the age cutoff are virtually identical to those born just after, so researchers can attribute any differences the study finds entirely to the program.

Discussion. The RDD approach has several big advantages. It can be cost-effective and is relatively easy to implement on a large scale. It is also considered the most rigorous research method besides randomized control trials. Because of these advantages, RDD studies are now the most commonly used method in pre-K research. But the method has four shortcomings, which are important to keep in mind when evaluating study results.

How an RDD Study Works

Take a state in which children must turn four years old before September 1 to start pre-K. Imagine that Alex turns four on August 31, 2013, and is admitted to the pre-K program that fall, while Jesse turns four on September 1, 2013, and must wait until fall of 2014 to enroll in pre-K.

A year later, Alex and Jesse are both turning five. Alex is now beginning kindergarten and, assuming he stayed in the program, has had a year of pre-K. But Jesse has not had pre-K yet, even though the two children are only a day apart in age.

In an RDD study, researchers find large groups of Alexes (the treatment group) and Jesses (the control group), test both groups of children, statistically adjust the scores for small age differences, and compare the test results of the two groups—one that has completed a year of pre-K and one that is just beginning the program.

If children in the treatment group (such as Alex) have higher average scores than children in the control group (such as Jesse), researchers conclude that this was caused by the pre-K program, because the only difference they have identified between the children in the two groups is that one has had a year of pre-K while the other has not.

Attrition. The first limitation of RDDs is the problem of program dropouts, or attrition. Attending pre-K is voluntary, and not all children who start a pre-K program finish it. Children may move, transfer to a different program, be withdrawn by their parents, quit because they are unable to handle the program, or be expelled, among other reasons.

In an RDD study, children who leave the pre-K program are eliminated from the treatment group; the outcomes reported for that group include test scores only for children who successfully completed the pre-K program and entered kindergarten. However, the control

group is tested when entering pre-K, so that group's outcomes include test scores for *two* distinct subgroups of children: those who will end up dropping out over the course of the year, along with those who will complete the program.

In other words, the children who cannot or choose not to complete the pre-K year are weeded out of the treatment group but are included in the control group. In the earlier example, for instance, if Alex dropped out midyear, he would be eliminated from the study altogether, but Jesse would be included regardless of whether she ended up dropping out. So the control group includes all the children who are *going to* drop out, while the treatment group includes only children who *did not* drop out.

From a research point of view, this is a comparison between apples and oranges, because we do not know if the type of child who completes the pre-K program is comparable to the type who does not. Therefore, when the two groups of children are evaluated, it is not possible to tell whether the results are because of the impact of the program or because dropout children are weeded out of the study. For public pre-K programs that target disadvantaged children, researchers in fact often find that lower-performing children disproportionately fail to enroll in the program, or enroll and then drop out.

In the Boston study discussed in the next section, for example, 18 percent of the children in the pre-K group dropped out before testing at kindergarten entry, and the researchers reported that the children who dropped out were more disadvantaged than those who completed the program.[5] Researchers can use statistical methods based on observed variables to adjust for attrition (which the Boston researchers did), but that information is crucial for interpreting study results. For this reason, US Department of Education research standards require that attrition in an RDD study be recorded and reported.

Regrettably, though, many studies do not adhere to this standard, which makes it more difficult to assess their results. The researchers who conducted the RDD studies of the Abbott, Georgia, and Oklahoma programs discussed in this paper did not report attrition rates, so we do not know if their dropout rates were higher, lower, or the same as Boston's; whether the characteristics of the dropout children were different from those who completed the program; or how the dropout factor may have affected the results.

Test Timing. The second common shortcoming of RDD studies is the time frame for testing the two groups of children being studied. In RDD studies, researchers report the treatment group's test scores as children's gains from "a year of pre-K" and those of the control group as the outcome of "no pre-K." Therefore, to accurately reflect the *impact of pre-K*, which is what is being reported, testing must occur before the school year begins.

In actuality, however, many studies test children months into the school year. For example, Georgia's school year starts in early August, but testing for the study did not begin until September 21 and continued until the end of December.[6] Researchers in Boston carried out testing throughout the fall, reporting that only one-third of the children were tested by the end of October and just 88 percent by the end of November.[7] In Oklahoma, researchers reported that testing was completed for the most part during the first week of school, although they did not clarify further.[8] Researchers from the Abbott study reported no information on test timing.

This is an important problem, because it means the average test scores reported for children in the treatment group do not actually reflect a "year of pre-K," even though that is how the data are presented. Rather, those scores include additional gains from weeks or even months of kindergarten. Similarly, the average test scores of children in the control group do not reflect "no pre-K," but instead include gains from weeks or months of pre-K.

In other words, researchers in a number of studies have attributed the treatment group's results entirely to the pre-K program even though some children attended kindergarten for months before being tested. Kindergarten is often more intensively focused on teaching basic skills than pre-K is, so children's time in kindergarten is likely to have a significant impact on their test results. Because the impact of kindergarten has been conflated with the impact of the pre-K year in these

studies, it is impossible to know what is really causing the measured gains.

Limited Generalizability of Findings. The third shortcoming of RDD studies is that their assignment mechanism—the forcing variable of age—means that the method's "randomization" is most effective in a narrow bandwidth immediately above or below the age cutoff. A child born on August 31 is essentially the same age as a child born on September 1, so age is not an important difference between those two children.

But comparability between groups declines as children get further from the age cutoff that determines the two study groups: in other words, preschoolers who are apart in age by one day are much more comparable than those who are apart in age by 364 days. For that reason, other kinds of research designs (such as RCTs) are needed to confirm the degree to which results from RDD studies hold true for a broader group of children.

Long-Term Impact. Finally, the most significant shortcoming of RDD studies is that they do not answer the question parents, the public, and policymakers really care about: how early childhood programs impact children's long-term success in school and life. Instead, RDD studies are only able to show whether children who attend a pre-K program have higher test scores in the first months of kindergarten. But higher test scores in kindergarten matter only if they are associated with better school performance in later grades and, ultimately, better life outcomes when those children become adults.

So do test scores in the first half of the kindergarten year predict the rest of children's lives? Some studies show some correlation between kindergarten test scores and later success. But other studies show that they are at best a weak predictor of positive outcomes down the road. Little research has been conducted on this question. And beyond the research, common sense suggests that changing the trajectory of disadvantaged children's lives—and knowing if it has been successfully changed—is going to require more than raising and measuring kindergarten test scores.

As noted, RDD studies are the most commonly used research method in studying the impact of pre-K and have made useful contributions to the pre-K field. Early childhood experts do not agree on the extent to which the methodological problems noted here undermine RDD study findings. Nevertheless, RDD studies can certainly be strengthened by implementing more rigorous procedures, such as limiting testing to the first week or two of the school year, reporting attrition rates, making the control group comparable to the treatment group by retroactively excluding the children who do not finish the pre-K year, following all students who initially enroll in the program, or even including children who do not enroll in pre-K at all as an additional comparison group.[9]

At the same time, rigorous studies that directly address meaningful longer-term impact are needed to establish the knowledge base that policymakers and the public can rely on to justify new public investments in pre-K.

Propensity Score Matching

A Propensity Score Matching study assesses a pre-K program's impact by comparing children who attended pre-K with children who did not, without using randomization. Instead, researchers construct a "matched" comparison group based on a set of observed variables the researchers believe are associated with a child's later school and life outcomes. The method has been used to examine school performance and longer-term outcomes (such as high school graduation rates, criminal activity, and adult employment) for people whose parents sent them to pre-K as children.

In a typical matching design study, researchers first identify the variables that they will use to create the two matching groups, such as neighborhood, family income, family structure, parents' education and employment status, race/ethnicity, and home language. They then identify a target group of people—kindergartners, fourth graders, 21-year-olds, or any other age group—and use a statistical method to "match" them on each characteristic. Of that whole group, they find out which ones went to pre-K and which ones did not. Researchers then determine if the subgroup that did attend pre-K is, on average, doing better than the subgroup that did not attend.

Why Unobserved Variables Matter

Imagine two children, David and Michael, who are living in the same Chicago housing project. David and Michael are similar in many other respects too: both are African American, poor, and born to a teenage single mother who did not finish high school, among other common characteristics.

Now imagine that their mothers are very different in crucial ways. David's mother has the emotional capacity and drive to help her child do better in life than she did. She makes a concerted effort to enroll him in pre-K, manages the logistics of getting him there and picking him up every day, stays in regular touch with his teacher, and helps resolve problems to make sure he successfully finishes the program. Michael's mother is disorganized, self-absorbed, and depressed and always hated school. She does not make the effort to send Michael to pre-K, so he does not attend.

David goes to pre-K, and Michael does not. Five years later, David is doing much better in fourth grade than Michael is. In a matching design study, David's stronger performance in fourth grade is attributed to the pre-K program, because David and Michael are so similar on all the variables researchers have observed and measured.

But it is quite possible that pre-K is not the crucial factor causing the differences in the two children's later school performance. It is even possible that their school outcomes would be similar if David had been unable to get a space in a limited-capacity program while Michael was enrolled by a concerned neighbor and was luckier in securing a spot. We just do not know whether David's success in school is because he went to pre-K, because he has a more engaged mother, or some combination of the two.

The David/Michael scenario may be exceptional, or it may not be. But a matching design study does not address these factors, which is why it is hard to come to definitive conclusions about the impact of pre-K using this method.

If researchers find a correlation between attending pre-K and later success, they conclude that pre-K caused the improved outcomes, because they have identified the children studied as so similar otherwise. In other words, the only difference between the two subgroups is assumed to be that one went to pre-K and the other did not. Any variation in later outcomes is thus attributed to the pre-K program.

Discussion. The advantage of this approach is that it enables researchers to study the long-term impact of programs for a range of important outcomes. Its Achilles' heel, though, is what are called "unobserved variables." In the case of pre-K research, this refers to the fact that children who attended pre-K and children who did not may actually be dissimilar in important ways that are not on the researchers' list of observable characteristics.

Unobserved variables are also a reason that attrition compromises the results of matching design studies, such as those of New Jersey's Abbott Preschool Program and the Chicago Parent-Child Centers, both described in the next section.[10] In that kind of study, researchers aim to follow and compare the progress of large groups of similar children who did or did not attend pre-K to determine its long-term impact. However, tracking down all children from the original matched treatment and comparison groups is usually not possible. For example, researchers were able to find just 72 percent of the original pre-K group and 66 percent of the original "no pre-K" comparison group when trying to assess the impact of the Abbott program on children's fifth-grade school performance.

Researchers do not know what causes children to disappear from study groups or whether there are any relevant differences between the children they can find and those they cannot. Attrition from study groups could occur just by chance or it could be associated with important, unobserved variables that bias study findings in ways that researchers are not able to account for. For example, at-risk children may be disproportionately likely to leave the study.[11] Results from PSM studies need to be evaluated with this limitation in mind.

Difference-in-Differences

The Difference-in-Differences (DD) method is often used to evaluate changes in outcomes associated with implementing a new state- or county-wide pre-K program. The method does not focus on a program's direct effect on participating children; instead, researchers compare a particular outcome (such as a county's average score on a state test) from before and after the implementation of the program. The researchers then compare any change in that outcome to those of a large group of children in a state or county that was not affected by the policy.

For example, Maria Donovan Fitzpatrick used the DD method to study the impact of a new universal pre-K program in Georgia when the proportion of four-year-olds attending pre-K grew from 14 percent in 1995 to 55 percent in 2008.[12] Fitzpatrick first determined the association between the increased availability of pre-K and any subsequent change in the state's fourth-grade scores on the National Assessment of Educational Progress (NAEP) exam, which the initial pre-K cohort took a few years after the policy's adoption. She then compared Georgia's results to the fourth-grade NAEP scores in states that did not have universal pre-K. (The study eventually concluded that the pre-K program did not have a statistically significant impact on child outcomes.)

Discussion. The DD method is relatively easy to implement, is low cost, and can provide information on a program's long-term impact. However, the approach has two big shortcomings. The first is that it tells us nothing about specific effects on individual children who actually participated in a program; it only assesses broad, aggregate association with limited state- or county-wide measures, such as test scores and high school graduation rates.

The second is that many social, economic, and education trends other than pre-K can affect children's later school and life outcomes. Because the DD method does not allow researchers to disentangle those other effects from the effect of pre-K, it is hard to produce convincing evidence that any observable gains were because of the pre-K program alone.

Table 2. Overview of Research Designs Discussed in Part II

		WHAT IS ASSESSED	
		Early Kindergarten Skills	**Long-Term Impact**
RIGOR OF RESEARCH DESIGN	**Stronger**	**Regression Discontinuity Design** Abbott, Boston, Georgia, Oklahoma	**Randomized Control Trial** Abecedarian, Head Start, NFP, Perry, Tennessee
	Weaker	–	**Matching Design** Abbott, Chicago

Source: Authors.

Part II

Ten Early Childhood Programs

The following are brief overviews of 10 of the best-known, most widely cited early childhood programs of the last half century: Abbott Preschool Program, Abecedarian Project, Boston Pre-K, Chicago Child-Parent Center Program, Georgia Pre-K, Head Start, Nurse-Family Partnership, Oklahoma Pre-K, Perry Preschool Program, and Tennessee Voluntary Pre-K.[13]

We do not intend to suggest that these are the 10 best or most effective programs, nor are the program overviews meant to be comprehensive evaluations or in-depth critiques. Our aim is simply to provide short sketches of 10 leading early childhood programs and how they have been studied, for nonexperts in particular.[14]

Key Points to Keep in Mind

When reading the 10 program overviews in the following pages, we encourage you to keep the following points in mind.

