SECTION 1, CHAPTER 2

WHAT ARE REASONABLE EXPECTATIONS FOR ECE PROGRAM EFFECTIVENESS?

By Jeanne Brooks-Gunn, Ph.D., Teachers College and College of Physicians and Surgeons, Columbia University

Sarah Lazzeroni, Teachers College, Columbia University

Early childhood education (ECE) programs in the U.S. have a long and rich history, as well as a robust evaluation literature. In fact, more well-designed evaluations have been conducted for ECE programs than for elementary or high school programs. Other chapters in this volume consider what we know about the quality of early childhood programs and child outcomes (Burchinal & Farran, Ch. 1), about instructional practices contributing to ECE quality (Farran, Ch. 4), and about how teacher training and professional development influence program quality (Pianta & Hamre, Ch. 5).

EVALUATING ECE EFFECTIVENESS

This chapter examines a slightly different but related topic: What are reasonable expectations for ECE program effectiveness? The overlap is evident in that asking about expectations raises questions about what is reasonable today given the state of ECE quality, as well as the variability in quality. In general, ECE program impacts are expected to be small-to-medium, but not large.[1] Our estimates are based on the current ECE evaluation literature (Elango, Garcia, Heckman, & Hojman, 2015; Love, Chazan-Cohen, Raikes, Brooks-Gunn, 2013; Marietta, 2010; Phillips, Gormley, & Anderson, 2016; Weiland & Yoshikawa, 2013; Yoshikawa, Weiland, Brooks-Gunn, 2016). We offer general, research-based estimates for ECE program effectiveness. We should see modest program effects for four-year-olds whose teachers receive continuous professional development, a BA or additional training, adequate wages, and training on well-defined curricula. Additionally, all ECE programs should offer full-day programming and strive for relatively low teacher turnover. Some programs should be expected to enhance child school readiness by at least one-sixth to one-third of a standard deviation (more on this metric below). These effects would be found in traditional evaluations (randomization to treatment or control); they would be most likely in communities that do not have preschool slots for all four-year-olds (i.e., where a significant proportion of children are being cared for by kith and kin or where there is an age-based cutoff for enrollment).

This effect size is most likely to be seen in measures of language, literacy, mathematics, cognition, and perhaps, executive function (EF), which encompasses attention, memory, and inhibition. Significant effects are not likely for general health or health care, as the vast majority of four-year-olds are relatively healthy and receive health care. By contrast, if ECE programs offer referrals to or are linked with dental care, we are likely to see effects (since most four-year-olds, especially those who are poor or near poor, don't receive dental care). We can't be sure whether ECE increases receipt of services for special needs, as two opposing counterfactuals exist. That is, if ECE improves language, literacy, and cognition, then the proportion of children classified as developmentally delayed would decrease; at the same time, ECE program staff are likely to identify children who could benefit from Individuals with

[1] Large effects are defined as differences of .40 or more of a standard deviation between a group receiving an ECE program and a group not receiving the program (control group), where each child is ideally put into one of the two groups by random assignment. Medium effects are defined as .25 to .40 of a standard deviation. A small effect would be between .15 and .25 of a standard deviation (yet statistically significant). For the purposes of this chapter, our expected range is between one-sixth and one-third of a standard deviation.

Disabilities Education Act (IDEA) services, while kith-and-kin caregivers are much less likely to have the knowledge or access to do so (Love, Chazan-Cohen, Raikes, & Brooks-Gunn, 2013). In any case, ECE aims to provide needed services to each child, which would favor the second outcome (providing IDEA services).

We chose one-third of a standard deviation or more based on the best ECE evaluation results to date; not all program evaluations achieve this, for a variety of reasons (Yoshikawa, Weiland, & Brooks-Gunn, 2016). Having a robust effect size is also important given the expected reduction in effect sizes throughout the elementary school years. Without additional services or improvements to early elementary school, the effect of ECE will fall to one-half of its initial size by the end of third or fourth grade. Therefore, an effect size of one-half will become one-quarter and an effect size of one-third will become one-sixth. Effect sizes that are lower than one-third are very unlikely to be sustained into the late elementary school years.