The Devil Is in the Details. The details of the program descriptions matter a great deal. All are called "early childhood programs" and serve children under age five, but designs and aims vary considerably across programs (see Table 1 on p. 4). While program outcomes are important, a clear picture of the specific program that produced those outcomes is at least as crucial to understanding research in the early childhood field. For example, a 10-month, school-based program that serves a diverse group of four-year-olds for 6 hours per day bears little resemblance to a five-year, center-based program that serves poor children for 10 hours per day from infancy to kindergarten.

Study Findings Vary Widely. Some studies find impacts on children's academic achievement alone, while others find a range of social-emotional, cognitive, and academic effects. Some fail to find gains that persisted beyond kindergarten, while others find positive effects lasting decades.

Research Methods Matter. A basic understanding of how studies are conducted is crucial to correctly interpreting their results. Three of the four commonly used research methods described in the previous section were used in the studies discussed here (see Table 2 on p. 11).

Some specific details of procedures used in the studies are not included in the following descriptions. We encourage you to read the studies themselves (see References on p. 42) to get a fuller picture of how the research was carried out.

Statistical Significance Doesn't Mean Policy Significance. While they are often conflated, there is a big difference between statistical significance and practical significance—and this distinction is crucial when using research to inform policy decisions. Statistical significance is a technical term researchers use to indicate that a study's result is very unlikely to have occurred by random chance.[18] The results reported for the studies are "significant" in that sense. A result can be large or tiny, relevant or irrelevant, and still be "statistically significant."

But a result found to be significant in a study is not necessarily important from a policy point of view. The practical significance of a study's results—its meaningful, real-world impact—has to be evaluated as an entirely separate question. Researchers determine whether a result is statistically significant; policymakers assess how much a program moves the needle on

Statistical Versus Policy Significance: An Illustration

Imagine that a large group of sixth graders in a particular school district scores an average of 75 on a year-end district math test, with a standard deviation of 10. Children who score below 55 are rated as low performing; from 55 to 64 as below average; from 65 to 84 as average; from 85 to 94 as above average; and above 95 as exceptional.

The gaps between lower- and higher-achieving children are substantial. A gap between a child in the middle of the below-average group and one in the middle of the above-average group is 30 points, or 3 SD. The gap between the top of the low-performing scores and the bottom of the exceptional scores is 40 points, or 4 SD.

The next year, the district sets up a yearlong tutoring program to raise the performance of the lowest-scoring children. Researchers conduct an end-of-year evaluation and find that the program increased those children's scores significantly, reporting a large effect size of 0.8 SD. The district concludes that it is a successful program that is closing the gap.

In the real world, though, the "significant, large effect" researchers found meant that the average score of children in the tutoring group was increased from 60 to 68—which is better, but makes only a small dent in the overall achievement gap and may not be the game-changer those children really need.

The point is that an impact researchers describe as "significant" and "large" may not actually be significant or large from a policy point of view—and policymakers need to take that into account when making decisions about how to address a particular problem.

an important social problem, whether outcomes justify the cost, and whether limited resources could be better spent on something else that more effectively addresses that particular problem.

Guidelines for Interpreting Study Findings

To interpret study findings correctly, three aspects of findings are important to understand: effect sizes, the meaning of statistical versus practical significance, and the specific nature of the outcomes a study is measuring.

Effect Sizes. A program's impact on participating children is reported in different ways depending on the data presented. Sometimes results are reported in terms of percentages—such as "children who participated in the program had 30 percent less special education placement by fourth grade"—which are relatively intuitive. But sometimes results are reported in terms of "effect sizes"—such as "an effect size of 0.38 in math achievement"—which are harder to interpret.

An effect size is expressed as a fraction of one standard deviation (SD)—that is, an effect size of 0.2 is 20 percent of an SD. (See the sidebar above for an explanation of standard deviation.) Conventional guidelines consider effect sizes of less than 0.3 SD as "small," of 0.3 to 0.8 SD as "moderate," and of 0.8 SD or more as "large."[15] For example, if researchers are investigating a program's impact on children's math achievement and find an effect size of 0.52 SD, they would usually describe that as a moderate effect on children's achievement.

However, this general rule can vary depending on the context. To address that, effect sizes are sometimes translated into practical information, such as how many months ahead an effect size represents in terms of children's average annual gains. For example, a moderate effect size of 0.5 SD in reading would roughly translate to three months of the average achievement gain in kindergarten: in other words, children reached a particular level in September that they otherwise would not have reached until December. Similarly, a small effect size of 0.2 SD would mean they reached a level in September that they otherwise would not have reached until

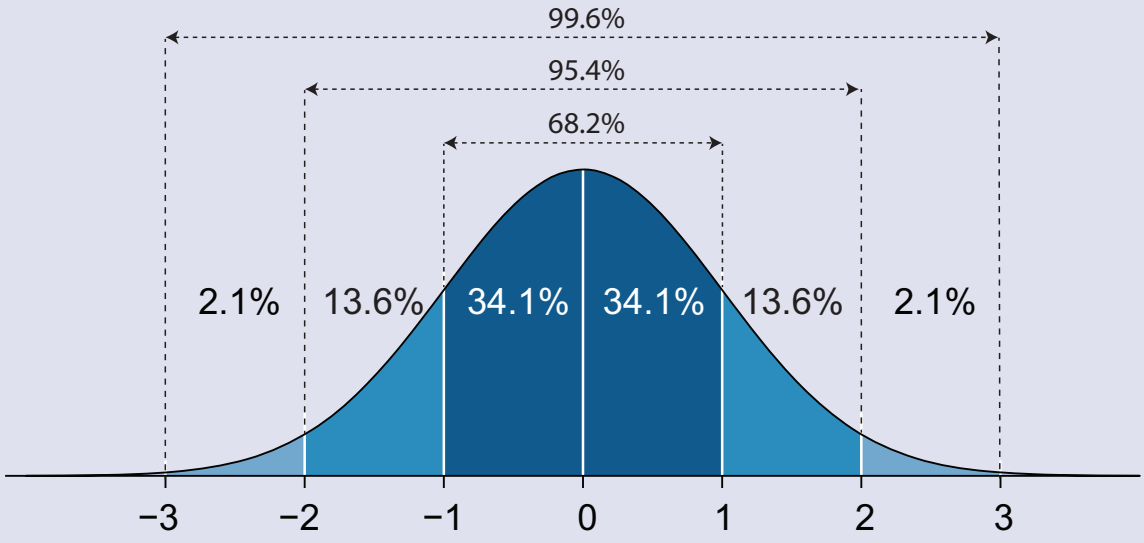
What Is a Standard Deviation?

"Standard deviation" is a statistical concept used to express how close or far a value is to a group's average. Statisticians use a bell curve (such as in Figure 1) to demonstrate a normal distribution of the data points for the group, with the data distributed around the highest point in the curve.

For instance, take a group of children that scored an average of 100 points on a state math achievement test with a standard deviation of 16. That means 68 percent of the children scored between 84 and 116 (that is, plus or minus one SD—16 points—from the average of 100); 95 percent scored between 68 and 133 (plus or minus two SDs, or 32 points); and 99 percent scored between 52 and 148 (plus or minus three SDs, or 48 points).

In this scenario, if the average test score for a subgroup of low-scoring children starts at 65 and is increased by 8 points to 73, that is a gain of one-half a standard deviation (because the SD is 16 for the whole group) and would be described as "an effect size of 0.5 SD."

Figure 1. Example of a Bell Curve



Source: Authors.

October. (See Table 3 on p. 15 for a translation of effect sizes into months of an average school year.)

Another useful metric that can be used to evaluate effect size is how much a program outcome narrows achievement gaps. For example, researchers have found that math and reading achievement gaps between children in the bottom and top income quintiles are more than a full standard deviation when they enter kindergarten.[16] Therefore, an effect size of 0.5 SD translates into reducing the achievement gap at kindergarten entry by almost one-half.[17]

Skills Measured in Studies. For six of the programs described in the next section, researchers measured program outcomes by using tests of children's basic academic skills. Studies use different tests and terminology, but they all assess skills in two general areas: pre-reading and mathematics.

Pre-reading skills include naming letters, writing letters, recognizing words, spelling words, knowing the meaning of words (vocabulary), and pronouncing words correctly. Studies report these skills using terms such as *language/literacy*, *print awareness*, *letter*

Table 3. Effect Size (in Standard Deviations)

		READING			MATH		
	Grade	*Small* 0.2 SD	*Medium* 0.5 SD	*Large* 0.8 SD	*Small* 0.2 SD	*Medium* 0.5 SD	*Large* 0.8 SD
ESTIMATED MONTHS OF AVERAGE ACHIEVEMENT GAINED	Pre-K	2	5	8	2	5	8
	K	1	3	5	1	3	5
	1st	1	3	5	1	3	5
	2nd	2	4	6	2	4	6
	3rd	2	6	10	2	5	8

Note: This table converts effect sizes into estimated months of average annual achievement for specific grades. In kindergarten, for example, a small 0.2 SD effect size roughly translates into one month of achievement. For more information, see http://gse.buffalo.edu/faculty/centers/gaps/calculator1.
Source: Jaekyung Lee, Jeremy Finn, and Xiaoyan Liu, "Time-Indexed Effect Size for P–12 Reading and Math Program Evaluation," Society for the Research on Educational Effectiveness, Spring 2012, http://gse.buffalo.edu/gsefiles/images/conf_methods_abstract_2012-revised.pdf.

knowledge, *letter-word identification*, *spelling*, *phonemic awareness*, and *receptive vocabulary*.

Math skills include counting and simple calculations. They are reported using terms such as *math, applied math problems*, *numeracy and geometry*, *counting*, *math problem solving*, and *quantitative concepts*.

In addition to measuring academic skills, one study (Boston pre-K) also tested children in the noncognitive skills of executive function and emotional development. The study reports those skills using the terms *working memory*, *cognitive inhibitory control*, *attention shifting*, and *emotion recognition*.

****

The following 10 program overviews are presented in alphabetical order. Each has four sections:

1. A description of the program itself, including the program's goals, when and where it has operated, and whom it served and how;

2. An overview of recent or widely cited research on the program, explaining what was investigated and how the research was carried out;

3. Highlights of study findings; and

4. A brief summary of key takeaways.

Abbott Preschool Program

New Jersey's Abbott Preschool Program is a state-funded, public preschool program free to all three- and four-year-olds in 35 of the state's lowest-income school districts (now known as the "Former Abbott and Expansion Districts"). The program's goal is to prepare children to enter school with the knowledge and skills necessary to meet New Jersey achievement standards.

Abbott is a mixed delivery system managed by the state's public schools: as of 2015, about 44 percent of children were served in public school classrooms, and the other 56 percent attended programs in private centers and Head Start agencies that contract with local boards of education. The program served approximately 43,000 children in 35 districts during the 2014–15 school year, constituting almost one-quarter of all the state's three- and four-year-olds and almost 85 percent of the three- and four-year-olds in the Abbott districts.

Children can attend Abbott for either one year beginning when they are four years old or two years beginning when they are three. The program runs on the public school schedule of six hours per day for a 180-day school year—a total of 1,080 program hours for one year and 2,160 hours for two years. The New Jersey Department of Education also coordinates with the state's Department of Human Services to provide before- and after-school child care and summer programs for up to 10 hours per day and 245 days per year, which are free to low-income families.

Each classroom is staffed by a state-certified lead teacher and an assistant teacher, and the maximum class size is 15 children. Staff are provided with ongoing supervision and coaching and receive the same salary and benefits as public school teachers.

Study Description. In the fall of 2005, researchers implemented a two-part research project—the Abbott Preschool Program Longitudinal Effects Study—to study the Abbott pre-K program's impact on children's academic performance, both when they enter kindergarten and over the long term.

For the first part of the study, the researchers used an RDD to compare two groups of children. That group was composed of 766 children who had already attended Abbott pre-K and were entering kindergarten in fall of 2005. The first group included 451 children who had attended one year of preschool and 303 who had attended two years. This part of the study examined the impact of one and two years of the program on children's academic skills when they were starting kindergarten. The second group was composed of 305 children who were just entering the pre-K program that fall because they had missed the age cutoff for the previous year.

For the second part of the study, the researchers used a propensity score matching design with children drawn from the original RDD. In a follow-up through second grade, they compared 754 children who had attended Abbott (451 attended for one year and 303 for two years) with a group of 284 children with similar demographic characteristics from the same kindergarten classrooms who had *not* attended Abbott. Subsequently, they conducted a follow-up through fifth grade, comparing 553 children who had attended Abbott with 201 children who did not attend.

The Bottom Line. Both the RDD and matching design parts of the study showed that children whose parents enrolled them in the Abbott program scored higher than their peers in language arts, literacy, and math at kindergarten entry. The RDD showed stronger impacts than the matching study, but whether gains are overstated by the RDD or understated by the matching study is unknown. The matching study also found somewhat larger gains at kindergarten entry for children who attended the program for two years rather than one.

Researchers were able to follow a little more than two-thirds of the original pre-K group into fifth grade: 553 of the 766 children who attended Abbott (72 percent) and 201 of the 305 children who did not attend (66 percent). Of those children, researchers found small to moderate gains on tests of basic academic skills. Children were also a few percentage points less likely to be retained in grade or placed in special education. By the end of fifth grade, children who had attended two years of pre-K had slightly larger gains in academic skills. However, they were slightly more likely to have been retained in grade or placed in special education than children who attended just one year of pre-K.

Study Findings

Part I: RDD
Effects at Kindergarten Entry for One Year of Pre-K

- 0.28 SD for language/literacy
- 0.56 SD for print awareness
- 0.36 SD for math

Part II: Matching Design Study
Effects at Kindergarten Entry

One Year of Pre-K

- 0.21 SD for language/literacy
- 0.29 SD for print awareness
- 0.20 SD for math

Two Years of Pre-K

- 0.42 SD for language/literacy
- 0.31 SD for print awareness
- 0.34 SD for math

Effects at the End of Fifth Grade

One Year of Pre-K

- 0.18 SD for language/literacy
- 0.14 SD for math
- 35 percent less special education placement (11 versus 17 percent)
- 42 percent less grade retention (11 versus 19 percent)

Two Years of Pre-K

- 0.22 SD for language/literacy
- 0.29 SD for math
- 24 percent less special education placement (13 versus 17 percent)
- 37 percent less grade retention (12 versus 19 percent)

While the study found modest academic gains, it is difficult to draw strong conclusions from the findings because of the lack of data on attrition and test timing for the kindergarten-entry group and the exceptionally high attrition levels from the study of the fifth-grade group.

Abecedarian Project

The Abecedarian Project was a small, carefully designed, educational child care program, run as a single-site research project in the mid-1970s at the University of North Carolina–Chapel Hill. The project provided 57 low-income, high-risk children with an intensive, full-time, year-round, early learning program beginning in infancy and continuing for five years. Its aim was to promote children's language, cognitive, social-emotional, and motor development.

Children entered the program when they were a few weeks old and remained until they entered kindergarten at age five. Four cohorts of children were admitted between 1972 and 1977. The program operated for 10 hours per day, 5 days per week, and 50 weeks per year (a total of 2,500 hours per year for 5 years) and used a special curriculum focused particularly on language development through high-quality adult-child interactions. Caregivers received intensive in-service training, and caregiver-child rations were high: from 1-to-3 for infants to 1-to-6 as children moved into preschool. The children's parents served on the center's advisory board and were provided a series of informative programs on parenting, periodic social events, and counseling on their child's health and development.

Study Description. The Abecedarian Project study was conducted using an RCT. It aimed to determine whether the program could lower the risk of developmental delays and academic failure for children born into low-income families.

The study randomly assigned 57 children to the program and 54 children to a control group that did not participate in the program. Of the 111 total children included in the study, 98 percent were African American, 76 percent lived in single-mother households, and 55 percent were on welfare. Participating children were born to mothers with an average age of 20 and an average IQ of 85. Sixty-six percent did not have a high school diploma.[19]

Follow-up studies were periodically conducted on both groups from 1984 to 2014 to determine the program's long-term impact. The most recent follow-up

Study Findings

Improved Academic Outcomes

- Children who attended the program were:
  - 37 percent less likely to be placed in special education (31 versus 49 percent);
  - 48 percent less likely to be retained in grade (34 versus 65 percent); and
  - Two years ahead in reading scores and more than one year ahead in math scores at age 21.
- Teenage mothers from the program group were two and a half times more likely to complete high school (46 versus 13 percent).

Increased Employment and Completion of Higher Education

- Children who attended the program were:
  - Almost three times more likely to have earned a bachelor's degree by age 30 (23 versus 6 percent); and
  - 42 percent more likely to be in full-time employment at age 30 (75 versus 53 percent).
- Mothers were 39 percent more likely to be employed when their children were 15 (92 versus 66 percent).

Better Family Planning

- Just one-quarter of the program group became parents as teenagers, compared with almost half of the control group (26 versus 45 percent).
- The program group had their first child an average of almost two years later than the control group (21.8 versus 20 years of age).

Decrease in Dependence on Public Assistance

- By age 30, program participants were six times less likely to have received welfare benefits for at least 10 percent of the prior seven-year period.

Better Physical Health

- By their mid-30s, no male program participants exhibited "metabolic syndrome"—a cluster of conditions associated with greater risk of heart disease, stroke, and diabetes—compared to one-quarter of the males in the control group (0 versus 25 percent).

study in 2014 tracked 101 of the original participants when they were in their mid-30s.

The Bottom Line. Along with Perry Preschool and Nurse-Family Partnership, both also discussed in this paper, the Abecedarian Project is considered one of the highest-quality early childhood programs of the past half century. It is well-known for its large, sustained effects on participants' educational attainment, employment, and other life outcomes and for the positive effects it had on participants' mothers.

Abecedarian is often cited in proposals for expanding pre-K. But it is important to note that Abecedarian was a high-quality, educational child care program, bearing little resemblance to pre-K programs for three- and four-year-olds. In addition, the program was a single-site model run by the researchers who designed it. Whether and how the program's quality could be maintained at a larger scale is not clear. The important takeaway from the Abecedarian study is that full-time, high-quality child care for disadvantaged children beginning when they are very young can have a powerful impact on their later life outcomes.

Boston Pre-K

Boston's pre-K program is a publicly funded, universal preschool program free to all four-year-olds living in the city's school district. Developed and operated by the Boston Public Schools (BPS) in 2005, the program currently serves approximately 2,500 (a little more than 40 percent) of the city's four-year-olds.

The BPS pre-K program runs on the regular public school schedule: 5 days a week, 6.5 hours per day, for a 180-day school year (a total of 1,170 hours for the year-long program). Lead teachers must have a minimum of a bachelor's degree, must obtain a master's degree within five years, and are members of the public school teaching force. All classrooms are also staffed with a paraprofessional (instructional assistant). Teacher-child ratios are a minimum of 1-to-11 with a maximum class size of 22 children.

Boston's pre-K program strongly emphasizes ongoing quality improvement driven by data from multiple sources including child outcomes and measures of classroom instructional quality. All classrooms use a uniform, research-based curriculum for both literacy (based on the 2005 version of *Opening the World of Learning*) and mathematics (*Building Blocks*). Teachers receive five days of initial training in implementing the curriculum and year-round, classroom-based coaching to ensure high-quality instruction.

Study Description. Researchers used an RDD to investigate the impact of BPS pre-K on participants' language, literacy, and mathematics skills—domains specifically targeted by the BPS pre-K curriculum—and noncognitive domains, such as executive function and emotional development. The study compared two groups of children: one group that had attended a full year of BPS pre-K in 2008–09 and was beginning kindergarten in fall 2009 and a second group that missed the age cutoff for the previous year and was beginning pre-K in fall 2009.

The final study group of 2,018 children was composed of 969 children who had completed pre-K and were beginning kindergarten in fall 2009 and 1,049 children who were just beginning pre-K that fall. Forty-one percent were Hispanic, 26 percent were black, 18 percent were white, 11 percent were Asian, and 3 percent were of mixed or other race. Fifty percent of the sample came from English-speaking homes; 28 percent came from Spanish-speaking homes; and 22 percent came from homes that spoke a language other than English or Spanish. Sixty-nine percent were eligible for free or reduced-price lunch.

Trained assessors tested participating children for receptive vocabulary, pre-reading and reading skills, numeracy and early math skills, working memory, cognitive inhibitory control, attention shifting, and emotional development. All children were tested in English, regardless of home language spoken. Testing began at the end of September 2009 (two weeks after the start of school) and continued throughout the fall. Approximately 33 percent of testing data was collected by the end of October, 88 percent by the end of November, and 98 percent by the end of December.

The Bottom Line. The Boston study offers a useful look at a universal, citywide pre-K program. Findings showed that the program had moderate to large impacts on children's language, literacy, and mathematics skills when they were entering kindergarten and smaller impacts on executive functioning and one measure of emotion recognition. While all children benefited, impacts were considerably larger for some subgroups, especially minority and low-income children.

Boston's program is very carefully run and of higher quality than typical city and state pre-K programs. BPS pre-K uses uniform, evidence-based curricula for math and literacy across all classrooms and provides teachers with intensive training and ongoing coaching. The program's teachers also have unusually high levels of education and experience: during the year studied, 78 percent of program teachers held master's degrees, and 75 percent had at least five years of teaching experience. The researchers were not able to identify which of the inputs—curricula, teacher education and experience, training and ongoing coaching, student attendance, or some combination of these—caused the program's impacts. Because of the program's uniqueness, study findings should be interpreted as specific to BPS's program design and management rather than as broadly applicable.

Study Findings

Overall Average Gains
Academic skills:

- 0.44 SD for receptive vocabulary
- 0.62 SD for letter-word identification
- 0.59 SD for applied math problems
- 0.50 SD for numeracy and geometry

Executive function and emotional development:

- 0.24 SD for working memory
- 0.21 SD for cognitive inhibitory control
- 0.28 SD for attention shifting
- 0.19 SD for emotion recognition

Gains Across Demographic Groups
Children eligible for free or reduced-price lunch benefited more:

- 0.66 versus 0.47 SD for applied math problems
- 0.34 versus –0.01 SD for cognitive inhibitory control
- 0.33 versus 0.03 SD for attention shifting

Asian children benefited more than white children:

- 0.62 versus 0.22 SD for receptive vocabulary
- 0.49 versus 0.00 SD for letter-word identification
- 1.04 versus 0.40 SD for applied math problems
- 0.76 versus 0.29 SD for cognitive inhibitory control
- 0.50 versus 0.01 SD for attention shifting

Hispanic children benefited more than white children:

- 0.50 versus 0.22 SD for receptive vocabulary
- 0.88 versus 0.00 SD for letter-word identification
- 0.70 versus 0.40 SD for applied math problems
- 0.51 versus 0.29 SD for cognitive inhibitory control
- 0.31 versus 0.01 SD for attention shifting

Black children benefited more than white children:

- 0.36 versus 0.22 SD for receptive vocabulary
- 0.68 versus 0.00 SD for letter-word identification
- 0.46 versus 0.40 SD for applied math problems
- 0.33 versus 0.29 SD for cognitive inhibitory control
- 0.19 versus 0.01 SD for attention shifting

In addition, three other aspects of the study should be kept in mind when evaluating the program's results. First, because children were tested well into the fall, the gains reported include gains from weeks or even months of kindergarten on top of those made in pre-K. Second, all children were tested in English, although 50 percent of them lived in non-English speaking homes. The treatment-group children had been exposed to a year of English before being tested. But some children from the control group may not yet have been exposed to English and were therefore being tested in a language that they did not speak, which could have led them to score especially poorly. Finally, like all RDDs, the study is unable to address whether and to what degree gains measured in kindergarten are sustained long term.

Chicago Child-Parent Center Program

Founded in Chicago in 1967, the Child-Parent Center program (CPC) provides: a preschool program for economically disadvantaged three- and four-year-olds living in high-poverty neighborhoods; family support services beginning in preschool; a kindergarten program; and early elementary school intervention for children in first through third grades. The program seeks to promote students' academic success, social competence, economic self-sufficiency, and overall health. In addition to Chicago, the program has recently expanded to other sites in Illinois, Minnesota, and Wisconsin.

In Chicago, the program is run by the Chicago Public Schools. It operates in 19 sites and enrolls just over 2,000 students in both a half-day (2.5 to 3 hours) and full-day (7 hours) program. Classes run 5 days a week for a 180-day school year. For the study described in the next section, all children participated in three years of CPC: two preschool years and one kindergarten year for a total of 1,350 to 3,780 program hours, depending on whether children attended half- or full-day programs. CPC also provides health and social services for children, including health screenings, nursing services, speech therapy, and free breakfast and lunch.

Teachers must have a bachelor's degree and certification in early childhood education, and they are paid the same salary and benefits as Chicago public school teachers. The CPC class size is 17 children for a half-day classroom with a teacher and teacher assistant and 20 students for a full day with a teacher and teacher assistant.

Each CPC is staffed by a team that includes a head teacher, a parent-resource teacher, and a school-community representative. Head teachers coach other teachers, coordinate the curriculum, and provide professional development. Parent-resource teachers provide parent workshops and other health, safety, and nutrition supports. The school-community representative recruits children from CPC neighborhoods for the program, refers families to community and social services agencies, and provides home visits. CPC especially emphasizes parent engagement: two and a half hours of parental involvement are required every week in either in-school or at-home activities.

Study Description. An ongoing propensity score matching study has compared a group of children that attended CPC preschool with a group that did not, to investigate the long-term effects of children's participation in three consecutive years of CPC (two years of preschool and one year of kindergarten).

The first group was composed of 989 children whose parents enrolled them in CPC for three years—beginning at age three—and who completed the CPC kindergarten program in 1986. The second group was composed of 550 children who did not attend CPC preschool: 374 of those children attended kindergarten in non-CPC schools, and 176 attended kindergarten in CPC schools but did not attend the preschool program. These two groups were matched on age, neighborhood, socioeconomic status, and eligibility for government-funded early childhood programs.

The study used data from the Chicago Longitudinal Study, which tracks 1,539 low-income, minority children (93 percent black and 7 percent Hispanic) who completed public school kindergarten in the spring of 1986. Data were first collected in 1985, and the program group has been followed for more than 20 years. The most recent results are from a follow-up with the two groups when they were 26 years old, which included about 90 percent of the original study participants.