It is likely that we would see smaller declines if changes, some of which we list below, were made in early elementary school. Without such changes, sustained ECE effects will be very modest or not present at all. Sixteen years ago, one of us wrote an article titled "Do You Believe in Magic?," with a thesis that no matter how wonderful a preschool program might be, one year of even the highest-quality services is not enough for children to succeed (Brooks-Gunn, 2003). Improvements must be made in the quality and often the quantity of education at both the preschool and elementary school levels (not to mention middle and high school, but that is beyond the scope of this chapter). More time in education settings may also be necessary (for example, full-day pre-K and kindergarten and after-school and summer programs during elementary school).

Asking about reasonable expectations is especially important because almost three-quarters of adults are in favor of preschool programs (Jones, 2014). Most people appreciate the idea that an early start is one of the most effective approaches to helping children learn. In this sense, developmental psychologists and early childhood educators have been wildly successful. A few benefit-cost analyses—underscoring the message that earlier is better—have cemented this belief. Economists James Heckman and Lynn Karoly have provided compelling evidence of long-term effects (Cannon, Kilburn, Karoly, Mattox, Muchow, & Buenaventura, 2017; Heckman, 2006; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Karoly, 2016). But underneath all the kudos lies a concern about what we should really expect from a preschool program in terms of children's later well-being. Our success as educators and social scientists in communicating that an early start matters may have some unintended consequences. That is, expectations may outstrip results. Today's ECE programs, even those showing short-term effects of one-third of a standard deviation, are unlikely to generate a 14:1 or even a 7:1 benefit-cost ratio, as the Perry Preschool Program did (Heckman, 2006). We believe that a more reasonable goal would be a 1.5 to 1 or 2 to 1 ratio of benefits to costs[2] (Karoly, 2016; Kilburn & Karoly, 2008; Steuerle, & Jackson, 2016).

Policy scholars debate (a) what effect sizes mean in terms of school achievement, (b) how large effect sizes need to be to translate into long-term indicators of success, (c) which ECE programs can deliver changes large enough to make a difference later on, and (d) whether our expectations for large effects are reasonable. We examine how ECE is defined, what types of evaluation are appropriate, how effect sizes are measured, what child outcomes are typically examined and what the results say (with a focus on differential effects), and what the implications are for pre-K to third grade education (Brooks-Gunn, 2003; Camilli, Vargas, Ryan, & Barnett, 2010; Duncan & Magnuson, 2013; Garces, Thomas, & Currie, 2002; Gormley & Gayer, 2005; Hill, Gormley, & Adelstein, 2015; Love, Chazan-Cohen, Raikes, & Brooks-Gunn, 2013; Reynolds, Magnuson, & Ou, 2010; Yoshikawa et al., 2013).

DEFINING ECE PROGRAMS

In this chapter, early childhood education refers to programs that provide center-based education to children from one to five years of age. Center-based programs for children under one year, although they exist (the most notable being the Abecedarian Programs, Early Head Start programs, and the current Educare programs) (Yazejian, Bryant, Hans, Horm, St. Clair, File, & Burchinal, 2017), serve only a very small fraction of infants, given both the high cost of care in the first year of life and parental preferences. At five years old, most U.S. children enter kindergarten or at least become eligible for kindergarten. Currently, the vast majority of four-year-olds attend preschool, and the number of three-year-olds in preschool is rapidly rising: about 60% of four-year-olds (Rathbun, Zhang, & Snyder, 2016) and 43% of three-year-olds (Weiland & Yoshikawa, 2013) are enrolled in preschool, according to recent estimates (Yoshikawa, Weiland, & Brooks-Gunn, 2016). One- and two-year-olds are much less likely to attend preschool. Therefore, we focus on four-year-olds and, to a lesser extent, three-year-olds. (Most evaluations focus on four-year-olds, although they are beginning to include more three-year-olds, who are receiving ECE in increasing numbers.)