The Bottom Line. The CPC study found that children who attended CPC for three years (two preschool years and one kindergarten year) had better long-term outcomes than children who did not attend CPC preschool. Those outcomes included reduced child maltreatment, less need for special education, lower levels of depression, reduced crime and delinquency, and reduced dependency on welfare.

The CPC study's matching design allowed researchers to examine long-term academic and social outcomes for participating children. But two limitations of the study must be kept in mind when evaluating its results.

First, 10 percent of study participants had dropped out of the study before the follow-up at age 26. It is not known why they dropped out or whether they are different in important ways from those who remained in the study. If the 10 percent excluded from the study is doing better than the other 90 percent, the study will

Study Findings

Improved Academic Outcomes

- 49 percent fewer years of special education by age 18 (0.73 versus 1.43 years)
- Almost 10 percent higher high school completion by age 26 (80 versus 73 percent)

Reduced Crime and Delinquency

- 28 percent fewer felony arrests by age 26 (13 versus 18 percent)
- 40 percent fewer violent arrests by age 20 (9 versus 15 percent)
- 19 percent lower incarceration rate by age 23 (21 versus 26 percent)
- 32 percent fewer arrests of any type by age 20 (17 versus 25 percent)

Reduced Dependency on Welfare

- 12 percent fewer months of receiving any form of public aid at age 23 (an average of 28.3 versus 32.1 months)
- 15 percent higher rate of health insurance coverage at age 26 (77 versus 67 percent)

Improved Emotional and Physical Outcomes

- 24 percent lower rate of depressive symptoms from ages 22 to 24 (13 versus 17 percent)
- 50 percent fewer children who experienced child maltreatment from ages 4 to 17 (10 versus 5 percent)

understate the gains from the program. If that 10 percent is doing worse, the study will overstate the gains from the program.

Second, the study was unable to take into account the degree to which children whose parents send them to preschool and agree to spend a minimum of two and a half hours per week may vary from children whose parents do not. A CPS study published in 1995 reported that parents who had sent their children to CPC preschool were considerably more likely to be involved with their child's schooling through elementary school and that children with more involved parents performed better than children with less involved parents.[20] Researchers do not know whether greater parental involvement was caused by the CPC preschool or whether more involved parents were more likely to send their children to CPC preschool in the first place. So it is not clear whether the most important factor in children's longer-term outcomes was their participation in pre-K or the kind of parents they have.

Georgia Pre-K

Georgia's pre-K program is a state-funded, universal, public preschool program free to all four-year-olds in the state. In 2014–15, 60 percent of the state's four-year-olds were enrolled in the program for a total of just over 80,000 children.

Georgia's pre-K is a mixed delivery program: as of 2015, approximately 47 percent of children were served in public school classrooms, and the other 53 percent attended programs in private centers and Head Start agencies that contract with local boards of education. The program runs on the public school schedule of 6.5 hours per day, 5 days per week, for a 180-day school year (a total of 1,170 program hours).

Lead teachers must have a state teaching license and a bachelor's degree in early childhood or a related field, and classroom assistants must have either a Child Development Associate (CDA) credential or another valid credential as approved by the state. The minimum teacher-child ratio is 1-to-11 with a maximum class size of 22 children.

Study Description. In 2011, the Georgia legislature funded a series of studies to evaluate the program. Two studies were conducted by University of North Carolina–Chapel Hill researchers from 2011 to 2013, and a third study, running from 2013 to 2018, is currently in progress.

The second of these recent studies, carried out in 2012–13, used an RDD design to assess how program participation affects children's kindergarten readiness skills and whether program effects vary across student subgroups. The study included a total of 1,181 children: 611 had completed the Georgia pre-K program in 2011–12 and were beginning kindergarten in the fall of 2012, and 570 children were beginning pre-K in fall 2012 because they had missed the age cutoff for the previous year. All 1,181 children were assessed between September 21 and December 21, 2012, in language/literacy, math, general knowledge, and behavior skills.

The Bottom Line. The study of Georgia's pre-K program showed that participating children scored higher on several measures of math and literacy when tested in kindergarten. Impacts were very large for literacy and moderate to large in math.

Study Findings

Improved Academic Skills

- Children in the program group scored significantly higher on several tests of math and literacy skills, with the largest effect sizes as follows:
  - 1.20 SD for phonemic awareness
  - 1.05 SD for letter-word identification
  - 0.89 SD for letter knowledge (naming letters)
  - 0.86 SD for counting
  - 0.51 SD for math problem solving

Improved Social-Emotional Development

- Teachers rated the program group as having more positive interactions in the classroom than the control group by 0.23 SD.

These results are impressive, but three limitations of the study are important to keep in mind. First, the study did not report attrition rates, so we do not know how many children dropped out of pre-K over the course of the year, whether those children differ in significant ways from those who successfully completed the program, and how that might have affected study findings.

Second, although Georgia's school year starts in early August, testing did not begin until September 21 and continued until the end of December. That means that the large gains attributed to the pre-K program are not from the pre-K year alone but also reflect additional gains made over several weeks or months of attending kindergarten. Because gains from the pre-K and kindergarten programs are conflated, it is not possible to determine what proportion of those gains are due to the pre-K program specifically.

Finally, as with all RDDs, we do not know how the study's findings of higher scores on academic skills tests in the first few months of kindergarten translate into important, longer-term outcomes.

Head Start

Head Start is a federally funded preschool program for economically disadvantaged three- and four-year-olds that provides educational, social, medical, dental, nutritional, and other services for low-income young children and their families. It aims to promote school readiness by advancing children's academic, social, emotional, and physical development and by improving parenting practices and family economic stability. The program was launched in 1965 as part of President Lyndon B. Johnson's War on Poverty, and today it serves almost 900,000 children in roughly 18,000 Head Start centers across the country.

Children can attend the program for either one year (entering at age four) or two years (entering at age three). Forty-six percent of Head Start children attend full-day programs for at least six hours a day, four or five days a week; 54 percent attend part-time programs. Local agencies are required by the federal Head Start Act to use research-based curricula, and programs must have child-to-teacher ratios of under 10 children per teacher. As of 2015, 73 percent of Head Start teachers had a bachelor's degree or higher, and nearly all had at least an associate's degree.

Study Description. The 1998 Congressional Head Start reauthorization mandated that the Department of Health and Human Services conduct a national RCT evaluating Head Start's impact on children who attend the program. The subsequent Head Start Impact Study assessed the effect of one year of Head Start participation on all four program goals: academic skills, social-emotional development, health, and parenting practices. The study used testing along with parent and teacher reports to examine these outcomes for two groups of children: one cohort that entered the program at age three and one cohort that entered at age four.

The study sample included 4,667 three- and four-year-old children applying for entry to one of a randomly selected, nationally representative sample of 383 Head Start centers across 23 states. About 2,600 children were randomly assigned to a Head Start group, and about 1,800 children were randomly assigned to a non–Head Start control group. Children in the control group could remain at home or attend other child care or preschool programs chosen by their parents. Data collection ran from fall 2002 through 2008, following children from their application to Head Start through the spring of their third-grade year.

The Bottom Line. Overall, the Head Start Impact Study (HSIS) showed small positive effects on reading and math test scores at the end of the preschool year, but few positive effects persisted into elementary school. The study found almost no impacts on the three-year-old or four-year-old Head Start cohorts at the end of kindergarten in any of the four domains of academic skills, social-emotional development, health, and parenting practices. Not surprisingly, when tested again at the end of third grade, the Head Start children fared no better than the "no Head Start" control group.

In interpreting these results, it is important to note that some children in the "no Head Start" control group did in fact attend other Head Start programs outside the study. In the first year of the study, about 14 percent of the four-year-old control group and 18 percent of the three-year-old control group attended a Head Start program; in the second year, nearly half of the three-year-old control group (then age four) attended a Head Start program outside of the study.

The implication of this is that the Head Start group was compared with a group of children that was not entirely "no Head Start" because some children from the control group also attended Head Start programs. This probably caused study findings to understate Head Start's effects. If the study had been conducted without these flaws, it is possible that somewhat larger short-term effects of Head Start might have been found; it is not clear to what degree such effects would have been sustained over the longer term.

Flaws aside, a crucial aspect of the HSIS findings is that they tell us only the *average* impact of Head Start centers across the country, and those aggregate findings obscure two important variables. First, both the quality and the effectiveness of the nation's roughly 18,000 Head Start centers vary greatly. Using the Early

Study Findings

Academic Skills

Three-Year-Old Cohort

- At the end of the Head Start year, the three-year-old cohort had the following gains:
  - 0.24 SD for letter naming;
  - 0.26 SD for letter-word identification;
  - 0.22 SD for preacademic skills; and
  - 0.15 SD for math applied problems.
- At the end of kindergarten, the Head Start group had slightly less math ability (–0.19 SD), as assessed by teachers.
- At the end of third grade, the Head Start group performed no better than the control group.

Four-Year-Old Cohort

- At the end of the Head Start year, the four-year-old cohort had the following gains:
  - 0.25 SD for letter naming;
  - 0.22 SD for letter-word identification; and
  - 0.19 SD for preacademic skills.
- At the end of third grade, children in the Head Start group had a slight advantage in reading (0.11 SD), with no other effects compared to the control group.

Social-Emotional Development

Three-Year-Old Cohort

- During the Head Start year, the three-year-old cohort showed slightly less hyperactive behavior (–0.21 SD) and less overall problem behavior (–0.14 SD), as reported by parents.
- At the end of third grade, the cohort had slightly greater social skills and positive approach to learning (0.12 SD), as reported by parents.
- No teacher-reported effects were found during the Head Start year or later.

Four-Year-Old Cohort

- No differences were found for the Head Start or kindergarten years.
- At the end of first grade, the Head Start group was slightly less withdrawn (–0.13 SD), as reported by parents. At the same time, teachers reported that they were slightly more socially reticent (0.19 SD) and had more problems with teacher interactions (0.13 SD).
- At the end of third grade, the Head Start group had several small negative outcomes on teacher-reported measures of social-emotional development:
  - –0.13 SD in closeness with teacher;
  - –0.14 SD in positive teacher-child relationships; and
  - –0.24 SD in expression of negative versus positive emotions.

Parenting Practices

Three-Year-Old Cohort

- During the Head Start year, parents in the Head Start group were slightly less likely to spank their child (–0.14 SD), more likely to read to their child (0.15 SD), and more likely to engage their child in "cultural enrichment activities" (0.18 SD).
- At the end of third grade, parents in the Head Start group were slightly more likely (0.16 SD) to use a positive parenting style (characterized by greater warmth and control).

Four-Year-Old Cohort

- At the end of third grade, parents in the Head Start group spent more time with their children (0.27 SD).

Childhood Environment Rating Scale (ECERS), a 2013 National Institute for Early Education study found that 40 percent of Head Start centers were high quality, 57 percent were medium quality, and 3 percent were low quality. Centers with predominantly African American children were found to be considerably worse: 7 percent were low quality and only 26 percent were high quality.[21] Researchers have

also found that some centers are much more effective at producing sustained positive outcomes, although the key drivers of effectiveness are not clear.[22]

Second, there is considerable evidence that some children benefit more from Head Start than others. For example, the HSIS found that high-risk children in the three-year-old cohort showed sustained cognitive benefits through the end of third grade. In other words, some children can benefit a great deal from some centers—but we do not know enough about which children or centers those are. This highlights a crucial limitation of the RCT: while it is the most rigorous research approach, an RCT often assesses only average impact, leaving unanswered the important question of what works best for whom.

Finally, the HSIS findings highlight the challenge of scaling quality. In theory, all Head Start centers across the country could be effective. However, in practice, they are not. Just like the public K–12 school system has shown us over decades, implementing high-quality education across the country—or even across a single city—is much easier said than done.

Nurse-Family Partnership

The Nurse-Family Partnership (NFP) is a nonprofit organization that supports the delivery of home visits by registered nurses to young, first-time, low-income mothers. NFP nurses establish a long-term relationship with expecting mothers throughout their pregnancy and the first two years of the child's life, with visits typically scheduled with mothers at home every other week during pregnancy, weekly for the first six weeks after the baby's birth, and then every other week until the child turns two.

The program has three main objectives: (1) to improve the outcomes of pregnancy by helping women enhance their health-related behaviors during pregnancy; (2) to improve the child's subsequent health and development by teaching parenting skills; and (3) to promote family planning and stability, maternal educational achievement, and self-sufficiency. The Nurse-Family Partnership began as a pilot program in Elmira, New York, in the 1970s and today serves approximately 30,000 women in 43 states, the US Virgin Islands, and 6 tribal communities.

Study Description. Beginning in the 1970s, NFP conducted three RCT studies, each in a different location with a distinct population, to assess the outcomes of the participating mothers and children. The three studies collectively included 2,273 low-income women and took place in Elmira, New York, from 1978 to 1982; Memphis, Tennessee, from 1990 to 1993; and Denver, Colorado, from 1994 to 1997.

The Elmira study included 400 women: 92 percent were white, 60 percent were low income, 60 percent were unmarried, and their average age was 19.

The Memphis study included 1,138 women: 92 percent were African American, 85 percent came from households at or below the poverty line, 98 percent were unmarried, and their average age was 18.

The Denver study included 735 women: 46 percent were Mexican American, 36 percent were white, 15 percent were African American, almost 100 percent were low income, 84 percent were unmarried, and their average age was 20.

In each study, the women were assigned either to a group that participated in the NFP home-visiting program or to a control group that did not receive home visits but received developmental screenings and, in two studies, transportation to medical appointments.

The Bottom Line. These three studies have shown sizable, sustained effects on child and maternal outcomes across three different program sites. The studies have been exceptionally rigorous; the program is one of a handful that has met the congressionally defined standard of "Top Tier Evidence."[23]

The strong evidence for NFP underscores the important impact that high-quality early childhood programs can have on long-term outcomes for disadvantaged children and their parents. But NFP explicitly focuses on young, low-income mothers from pregnancy through the first two years of their babies' lives, so its results cannot inform debates about the potential effectiveness of pre-K.

Study Findings

Elmira

- Effects on children of nurse-visited women when they reached adolescence:
  - 48 percent fewer verified incidents of child abuse and neglect by age 15 (average of 0.26 incidents per child in the nurse-visited group versus 0.50 incidents in the control group)
  - 57 percent less likely to have been convicted for criminal activity by age 19 (12 versus 28 percent)

- Effects on unmarried, low-income nurse-visited women when their children reached age 15:
  - 33 percent less time spent on welfare (average of 60 versus 90 months)
  - 31 percent fewer subsequent births (average of 1.1 versus 1.6 births)
  - 82 percent fewer arrests (average of 0.16 versus 0.90 arrests)
  - 81 percent fewer convictions (average of 0.13 versus 0.69 convictions)

Memphis

- Effects on children of nurse-visited women at age two:
  - 23 percent fewer health care encounters for injuries or dangerous ingestions (average of 0.43 incidents per child in the nurse-visited group versus 0.56 in the control group)
  - 78 percent fewer days hospitalized for injuries or dangerous ingestions (average of 0.04 versus 0.18 days)

- Effects on children of nurse-visited women at age 12:
  - 29 percent less likely to have psychological problems such as depression or anxiety (22 versus 31 percent)

Denver

- Effects on the subsample of children at age four whose mothers had low psychological resources before program participation:
  - 0.31 SD in language development
  - 0.38 SD in behavioral adaptation (e.g., attention, impulse control, and sociability)
  - 0.47 SD in executive functioning

- Effects on nurse-visited women when their children reached age four:
  - 20 percent longer interval between the women's first and second births (average of 24.5 versus 20.4 months)
  - 50 percent fewer women experienced domestic violence from their partner in the prior six months (7 versus 14 percent)

Oklahoma Pre-K

Oklahoma's pre-K program is a state-funded, universal preschool program free to all four-year-olds in the state. In 2014–15, 76 percent of the state's four-year-olds were enrolled in the program for a total of more than 40,000 children.

The program is run by the public schools and operates on the regular school schedule: 5 days per week for a 180-day school year. The half-day program runs for 2.5 hours per day, and the full-day program runs for 6 hours per day (a total of 450 hours for the part-time program and 1,080 hours for the full-time program).

Lead teachers must have a bachelor's degree, must be certified in early childhood education, and are members of the regular public school teaching force. Most classrooms have an assistant teacher who does not have to meet specific education or training requirements. The minimum teacher-child ratio is 1-to-10 with a maximum class size of 20.

Study Description. A series of studies using data from Tulsa have examined the impact of Oklahoma's program on children's kindergarten readiness. The researchers chose Tulsa for the study location because at the time of the evaluation it was the largest school district in the state, had a racially and ethnically diverse student body, and routinely tested four- and five-year-olds at the same point early in the school year.

In one of the most commonly referenced Tulsa studies, researchers used an RDD to compare a group of 3,727 children who had completed the pre-K program in 2002–03 and were entering kindergarten in the fall

Study Findings

Improved Academic Skills

- Overall, program participants had the following gains:
  - 0.79 SD for letter-word identification
  - 0.64 SD for spelling
  - 0.38 SD for applied math problems

Gains Across Demographic Groups

- Hispanic children:
  - 1.50 SD for letter-word identification
  - 0.98 SD for spelling
  - 0.99 SD for applied math problems
- Black children:
  - 0.74 SD for letter-word identification
  - 0.52 SD for spelling
  - 0.38 SD for applied math problems
- Native American children:
  - 0.89 SD for letter-word identification
  - 0.72 SD for spelling
  - 0.60 SD for applied math problems
- White children:
  - 0.76 SD for letter-word identification
  - 0.72 SD for spelling

Gains Across Socioeconomic Backgrounds

- Children receiving reduced-price lunch:
  - 1.04 SD for letter-word identification
  - 0.97 SD for spelling
- Children receiving free lunch:
  - 0.81 SD for letter-word identification
  - 0.65 SD for spelling
  - 0.45 SD for applied math problems
- Children receiving full-price lunch:
  - 0.63 SD for letter-word identification
  - 0.54 SD for spelling
  - 0.29 SD for applied math problems

of 2003 with a group of 1,843 children who were just beginning the pre-K program that fall because they had missed the age cutoff for the previous year. Of those totals, 3,149 (84.5 percent) of the kindergarten students and 1,567 (85 percent) of the pre-K students were tested in basic literacy and math skills.

Tests were administered in English by pre-K and kindergarten classroom teachers who had been trained in test administration for the project. The researchers report that testing was conducted for the most part during the first week of school.

The Bottom Line. The Oklahoma pre-K program is one of the most commonly cited state programs, highlighted as a model for the Obama administration's "Preschool for All" proposal. The Tulsa study found moderate to large effects on participating children's academic skill test scores when they were entering kindergarten. Overall, kindergarten students who had attended the Tulsa pre-K program were found to be six to eight months ahead of their peers on tests of basic literacy and math skills. The study showed especially large effects on Hispanic and low-income students' test scores.

In interpreting these findings, it is important to remember that they are limited to children living in Tulsa and whose parents chose to enroll them in the program. Those children may not be representative of all eligible children across the state. In addition, although researchers did not report attrition data, they noted that the group tested at kindergarten entry was significantly different from the group just beginning pre-K, which suggests that children from some backgrounds may have been more likely than others to drop out of the program during the year.

Further, all children were tested in English. If children from the control group had not yet been exposed to English, they were being tested in a language they did not speak, which could have led them to score especially poorly. Finally, like all RDDs, the Tulsa study cannot determine the degree to which short-term outcomes in basic academic skills predict longer-term academic, social, and economic outcomes.

The Perry Preschool Program

The Perry Preschool Program was a small preschool and home-visiting program for three- and four-year-olds, run as a pilot project from 1962 to 1967 in Ypsilanti, Michigan. It was a single-site model program that provided 58 low-income, high-risk, African American children with two years of the research-based High-Scope preschool curriculum combined with weekly home visiting, beginning when they were three. The program's goal was to improve education outcomes for children living in poverty and at risk of school failure in order to give them the skills necessary to succeed in school and life.

The program included both a morning preschool for children and a weekly home-visiting component targeting their mothers. Children attended the preschool program for 2.5 hours a day, 5 days per week, for 8 months (October to May), for a total of about 900 hours over 2 years. The program curriculum was based on active learning, meaning that children were encouraged to develop and learn from their own activities. In addition, teachers visited each child's home for 1.5 hours every week, providing ongoing coaching to mothers to help them carry out the program curriculum at home.

The program had a total of four teachers—certified in elementary education, early childhood education, and special education—who together served 20 to 25 children each school year. Teacher-child ratios were high to enable teachers to conduct the weekly home visits to all participating children.

Study Description. The Perry Preschool Study was conducted as an RCT. Fifty-eight children were assigned to the program group and 65 children to a control group that did not participate in the program, for a combined total of 123 children. All children included in the study were African American with IQs between 70 and 85. Of the participating families, 45 percent were headed by a single parent, 58 percent were receiving welfare benefits, and in 47 percent neither parent was employed.

Children were assessed annually from ages 3 through 11, and follow-up studies of participants have

Study Findings

Improved Academic Outcomes

- 1.3 fewer years in special education services (average of 3.9 versus 5.2 years)
- 31 percent more likely to have completed high school or received a GED by age 27 (71 versus 54 percent)
- Females were almost one and a half times more likely to have completed high school or received a GED by age 27 (84 versus 35 percent)

Improved Family Planning and Stability Among Females at Age 27

- 83 percent fewer reported abortions (4 versus 23 percent)
- 31 percent less likely to be single parents (57 versus 83 percent)
- Four times more likely to be married (40 versus 8 percent)

Improved Economic Status at Age 27

- 42 percent of males in the program had monthly earnings of at least $2,000 as compared to 6 percent for the non-program group
- 45 percent more females in the program were employed (80 versus 55 percent)
- Those in the program were more than one and a half times more likely to own their home (36 versus 13 percent)

Lower Crime at Age 40

- 46 percent less likely to have been in jail or prison (28 versus 52 percent)
- 33 percent less likely to have been arrested for violent crimes (32 versus 48 percent)
- 72 percent less likely to have been arrested for drug dealing (7 versus 25 percent)

been conducted at ages 14, 15, 19, 21, 27, and 40 comparing program participants to the control group. A follow-up at age 50 is currently underway.

The Bottom Line. The ongoing Perry Preschool Study has found important positive impacts on participants' educational and life outcomes decades later, such as increased rates of high school graduation and marriage; higher earnings; and reductions in crime, teen pregnancy, and out-of-wedlock births. The study shows that a high-quality early childhood program can significantly improve the lives of poor children over the long term.

Because of these impressive findings for a program described as "preschool," Perry is often cited in support of expanding public pre-K. Yet the program differs from proposed large-scale pre-K programs in two important ways. First, like Abecedarian, Perry was a small program run in a single site by those who designed it. Teacher-child ratios were very high—five children or fewer per teacher—and the program ran for two years. It is unclear if, or how, the program's quality and intensity could be scaled.

Second, the program had two key components: a preschool program and weekly home visit. Given the crucial role that parents play in their child's development, there is no reason to expect that preschool alone would have the same kind of impact. Indeed, NFP shows that it is possible to achieve big impacts with home visiting alone, so we do not know which of Perry's two major program components caused its strong results. It seems likely, in fact, that the program's impressive impact was caused by the two components operating in combination.

The bottom line is that replicating Perry's results is likely to require replicating Perry's program, not a lower-cost, lower-quality version that lacks one of the model's crucial components.

Tennessee Voluntary Pre-K Program

The Tennessee Voluntary Pre-K program (TN-VPK) is a targeted state-funded program for the state's neediest four-year-olds that seeks to develop children's school readiness skills. Launched in 2005, the program uses a tiered admission process that gives low-income children priority in admission. Today, the program serves 22 percent of the state's four-year-olds, for a total of 18,000 children annually.

The program runs on the public school schedule: 5.5 hours per day, 5 days per week, for a 180-day school year (a total of 990 program hours). About 93 percent of VPK classrooms are in public schools; the remaining classrooms are in Head Start centers or other private programs.

Teachers must have a state teaching license in early childhood development and education. Classroom teaching assistants must have a CDA credential or be working toward one. Staff are paid the same salary and benefits as public school teachers. The minimum teacher-child ratio is 1-to-10 with a maximum class size of 20 students.

Study Description. In 2009, Vanderbilt's Peabody Research Institute, in conjunction with the Tennessee Department of Education, launched the TN-VPK Effectiveness Study: a five-year RCT to investigate the effects of program participation as children progress through elementary school. To date, the TN-VPK study is the only randomized control trial of a scaled-up, state-funded pre-K program.

The Effectiveness Study randomly placed children who applied to TN-VPK into two groups: those who were given places in the program (the program group) and those who were not admitted (the control group). This procedure was used for more than 3,000 children across TN-VPK applicants for the 2009–10 and 2010–11 school years. Both the children who participated in TN-VPK and those who did not are being tracked through Tennessee's education database, and information on their school performance is being collected each year. State achievement test data for all 3,000 children were collected in late fall of 2015, and a report on those findings is forthcoming.

The study also included an "Intensive Substudy": a subsample of 1,076 children (773 who were randomly assigned to a VPK classroom and 303 who were not admitted) who have been directly assessed by the research team from the beginning of the pre-K year through the end of third grade using standardized language, literacy, and math achievement tests. Researchers have reported a high degree of variability in test timing and an average lag time of about two and a half months into the pre-K year before obtaining baseline pretest assessments for both VPK children and the control group.

The most recent study report, released in September 2015, presents the effects of TN-VPK on children's academic achievement and behavioral outcomes from pre-K through third grade for the Intensive Substudy group. More than 90 percent of the study sample has remained in the study across the four years.

The Bottom Line. The TN-VPK study shows that children who attended Tennessee's pre-K program made significant gains on tests of basic math and literacy skills by the end of their pre-K year. But the study subsequently found that children who did not attend pre-K made bigger gains in kindergarten and, as a result, by the end of the kindergarten year, the non-VPK children had caught up to the VPK children. By the end of second grade, the VPK children were actually scoring slightly lower than the non-VPK group on most measures.

The Intensive Substudy has an important flaw to keep in mind. The school district required that researchers obtain parental consent for children to participate in the substudy, which was requested after randomly placing children into VPK and non-VPK groups. Researchers were unable to obtain consent for a large number of children (58 percent of the first cohort and 29 percent of the second cohort) because of administrative complications. Those children were therefore withdrawn from the substudy, leaving 1,076 children—36 percent of the study's full sample of children—who were representative of the full original sample but were not randomly selected from it.