ECE programs have many goals. The primary goal is to envelop children in a learning milieu that provides opportunities to master age-appropriate social, emotional, linguistic, physical, and cognitive skills. A closely related focus is the relatively low levels of school readiness among some groups. Children whose parents have low education, low income, and/or are from minority ethnic groups are, on average, likely to enter kindergarten with lower skills than children from other backgrounds (Duncan & Magnuson, 2005; Reardon & Portilla, 2016). They are also less likely to receive high levels of learning stimulation at home (Brooks-Gunn, Markman-Pithers, & Rouse, 2016; Hoff, 2006; Hoff, 2012; Kalil, Ziol-Guest, Ryan, & Markowitz, 2016; Votruba-Drzal, 2003), in large part because of

[2] Sibling and county comparisons have been used to follow children into adulthood, in order to look at long-term sustained effects of ECE. A handful of the small-program evaluations have also done so (Abecedarian Project and the Perry Preschool Program) (Belfield, Nores, Barnett, & Schweinhart, 2006; Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Hill, Gormley, & Adelstein, 2015). The estimates of effect sizes from these two programs are frequently cited by ECE policymakers as well as by politicians (one mention being made by President Obama in a State of the Union address). Although impressive, these benefit-cost estimates are based on fewer than 150 individuals who were born in the 1960s and 1970s.

their parents' education, income, and/or cultural beliefs about parenting. Many ECE programs have been designed to enhance poor children's school readiness, Head Start being the most salient. Sometimes the focus is on the gaps or discrepancies between more and less disadvantaged children. However, these terms do not address the goal of raising skills in one group (gaps could also be closed by reducing skill development in more advantaged groups).

Another goal is to help children who speak a language other than English at home to become proficient in both English and their native language. Policy scholars disagree about whether education for English language learners (ELLs) should focus more on helping students become bilingual or on helping students become proficient in English as quickly and efficiently as possible (Barrow & Markman-Pithers, 2016). Depending on which objective they emphasize, educational programs for young ELLs are generally divided between programs that are taught in both English and another language and programs that are taught solely in English (Barrow & Markman-Pithers, 2016).

Higher labor force participation among women and more work hours have led to a need for safe, affordable, and educational child care, yet such care is not available to many… Our so-called polyglot system of early care and education is not conducive to supporting working parents.

Last, and often overlooked, is the need for quality care for young children whose parents work. The proportion of working mothers with children age five and younger is at an all-time high in the U.S. (Bureau of Labor Statistics, 2016a; Bureau of Labor Statistics, 2016b; Wen, Hetzner, Brooks-Gunn, 2019). About 70% of all mothers with children under 18 are in the labor force, including 64% of mothers with children between the ages of one and five years (Bureau of Labor Statistics, 2016a; Bureau of Labor Statistics, 2016b). Many mothers in the U.S. also return to work quite soon after giving birth—almost 60% are back at work within nine months, 26% within 2 months, and 7% within one month (Wen, Hetzner, & Brooks-Gunn, 2019). Working hours have also increased, by 35% in single-parent households with children under age 18 and by 16% in two-parent households with children under age 18. Higher labor force participation among women and more work hours have led to a need for safe, affordable, and educational child care, yet such care is not available to many (Chaudry, Morrissey, Weiland, & Yoshikawa 2017). Our so-called polyglot system of early care and education is not conducive to supporting working parents (Chaudry et al., 2017).

Although definitions vary, many use the term pre-K to refer to all early childhood educational programs (Brooks-Gunn, Markman-Pithers, & Rouse, 2016). Four categories of programs can be identified, depending on who administers the program and how it is funded. (Sometimes these lines are blurred since programs may be funded by more than one source and may be subject to multiple administrative rules; for example, see New York City's Pre-K for All program [Reid, Melvin, Kagan, & Brooks-Gunn, 2019]).