Because the substudy group was not randomly selected, RCT methods were no longer valid.

Study Findings

At the End of Pre-K

- By the end of the pre-K year, VPK children scored higher than the control group on the composite literacy, language, and math achievement measure and on each of six achievement subtests, with the largest effects as follows:
  - 0.32 SD for the composite literacy, language, and math achievement measure
  - 0.41 SD for letter-word identification
  - 0.29 SD for spelling
  - 0.27 SD for quantitative concepts
- Teachers rated the VPK group as more prepared for kindergarten (0.22 SD).
- Effects were much larger for children who were English language learners (ELL) and had less educated mothers:
  - 0.88 SD for ELL children whose mothers had less than a high school degree
  - 0.55 SD for ELL children whose mothers had at least a high school degree
  - 0.22 SD for native English speakers whose mothers had at least a high school degree

At the End of Kindergarten

- By the end of the year there were no longer significant differences between the two groups on any achievement measures because the control group children made greater gains in kindergarten than the VPK children.

At the End of First Grade

- No significant differences between the program and control groups were observed on the six achievement measures.
- First-grade teachers rated the VPK children lower on several behavioral measures:
  - Lower preparedness for grade (–0.17 SD)
  - Poorer work skills in the classrooms (–0.20 SD)
  - More negative about school (–0.21 SD)

At the End of Second and Third Grade

- VPK children scored lower than the control children on the composite achievement measure (–0.15 SD at the end of second grade and –0.13 SD at the end of third grade).
- Teachers reported no differences on behavioral measures at the end of second grade. At the end of third grade, they reported slightly more positive peer relations (0.19 SD) and slightly fewer behavior problems (–0.16 SD) among VPK children.

Researchers therefore had to use quasi-experimental methods to evaluate program results for children in the substudy, yielding the findings reported. State achievement test data for the full sample are accessible without parental consent, however, and became available for the first time in late fall of 2015. Analysis of that new data is a crucial part of the larger study and may yield different results than those found to date.

The TN-VPK study findings have important implications for interpreting results from other pre-K studies. Like the RDD studies of other pre-K programs described in this paper, children who participated in TN-VPK showed big gains after a year of pre-K. But, unlike those studies, the TN-VPK study followed children for several years and found that those initial gains were not sustained. The TN-VPK study findings do not show that gains measured at kindergarten entry are never sustained. But they do show that early gains cannot be taken as conclusive evidence that desirable longer-term outcomes will follow.

Finally, the study does not prove that no children benefited from Tennessee's pre-K program—the results reported are an average across all children, and it is reasonable to assume some children benefited. Nor does the study prove that pre-K *cannot* benefit children. It does, however, underscore the need for more rigorous pre-K research: to examine the relationship between short-term and longer-term outcomes, to better understand which children benefit most from pre-K, and to determine what program designs yield sustained positive effects for whom.

Conclusion

Each of the 10 programs described in this paper is called an early childhood program and serves children under five. However, what is most notable about these programs is not how similar they are but how much they differ in both design and results. Some of the 10 focused on four-year-olds, some on three-year-olds, and some solely on infants and toddlers. Some ran for just one year, others for two, and one served children from infancy to kindergarten. Some were school-based while others were home-based. Some targeted children alone while some targeted their families too. Some increased the number of alphabet letters children knew when they were five; others led to large increases in social, economic, and health outcomes decades later.

The research conducted on these programs also varied greatly. Researchers used different methods to investigate a range of questions: some evaluated basic academic skills in kindergarten, some examined children's performance in elementary school, and still others tracked a broad spectrum of effects into adulthood. Some studies were more rigorous than others.

The information provided by this body of research is less useful than commonly assumed. The research shows neither that "pre-K works" nor that it does not; rather, it shows that some early childhood programs yield particular outcomes, sometimes, for some children. It shows that early childhood programs *can* have a significant, sustained impact on the lives of children born into disadvantaged circumstances. But it falls far short of showing that all programs have that impact.

The most meaningful, far-reaching effects occurred with intensive, carefully designed, well-implemented programs that target infants and toddlers and include a strong focus on parents: Abecedarian, Nurse-Family Partnership, and Perry. Yet we still do not know nearly enough about how and why these exceptionally effective programs had the impact they did. In fact, what this research makes clearest is not what we *do* know, but rather what we do not.

Overall, two important policy implications emerge. To move the early childhood field forward, we must:

1. Strengthen and accelerate early childhood research; and

2. Advance high-quality child care and home visiting programs for disadvantaged children.

Strengthen and Accelerate Rigorous Research in Early Childhood

The early childhood research base is often characterized as rigorous and extensive, and it indeed includes hundreds of studies published over the last several years.[24] Yet a close look reveals that both the relevance and rigor of this research is considerably weaker than most realize.

Advocates emphasize that research overwhelmingly shows positive results from pre-K. But almost all published research shows positive results because studies with null results rarely get published. Research methods used are often less rigorous than acknowledged, as discussed in Part I, and few pre-K studies are replicated to test the real strength of their findings.

Further, the positive results these studies report, while significant from a *statistical* point of view, often have minor importance from a *policy* point of view. Many studies include only children whose parents send them to pre-K—excluding the vast majority of children under age five, who are at home or in child care. Many measure only rudimentary academic skills—such as recognizing letters, knowing how to hold a book

right-side-up, and performing basic counting—in the first months of kindergarten.

Pre-K advocates often claim that small gains in these basic kindergarten skills lead to large gains in children's cognitive and noncognitive capacities, which in turn lead to graduating from high school and staying out of prison. But this is a little-tested assumption that should not drive important policy decisions until it has been carefully investigated.

In short, while we have scores of studies that examine the impact of conventional pre-K, the core policy question remains unanswered: what are the most effective early interventions for improving disadvantaged children's lives?

To guide policy effectively, early childhood research must be strengthened in three ways. First, it must focus on the most important questions instead of the most fashionable or convenient ones. Second, researchers must increase data transparency and replication of prior work. Third, greater investment in new approaches to rigorous, policy-relevant research is needed.

Focus on the Important Questions. To build a knowledge base that can move the field forward, research must:

- *Measure what's important rather than what's easy.* Pre-K studies primarily assess children's short-term gains in basic academic skills, such as identifying letters of the alphabet, recognizing vocabulary words, and counting small numbers. Those academic gains are simply assumed to be a proxy for other, more important capacities that are harder to measure, such as language and executive function skills, reasoning, critical thinking, problem solving, persistence, and the ability to get along well with others. But researchers must explicitly investigate which early capacities are linked to long-term success and figure out how to measure whether children have acquired them.

- *Focus on long-term, not short-term, impacts.* Longitudinal studies on early childhood interventions are needed. How children who attend a pre-K program fare in kindergarten is not what matters—and the prevailing assumption that kindergarten achievement test scores are a sufficient proxy for long-term cognitive and noncognitive effects must be carefully tested. Research also often shows "fade-out" of early academic gains, and the implication of that phenomenon for children's long-term outcomes must be rigorously examined.

- *Investigate what works for which children, when, and how.* We need more precise information than whether Program X "works" or not. What child outcomes are the most important? What interventions work best for which children to affect those outcomes? What is the optimum age for intervention? What are the "active ingredients" of successful programs, how do they work, and why? Answers to these questions are crucial to designing effective and efficient programs.

- *Dive into the "black box" of program quality.* Our thinking about quality is often circular: we describe a program as "high quality" when it produces results—and producing results is how a high-quality program is defined in the first place. But early childhood programs are complex, with many moving parts. What drives quality, how to measure quality, and how to ensure quality in an early childhood setting has largely remained a "black box." While the field has taken initial steps to improve measures of quality, we need much better knowledge on *what* specific program inputs and practices are linked to *which* outcomes for children. We cannot invest in—or improve—quality when we do not understand what it is.

- *Investigate how to effectively implement programs at scale.* Ineffective scale-ups of small, effective programs become large, ineffective programs. And the most easily scalable program components may not be the ones that are most important to a program's impact. We need a much better understanding of *how* good programs can successfully be taken to scale and how effective we can expect those programs to be.

Increase Research Transparency and Replication. Like all academic research in policy-relevant fields, the nature and quality of early childhood research suffers from a disconnect between the interests of researchers on one hand and those of policymakers and the public on the other. An academic researcher's professional success depends on appearing in academic journals, not on figuring out what works best for children. And academic journals are strongly biased toward publishing studies with both positive and novel findings.

Just how bad is that bias? A 2010 analysis of research articles of almost 11,000 social science journals showed that almost 90 percent of the published studies reported positive results.[25] And most of these findings are never replicated because journals overwhelmingly publish new studies rather than replications of previous ones. A 2014 study of the entire publication history of the top 100 education journals found that 43 published no replications at all, and only six journals had a replication rate of more than 1 percent.[26] Without rigorous replication, published findings remain untested, and false findings go unchallenged.

Furthermore, researchers often share neither their data nor important methodological details, making it impossible for colleagues to verify and replicate their findings. The early childhood research field must improve recording and sharing of research data and methods, enabling researchers to test and build on prior work.[27] In fact, the US Institute for Education Sciences (IES) now requires that researchers publicly share data collected for IES-funded research. Increased transparency is essential to advancing the accurate, relevant knowledge needed for effective policymaking.

Invest in New Research Approaches. Because academic journals are biased toward research that shows positive impacts, researchers are incentivized to design studies that remain on familiar terrain and are likely to show statistically significant results. At the same time, they are discouraged from investigating promising but uncertain new approaches. In early childhood, these incentives create too many studies of pre-K's impact on academic skills in kindergarten and too few on other interventions—some of which may be much more effective and efficient.

In other words, while studying pre-K may be the best path to success for academic researchers, it is not the best path to building policy-relevant knowledge for the early childhood field. We need broader, riskier research that is focused on generating and testing new ideas, grounded in the best science of early development.

For example, the Frontiers of Innovation project at the Harvard Center for the Developing Child uses micro-trials to test new, scientifically developed interventions with small numbers of participants, quickly modifying those interventions based on initial evidence and then testing them again.[28] This kind of rapid-cycle approach can experiment with promising new ideas: taking risks, sharing results early, and learning quickly from ideas that do not work.

The federal government is uniquely positioned to advance stronger research in early childhood. Rather than spending tens of billions of dollars to scale up unproven programs, the federal government can contribute most effectively by helping build the knowledge base needed to support future investment. The government should:

- *Launch a research program in early childhood to promote innovation and experimentation.* A federal Early Learning Research Program—modeled on the successful Small Business Innovation research program for technology—could be funded with a small percentage of agency budgets to support the development and rigorous testing of innovative interventions. Setting aside just 1 percent of the roughly $20 billion annual federal expenditures on early learning and child care would provide a yearly budget of $200 million to investigate promising new approaches.[29]

- *Establish an online knowledge clearinghouse*. An online Federal Clearinghouse on Early Learning could disseminate evidence on existing initiatives and share ideas and best practices. This would promote transparency and knowledge sharing and would spark new thinking on how to advance children's early learning and development.

Advance Voluntary High-Quality Child Care and Home Visiting

Our most important policy goal in early childhood is to improve life outcomes for disadvantaged children. While expanding school-based pre-K is currently the primary focus in early childhood policy, advancing high-quality, educational child care and more effective parenting are the most practical and promising avenues to accomplish that goal.

Both existing program research and the best science point us in this direction. Of the five programs described in this paper that were rigorously evaluated by RCTs, three showed big, long-term impacts: Abecedarian, a five-year, full-time, educational child care program; Nurse-Family Partnership, a two-and-a-half-year home-visiting program targeting infants, toddlers, and their mothers; and Perry, a two-year program combining ongoing parent coaching with a high-quality preschool for three- and four-year-olds that stressed both cognitive and noncognitive development.[30]

The findings for these three programs are consistent with our scientific knowledge of early childhood development. We know that the most important period of children's development spans from conception through age two. We know that parents and early environments play by far the most crucial role in shaping a child's development. We know that both noncognitive and cognitive skills are essential to children's success. So it makes a great deal of sense that programs that target very young children, engage parents, and teach a broad range of skills are likely to have the largest impact on children's long-term school and life outcomes.

At the same time, we do *not* know whether school-based pre-K programs actually affect the outcomes that really matter. A growing body of pre-K research shows academic skill gains at the beginning of kindergarten, such as knowing more letters of the alphabet or being better able to count. But we have no idea whether those academic gains are associated with the range of competencies that are crucial to children's later life success. The claim that short-term gains predict long-term effects has not been systematically investigated, and the current evidence base on that question remains weak.

In addition, while research studies overwhelmingly show that pre-K has statistically significant positive results, this information is less meaningful than it seems at face value. A small increase in children's test scores is often described as "significant" in an academic study without being at all significant from a policy point of view. Pre-K is in the political spotlight and is a popular focus for research. But policymakers must look carefully at specific research findings when making high-stakes policy decisions.

The key to improving outcomes for at-risk children is to enrich their earliest environments. Those environments are primarily home—where they live—and for many children, child care, where they are placed while their parents work to support them. Given what we know and where young children spend most of their time, it makes sense to direct new investment toward voluntary home visiting to shore up vulnerable families and high-quality, educational care for disadvantaged children. This does not require setting up new institutions because homes and child care already exist. It simply requires improving the environments where children already spend the first years of their lives.

Concluding Thoughts

So does pre-K work? We don't know—and it is the wrong question to be asking in the first place. Because resources are limited, we are faced with tough decisions about the best use of public funds. The crucial policy question is not if pre-K is a *good* thing, but whether it is the *right* thing to address the greatest needs of our nation's most vulnerable children.

The leading science and strongest research to date indicate that the clearest avenue to help disadvantaged children is not to send them to school a year earlier but to improve child care and support parents in better fulfilling their role as their children's first teachers. Our current knowledge is insufficient to justify a large expansion of pre-K as the best path forward. And the growing pre-K push may well do more harm than good by diverting attention and scarce resources from other more effective approaches.

A stronger knowledge base is urgently needed to guide policy. We need innovative, rigorous research directed at the key policy question: what early interventions can substantially improve children's lives? An answer to that question, not whether pre-K can increase children's skills in kindergarten, is what the field needs to move forward.

Early childhood is gathering public and political momentum as one of the most important domestic policy areas of our time. But what America's most disadvantaged children are facing is not an achievement gap; it's a life gap. To close that gap, we must move beyond a narrow focus on improving academic skills as the aim and expanding pre-K as the solution. Researchers, policymakers, and the public must remain focused on the core goal: to give all children, no matter the circumstances of their birth, a fair start in life.

Glossary

Attrition Bias/Selective Attrition
These terms refer to the possibility that the children who drop out of a program before completing it may be different in important ways from those who stay in the program. In other words, attrition from study groups can be non-random and associated with important, unobserved variables that bias study findings in ways researchers cannot account for.

For example, children who have more supportive parents or a greater ability to deal with new situations may be more likely to successfully complete a pre-K program than those who do not. If this is the case, a study's results will reflect both a pre-K program's effects and other differences between children, such as their support system at home or their adaptability. If children who stick with the program are different from children who drop out, that will bias study results, because those results will reflect differences between children in addition to the impact of attending pre-K.

Confounding Variable
A confounding variable is something that is not measured in a given study but that affects the result. For example, an early childhood program such as Perry Preschool has a preschool component and a home-visiting component. In a study assessing the impact of just the preschool program, the home-visiting component would be described as a "confounding variable," because the relative effects of the preschool and home-visiting components cannot be disentangled.

Counterfactual
A counterfactual is what would have occured if participants were not given a particular intervention. For example, in the absence of a public pre-K program, the counterfactual for some children might be staying home with a family member. For others it might be spending weekdays with a neighbor, and for others it might be attending a private preschool their parents pay for.

High Risk/At Risk
Children's risk for future academic and social problems is assessed using several indicators of socioeconomic status and family stability. The most common indicators are: parent educational levels, family income, absence of the father from the home, use of welfare, parental unemployment, and birth to a teenage mother. A child with zero to two of these indicators would usually be considered low or no-risk. A child with three risk factors would be considered moderate risk, and a child with four to five factors would be considered high risk.

Other risk factors that are sometimes considered include residential instability (families moving one or more times in the last year), living in a household with no adult English speakers, and living in a family with four or more children.

Independent Versus Dependent Variables
Research studies often aim to determine cause-and-effect relationships: for example, a study might investigate the question: "Does attending pre-K cause children to do better in elementary school?" The independent variable is the cause that is being investigated (in this example, whether or not children attend pre-K) to see what impact it has on the dependent variable (performance in elementary school).

Any study can have multiple independent and dependent variables, although a particular study can seldom examine all the variables that may be at play. For example, attending pre-K might have several other effects (dependent variables) on children besides their performance in first grade, such as self-esteem or their feelings toward school. In addition to pre-K attendance, independent variables could also include the

quality of a child's specific classroom and the number of his absences during the year. Those independent variables may interact to affect children's performance in first grade.

Researchers can use statistical models to assess the relative impact of multiple independent variables, but no study can look at all possible variables. Consequently, the resulting findings may oversimplify a complex phenomenon and miss crucial causes of a particular effect. Sometimes a different variable (such as having an especially engaged mother) is the primary cause of both the dependent variable (performance in first grade) *and* the independent variable (attending pre-K). In fact, a big challenge for social science researchers is that their independent variable might actually be the dependent variable of something else they are not even measuring.

Intention-to-Treat Versus Treatment-on-Treated

An intention-to-treat approach is used when researchers collect outcome data for all the children who were assigned to or at least started a program. For example, an RDD study using an intention-to-treat approach collects data on all children who enrolled in a pre-K program, whether or not they completed the program. In a treatment-on-treated approach, researchers collect data only for children who finish (or were "treated" by) the program being studied.

Main Effect

A main effect is the average impact that researchers determine is caused by the independent variable, ignoring the possible impact of other variables. For example, the Head Start Impact Study investigated the average impact of attending Head Start on children's later performance in elementary school, finding that Head Start had virtually no impact on their academic achievement.

At the same time, Head Start may have had a big impact on some children and no impact on others, depending on several other factors, such as the quality of the center attended, family background, and the specific needs of the child. But the Impact Study tells us only the main effect—that is, the average impact that Head Start had on all enrolled children.

Mixed Delivery System

A mixed delivery system is a public preschool program that is provided in a mix of private settings and public schools. In this model, private organizations—including nonprofit and for-profit centers, Head Start agencies, and community-based organizations—contract with the government to administer publicly funded preschool programs.

Observed Versus Unobserved Variables

Observed variables are those that researchers are able to identify and measure, such as children's age, race, family income, family structure, neighborhood, and achievement test scores. Unobserved variables are characteristics that vary among children but that researchers cannot measure—such as personality and motivation—and therefore cannot account for in their results.

Selection Bias

Selection bias refers to the fact that individuals self-select into a program, which may affect that program's outcomes. Because pre-K is voluntary, families decide whether or not to enroll their children, and families that enroll their children in pre-K may have different characteristics from families that do not.

Studies that compare children who went to pre-K with those who did not often find that children who attended pre-K do better in elementary school. But it is hard to know how much weight to give those findings, because it is impossible to know how much the improved outcomes should be attributed to attending pre-K and how much to having engaged, knowledgeable parents who chose to enroll their child in the program.

Single-Site Versus Scaled-Up Program

Single-site refers to an early childhood education program that is carried out at a single location. Both the Abecedarian Project and the Perry Preschool Program were single-site programs.

Scaled-up refers to a program that operates at multiple sites. There is no minimum number of sites for a program to be deemed scaled-up, but the term is conventionally used to describe systems that aim to

serve all eligible children across a city, state, or even the country. Head Start and Boston pre-K are examples of scaled-up programs.

Treatment Group Versus Control Group

The treatment group refers to the children in a study who receive a particular intervention, such as pre-K. The control group refers to the children who do not receive that intervention. Studies compare the outcomes of the treatment group with those of the control group to determine the intervention's impact.

Universal Versus Targeted Pre-K

A universal pre-K program is one that is free and available to all children who live in a particular geographic area, regardless of family income. The pre-K programs in Boston, Georgia, and Oklahoma are examples of universal programs.

A targeted pre-K program is one that is focused specifically on low-income children. Some targeted programs are open exclusively to low-income children (for example, Head Start); others also admit higher-income children, but only if there are spaces left over (for example, Tennessee's Voluntary Pre-K Program).

References

Barnett, W. S., M. E. Carolan, J. H. Squires, K. Clarke Brown, and M. Horowitz. *The State of Preschool 2014: State Preschool Yearbook*. National Institute for Early Education Research, 2015. http://nieer.org/sites/nieer/files/Yearbook2014_full3.pdf.

Barnett, W. Steven, Kwanghee Jung, M. Youn, and Ellen C. Frede. *Abbott Preschool Program Longitudinal Effects Study: Fifth Grade Follow-Up*. National Institute for Early Education Research, Rutgers University, 2013. http://nieer.org/sites/nieer/files/APPLES%205th%20Grade.pdf.

Campbell, Frances A. and Craig T. Ramey. "Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence: Positive Effects of Early Intervention." *American Educational Research Journal* 32, no. 4 (Winter 1995): 743–72.

Campbell, Frances A., Craig T. Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson. "Early Childhood Education: Young Adult Outcomes from the Abecedarian Project." *Applied Developmental Science* 6, no. 1 (2002): 42–57. https://www.researchgate.net/profile/Elizabeth_Pungello/publication/240519150_Early_Childhood_Education_Young_Adult_Outcomes_From_the_Abecedarian_Project/links/0f317534d44d7c4b2e000000.pdf.

Campbell, Frances A., Elizabeth P. Pungello, Margaret Burchinal, Kirsten Kainz, Yi Pan, Barbara H. Wasik, Oscar A. Barbarin, Joseph J. Sparling, and Craig T. Ramey. "Adult Outcomes as a Function of an Early Childhood Educational Program: An Abecedarian Project Follow-Up." *Developmental Psychology* 48, no. 4 (July 2012): 1033. http://static.vtc.vt.edu/media/documents/245_-_Adult_outcomes_as_a_function.pdf.

Campbell, Frances A., Gabriella Conti, James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, Elizabeth Pungello, and Yi Pan. "Early Childhood Investments Substantially Boost Adult Health." *Science* 343 (2014): 1478–85. http://home.uchicago.edu/rodrig/ScienceABC.pdf.

Coalition for Evidence-Based Policy. "Social Programs That Work: Nurse-Family Partnership." http://evidencebasedprograms.org/1366-2/nurse-family-partnership.

———. "Social Programs That Work: Perry Preschool Project." http://evidencebasedprograms.org/1366-2/65-2.

Eckenrode, John, Mary Campa, Dennis W. Luckey, Charles R. Henderson, Robert Cole, Harriet Kitzman, Elizabeth Anson, Kimberly Sidora-Arcoleo, Jane Powers, and David Olds. "Long-Term Effects of Prenatal and Infancy Nurse Home Visitation on the Life Course of Youths: 19-Year Follow-Up of a Randomized Trial." *Archives of Pediatrics & Adolescent Medicine* 164, no. 1 (May 2010): 9–15.

Fitzpatrick, Maria D. "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement." *BE Journal of Economic Analysis & Policy* 8, no. 1 (December 2008). http://www-siepr.stanford.edu/Papers/pdf/08-05.pdf.

Gormley Jr., William T., Ted Gayer, Deborah Phillips, and Brittany Dawson. "The Effects of Universal Pre-K on Cognitive Development." *Developmental Psychology* 41, no. 6 (November 2005): 872. http://204.14.132.173/pubs/journals/releases/dev-416872.pdf.

Kitzman, Harriet, David L. Olds, Charles R. Henderson, Carole Hanks, Robert Cole, Robert Tatelbaum, Kenneth M. McConnochie et al. "Effect of Prenatal and Infancy Home Visitation by Nurses on Pregnancy Outcomes, Childhood Injuries, and Repeated Childbearing: A Randomized Controlled Trial." *JAMA* 278, no. 8 (August 27, 1997): 644–52.

Kitzman, Harriet J., David L. Olds, Robert E. Cole, Carole A. Hanks, Elizabeth A. Anson, Kimberly J. Arcoleo, Dennis W. Luckey, Michael D. Knudtson, Charles R. Henderson, and John R. Holmberg. "Enduring Effects of Prenatal and Infancy Home Visiting by Nurses on Children: Follow-Up of a Randomized Trial Among Children at Age 12 Years." *Archives of Pediatrics & Adolescent Medicine* 164, no. 5 (May 2010): 412–18.

Lipsey, M. W., D. C. Farran, and K. G. Hofer. *A Randomized Control Trial of the Effects of a Statewide Voluntary Prekindergarten Program on Children's Skills and Behaviors Through Third Grade*. Peabody Research Institute, Vanderbilt University, 2015. http://peabody.vanderbilt.edu/research/pri/VPKthrough3rd_final_withcover.pdf.

Lipsey, Mark W., Kerry G. Hofer, Nianbo Dong, Dale C. Farran, and Carol Bilbrey. *Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow-Up Results from the Randomized Control Design*. Peabody Research Institute, 2013. https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf.

Luckey, Dennis W., David L. Olds, Weiming Zhang, Charles Henderson, Michael Knudtson John Eckenrode, Harriet Kitzman, Robert Cole, and Lisa Pettitt. *Revised Analysis of 15-Year Outcomes in the Elmira Trial of the Nurse-Family Partnership*. Prevention Research Center for Family and Child Health, University of Colorado Department of Pediatrics, 2008.

Olds, David, Charles R. Henderson Jr., Robert Cole, John Eckenrode, Harriet Kitzman, Dennis Luckey, Lisa Pettitt, Kimberly Sidora, Pamela Morris, and Jane Powers. "Long-Term Effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior: 15-Year Follow-Up of a Randomized Controlled Trial." *JAMA* 280, no. 14 (October 14, 1998): 1238–44.

Olds, David L., Charles R. Henderson, and Harriet Kitzman. "Does Prenatal and Infancy Nurse Home Visitation Have Enduring Effects on Qualities of Parental Caregiving and Child Health at 25 to 50 Months of Life?" *Pediatrics* 93, no. 1 (January 1994): 89–98. https://www.researchgate.net/profile/David_Olds3/publication/14934622_Does_prenatal_and_infancy_nurse_home_visitation_have_enduring_effects_on_qualities_of_parental_caregiving_and_child_health_at_25_to_50_months_of_life_Pediatrics_9389-98/links/54c7bf000cf289f0cecdb99c.pdf.