1. State or city pre-K programs are, for the most part, overseen by state or city education departments; they are often universal, although some may be targeted to low-income children (Friedman-Krauss, Barnett, Weisenfeld, Kasmin, DiCrecchio, & Horowitz, 2018).
2. Federally funded programs include Head Start and its younger sibling, Early Head Start. The U.S. Department of Health and Human Services administers these programs, which are targeted to families with income below the federal poverty threshold (with 10% of the Head Start children having special needs, as a mandated set-aside) (Elango, Garcia, Heckman, & Hojman, 2015).
3. Community programs include a panoply of not-for-profit programs. They may be subsidized by community organizations or by the Child Care Development Block Grant program, in which federal money is passed on to the states to subsidize child care costs for low-income working parents (Matthews, Schulman, Vogtman, Johnson-Staub, & Blank, 2015).
4. For-profit early childhood programs have not been studied extensively, although the few observations available suggest that their overall quality is lower than that of the other three categories (Burchinal, Nelson, Carlson, & Brooks-Gunn, 2008; Rathbun, Zhang, & Snyder, 2016).

In the 1960s through the 1980s, ECE programs were developed mostly for children whose parents had low incomes and/or low education. Children from such families were observed to be less prepared for kindergarten (academically and socially) than children from more advantaged backgrounds. In fact, gaps in language skills are seen as early as age two, and perhaps even earlier (Fryer & Levitt, 2013; Klebanov, Brooks-Gunn, McCarton, & McCormick, 1998). It was thought that children from educationally and economically disadvantaged households received fewer opportunities—in their families, neighborhoods, and child care settings—to develop early skills that predict literacy and numeracy (Blau, 2003; Johnson, Martin, & Brooks-Gunn, 2013; Noble, Houston, Brito, Bartsch, Kan, Kuperman, Akshoomoff, et al., 2015). Families who have low incomes or live in low-income neighborhoods are also constrained in their child care choices, due to both income and availability. These ECE programs were premised on the idea that an educationally oriented preschool would provide experiences that would reduce the gaps between economically disadvantaged kindergarteners and their more advantaged peers. Hence the term "Head Start," the goal of which was to level the playing field by enhancing the skills of poor preschoolers. Consequently, programs from this era targeted children from low-income backgrounds. Thus, almost all the program evaluations through the last century involved children from low-income families. Our knowledge about program efficacy, especially long-term efficacy, is based on poor and, to a lesser extent, minority children. As more universal state and local pre-K programs have been implemented, we've seen debates arise about whether programs are equally effective for children from more advantaged families.

CONSIDERING EVALUATION DESIGNS

Commonly used designs

About 80% of the evaluations of ECE programs focus on four-year-olds (Camilli, Vargas, Ryan, & Barnett, 2010; Yoshikawa, Weiland, & Brooks-Gunn, 2016). Almost all the evaluations have been based on random assignment to a treatment or a control group. (A few well-known evaluations were not experimental—children were not randomized, and no data were collected prior to the treatment. The Chicago Parent-Child Program is the notable example [Reynolds, Temple, Robertson, & Mann, 2001].) These traditional evaluations are useful because they compare two equivalent groups of randomly assigned children. Therefore, any effects are unlikely to be due to unobserved differences between the two groups.

A few other designs have been used to evaluate ECE programs. One is based on sibling comparisons (looking at adolescent or adult outcomes of siblings who did and did not go to Head Start, for example) (Currie & Thomas, 1995), based on the premise that such comparisons control for family differences to a large extent. A few clever comparisons have employed variation in how programs were rolled out in a set of counties that were similar in poverty status, some of which received funding and technical assistance to open Head Start centers and some of which did not; this approach is a variant of the regression discontinuity design (Ludwig & Miller, 2007).

But such designs have limitations. Since parents voluntarily choose to send their children to ECE programs, the sample does not include families whose parents are unaware of a program, are distrustful of sending their children to a program, have few ECE programs available in their neighborhoods, do not speak English, or are concerned about immigration or child welfare scrutiny, to name a few of the reasons parents don't send their children to ECE programs. Consequently, we don't know how well an intervention may fare with all children of a specific age group. (Although citywide universal pre-K programs alleviate this concern to some extent, even in these circumstances, not all children are served.) And until recently, evaluations have focused on relatively small programs, offered in either just one site or in just a handful of sites. The national Head Start Impact Study (begun in 2002, even though Head Start itself began in 1965), which used a waitlist design, was the first to look at treatment and control children in hundreds of Head Start centers.