Olds, David L., Charles R. Henderson Jr., Robert Tatelbaum, and Robert Chamberlin. "Improving the Life-Course Development of Socially Disadvantaged Mothers: A Randomized Trial of Nurse Home Visitation." *American Journal of Public Health* 78, no. 11 (November 1988): 1436–45. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1350235/pdf/amjph00250-0050.pdf.

Olds, David L., John Eckenrode, Charles R. Henderson, Harriet Kitzman, Jane Powers, Robert Cole, Kimberly Sidora, Pamela Morris, Lisa M. Pettitt, and Dennis Luckey. "Long-Term Effects of Home Visitation on Maternal Life Course and Child Abuse and Neglect: Fifteen-Year Follow-Up of a Randomized Trial." *JAMA* 278, no. 8 (August 27, 1997): 637–43.

Olds, David L., JoAnn Robinson, Lisa Pettitt, Dennis W. Luckey, John Holmberg, Rosanna K. Ng, Kathy Isacks, Karen Sheff, and Charles R. Henderson. "Effects of Home Visits by Paraprofessionals and by Nurses: Age 4 Follow-Up Results of a Randomized Trial." *Pediatrics* 114, no. 6 (December 2004): 1560–68. https://www.researchgate.net/profile/David_Olds3/publication/8152273_Effects_of_home_visits_by_paraprofessionals_and_by_nurses_age_4_follow-up_results_of_a_randomized_trial/links/00b4952d4172498481000000.pdf.

Peisner-Feinberg, Ellen S., J. M. Schaaf, D. R. LaForett, L. M. Hildebrandt, and J. Sideris. *Effects of Georgia's Pre-K Program on Children's School Readiness Skills: Findings from the 2012–2013 Evaluation Study*. FPG Child Development Institute, University of North Carolina at Chapel Hill, 2014. http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/GAPreKEval_RDDReport%203-4-2014.pdf.

Promising Practices Network. "Child-Parent Centers." September 2008. http://www.promisingpractices.net/program.asp?programid=98.

Puma, Mike, Stephen Bell, Ronna Cook, Camilla Heid, Pam Broene, Frank Jenkins, Andrew Mashburn, and Jason Downer. *Third Grade Follow-Up to the Head Start Impact Study: Final Report. OPRE Report 2012-45*. Administration for Children & Families, 2012. https://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf.

Ramey, Craig T., Frances A. Campbell, Margaret Burchinal, Martie L. Skinner, David M. Gardner, and Sharon L. Ramey. "Persistent Effects of Early Childhood Education on High-Risk Children and Their Mothers." *Applied Developmental Science* 4, no. 1 (2000): 2–14. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.568.781&rep=rep1&type=pdf.

Reynolds, Arthur J. "One Year of Preschool Intervention or Two: Does It Matter?" *Early Childhood Research Quarterly* 10, no. 1 (1995): 1–31. http://www.sciencedirect.com/science/article/pii/0885200695900241.

Reynolds, Arthur J., Judy A. Temple, Barry A. B. White, Suh-Ruu Ou, and Dylan L. Robertson. "Age 26 Cost-Benefit Analysis of the Child-Parent Center Early Education Program." *Child Development* 82, no. 1 (January/February 2011): 379–404. https://www.researchgate.net/profile/Judy_Temple/publication/49807965_Age_26_Cost-Benefit_Analysis_of_the_Child-Parent_Center_Early_Education_Program/links/0fcfd5090294947649000000.pdf.

Reynolds, Arthur J., Judy A. Temple, Dylan L. Robertson, and Emily A. Mann. "Long-Term Effects of an Early Childhood Intervention on Educational Achievement and Juvenile Arrest: A 15-Year Follow-Up of Low-

Income Children in Public Schools." *JAMA* 285, no. 18 (May 9, 2001): 2339–46. http://www.precaution.org/lib/effects_of_early_intervention.010509.pdf.

———. "Age 21 Cost-Benefit Analysis of the Title I Chicago Child-Parent Centers." *Educational Evaluation and Policy Analysis* 24, no. 4 (December 2002): 267–303. https://www.researchgate.net/profile/Judy_Temple/publication/228541174_Age_21_cost-benefit_analysis_of_the_title_I_Chicago_Child-Parent_Centers/links/0fcfd5069faefc3e21000000.pdf.

Reynolds, Arthur J., Judy A. Temple, Suh-Ruu Ou, Dylan L. Robertson, Joshua P. Mersky, James W. Topitzes, and Michael D. Niles. "Effects of a School-Based, Early Childhood Intervention on Adult Health and Well-Being: A 19-Year Follow-Up of Low-Income Families." *Archives of Pediatrics & Adolescent Medicine* 161, no. 8 (August 2007): 730–39. http://www.kinderjugendgesundheit.at/uploads/low_income_01.pdf.

Schweinhart, Lawrence J. "Benefits, Costs, and Explanation of the High/Scope Perry Preschool Program." Paper presented at the Society for Research in Child Development, Tampa, Florida, April 26, 2003. http://www.highscope.org/file/Research/PerryProject/Perry-SRCD_2003.pdf.

———. *The High/Scope Perry Preschool Study Through Age 40: Summary, Conclusions, and Frequently Asked Questions*. High/Scope Educational Research Foundation, 2004. http://www.highscope.org/file/Research/PerryProject/specialsummary_rev2011_02_2.pdf.

———. "Historical Narrative." High Scope. http://www.highscope.org/Content.asp?ContentId=232.

Schweinhart, Lawrence J., Helen V. Barnes, and David P. Weikart. *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27.* Ypsilanti, Michigan: High/Scope Press, 1993.

Weiland, Christina, and Hirokazu Yoshikawa. "Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills." *Child Development* 84, no. 6 (November/December 2013): 2112–30. http://www.viriya.net/jabref/impacts_of_a_prekindergarten_program_on_childrens_mathematics_language_literacy_executive_function_and_emotional_skills.pdf.

Notes

1. Center for Public Education, "Pre-Kindergarten: What the Research Shows," March 2007, http://www.centerforpubliceducation.org/Main-Menu/Pre-kindergarten/Pre-Kindergarten/Pre-kindergarten-What-the-research-shows.html; Education Commission of the States, "50-State Review," January 2016, http://www.ecs.org/ec-content/uploads/01252016_Prek-K_Funding_report-4.pdf; National Institute for Early Education Research, *The State of Preschool: 2003 State Preschool Yearbook*, 17, http://nieer.org/sites/nieer/files/2003yearbook.pdf; Education Commission of the States, "50-State Review"; and National Center for Education Statistics, "Table 202.10. Enrollment of 3-, 4-, and 5-Year-Old Children in Preprimary Programs, by Age of Child, Level of Program, Control of Program, and Attendance Status: Selected Years, 1970 Through 2013," http://nces.ed.gov/programs/digest/d14/tables/dt14_202.10.asp.

2. First Five Years Fund, "2014 Poll Results: Research Summary," http://ffyf.org/resources/america-speaks-2014-poll-results-research-summary/.

3. Our focus in this paper is on program goals, design, and impact. The consideration of program costs relative to impact is essential to policymaking, but that is not the focus of this paper.

4. See, for example, Hirokazu Yoshikawa et al., *Investing in Our Future: The Evidence Base on Preschool Education*, Society for Research in Child Development and Foundation for Child Development, October 2013, http://fcd-us.org/sites/default/files/Evidence%20Base%20on%20Preschool%20Education%20FINAL.pdf; "What Constitutes Strong Evidence of a Program's Effectiveness?" https://www.whitehouse.gov/sites/default/files/omb/part/2004_program_eval.pdf; Sylvia M. Burwell, Cecilia Muñoz, John Holdren, and Alan Krueger to the Heads of Departments and Agencies, "Next Step in the Evidence and Innovation Agenda," Executive Office of the President, Office of Management and Budget, July 26, 2013, https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-17.pdf; and *What Works Clearinghouse: Procedures and Standards Handbook Version 3.0*, http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.

5. 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): 2112–30.

6. Ellen S. Peisner-Feinberg et al., *Effects of Georgia's Pre-K Program on Children's School Readiness Skills: Findings from the 2012–2013 Evaluation Study*, University of North Carolina–Chapel Hill Frank Porter Graham Child Development Institute, 2014, http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/GAPreKEval_RDDReport%203-4-2014.pdf.

7. Weiland and Yoshikawa, "Impacts of a Prekindergarten Program."

8. W. T. Gormley Jr. et al., "The Effects of Universal Pre-K on Cognitive Development," *Developmental Psychology* 41, no. 6 (2005): 872.

9. For a thoughtful discussion on how to address common methodological problems with RDD studies, see Mark W. Lipsey et al., "The Prekindergarten Age-Cutoff Regression-Discontinuity Design: Methodological Issues and Implications for Application," *Educational Evaluation and Policy Analysis* 37, no. 3 (September 2015): 296–313.

10. The Abbott Program study used both Propensity Score Matching and RDD designs.

11. Researchers call this "non-random attrition." For more on that term and several others used in this paper, see Glossary on p. 39.

12. In her evaluation, Fitzpatrick concludes: "In summary, estimates of the statewide effects of Universal Pre-K in Georgia generally indicate that the program's availability improved child outcomes by as much as one to three percent of a standard deviation but the use of appropriate control groups and methods of inference renders the estimated relationship statistically insignificant." See

Maria Donovan Fitzpatrick, "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement," *BE Journal of Economic Analysis & Policy* 8, no. 1 (2008): 1–38.

13. We identified these 10 programs through an informal poll of 40 early childhood experts, requesting their recommendations of 5 to 10 leading research studies in the field. Selecting the 10 programs to highlight in the paper was straightforward, because there was a great deal of consensus in the responses.

14. For readers who want to dig deeper, please see References on p. 42 to find the specific studies discussed.

15. See Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (New York, Lawrence Erlbaum Associates: 1988).

16. Greg J. Duncan and Katherine Magnuson, "Investing in Preschool Programs," *Journal of Economic Perspectives* 27, no. 2 (Spring 2013): 109–32, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318654/.

17. For a useful discussion of various approaches to evaluating effect sizes, see Mark W. Lipsey et al., *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*, National Center for Special Education Research, November 2012, https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf.

18. The level of statistical significance is expressed by what is called the "p-value," which describes the probability that a particular finding has occurred by chance. A p-value of 0.05 means that there is less than 5 percent probability that a finding was just chance; a p-value of 0.01 means that there is less than a 1 percent probability that a finding was chance; a p-value of 0.001 means that there is less than a one in a thousand probability that a finding was chance.

19. There are different methods for calculating IQ, but in general, scores in the neighborhood of 100 are considered average.

20. Arthur J. Reynolds, "One Year of Preschool Intervention or Two: Does It Matter?" *Early Childhood Research Quarterly* 10 (1995): 1–31.

21. There is disagreement over whether ECERS measures the important aspects of quality, but this study reveals big variation between centers nonetheless.

22. Christopher Walters, *Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start*, NBER Working Paper No. 20639, October 2014, http://www.nber.org/papers/w20639.pdf.

23. Coalition for Evidence-Based Policy, "Top Tier Evidence," http://toptierevidence.org/.

24. Lynn A. Karaly and Anamarie Auger, "Informing Investments in Preschool Quality and Access in Cincinnati," RAND Corporation, 2016, https://www.rand.org/content/dam/rand/pubs/research_reports/RR1400/RR1461/RAND_RR1461.pdf.

25. Fanelli Daniele, "'Positive' Results Increase Down the Hierarchy of the Sciences," *PLoS One* 5, no. 3 (April 7, 2010), http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010068.

26. M. C. Makel and J. A. Plucker, "Facts Are More Important Than Novelty: Replication in the Education Sciences," *Educational Researcher* 43, no. 6 (2014): 304–16.

27. New resources are available that can greatly strengthen research transparency and reproducibility. The Dataverse Project, based at Harvard University, enables researchers to share and analyze research data, and the Open Science Framework allows them to track and share their entire research process.

28. Center on the Developing Child, Frontiers of Innovation, Harvard University, http://developingchild.harvard.edu/innovation-application/frontiers-of-innovation/.

29. This figure is calculated using the most recent data on federal budget expenditures on Head Start, Tax Credits, the Child Care and Development Block Grant, the Child Care Mandatory and Matching Funds of the Child Care and Development Fund, and federal TANF funds to states used for child care. Grover J. "Russ" Whitehurst and Ellie Klein, "Do We Already Have Universal Preschool?," Brookings Institution, September 17, 2015, http://www.brookings.edu/research/papers/2015/09/17-do-we-already-have-universal-preschool-whitehurst-klein#ref25. The US Department of Education's Investing in Innovation (i3) program uses a tiered evidence framework to support three types of research: developing and testing innovative models, validating promising approaches, and evaluating scale-ups of programs with strong evidence of effectiveness. See US Department of Education, "Investing in Innovation (i3)," http://www.ed.gov/open/plan/investing-innovation-i3.

30. Chicago CPC, while evaluated with a less rigorous methodology, also showed big impacts, and it, too, included a strong focus on children's parents.

About the Authors

Katharine B. Stevens is the research fellow in early childhood at the American Enterprise Institute, where she focuses on the research, policy, and politics around early childhood; the role of early learning and development in expanding opportunity for low-income Americans; and the implementation challenges of rapidly growing early education initiatives, especially issues of program and teacher quality. Elizabeth English is a research assistant in education policy at the American Enterprise Institute.