Evaluations of small programs are influenced by the community in which they are conducted. From an evaluation perspective, the biggest concern is the availability and quality of other ECE programs. If most children in a control group are likely to attend a different ECE program, then the effect sizes will be smaller than in situations where children in a control group do not attend an ECE program (Zhai, Brooks-Gunn, & Waldfogel, 2011). The other design that has been used to evaluate ECE programs is regression discontinuity, typically comparing children whose birthdays are near the mandated age cutoff for pre-K on either side. That is, children who receive ECE because their

birthdays are just before the age cutoff are compared to those who do not receive ECE because their birthdays fall right after it. Boston and Tulsa have used this evaluation design (Gormley & Gayer, 2005; Yoshikawa et al., 2013).

The Counterfactual

The sibling and county comparisons also suffer from being based on ECE conditions almost 50 years ago. The sibling comparison analyses have tapped the Panel Study of Income Dynamics and the National Longitudinal Study of Youth–Child Supplement, which began in the late 1960s or the 1970s. The county comparison analyses were based on the first Head Start programs from the 1960s. Also, these studies focused on Head Start, which offers early childhood education only for children whose family incomes are at or below the poverty threshold. At the time, families with low income usually had no other options (few other programs were available in low-income neighborhoods, and even when other programs were available, families were often unable to afford them). Therefore, children who were not in Head Start were unlikely to be in other preschool programs or were in programs for only a few hours a day (see the ETS Head Start Evaluation from the 1970s as an example) (Lee, Brooks-Gunn, & Schnur, 1988; Lee, Brooks-Gunn, Schnur, & Liaw, 1990).

Today, children from low-income families have access not only to Head Start but also, in many cities and states, to universal pre-K programs, often run by or in collaboration with a department of education. Other partially subsidized programs also exist (some funded through the Child Care Development Block Grant). The two best-known, small-scale evaluations, the Perry Preschool and Abecedarian projects, also were initiated in the 1960s and 70s and also targeted poor children; very few of the children in the control groups received any other preschool experiences (Belfield, Nores, Barnett, & Schweinhart, 2006; Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Hill, Gormley, & Adelstein, 2015).

All of this suggests that the counterfactual for treatment today is different from what it was previously. If children in control groups are enrolled in other preschool programs, the counterfactual is no longer preschool versus no preschool; it is a particular program (Head Start, universal pre-K) versus whatever other programs exist in a particular community. The heterogeneity within the control group vis-à-vis preschool experiences is important to quantify, and several nonexperimental analyses have been conducted to address it. Our group has done analyses with the Infant Health and Development Program (IHDP), the Head Start Impact Study, the Fragile Families and Child Wellbeing Study, and the Early Childhood Longitudinal Study–Kindergarten Cohort and Birth Cohort (Hill, Waldfogel, & Brooks-Gunn, 2002; Hill, Brooks-Gunn, & Waldfogel, 2003; Lee, Zhai, Brooks-Gunn, Han, & Waldfogel, 2014; Lee, Zhai, Han, Brooks-Gunn, & Waldfogel, 2013; Lee, Brooks-Gunn, Han, Waldfogel, & Zhai, 2014; Lee, Han, Waldfogel, & Brooks-Gunn, 2018). In all cases, we find the largest effects of Head Start, pre-K, or Learning Games (IHDP) occur in comparisons with children who received only parental or relative care, as well as in comparisons with home-based family care and home-based care with a nonrelative. These comparisons are

more akin to the analyses from the 1960s and 70s. Such findings—and their consistency across data sets—suggest that the effect sizes seen in the past are unlikely in current evaluations because more children in a control group are receiving some sort of preschool. Interestingly, comparisons of children in preschool or Head Start against children in kith-and-kin care show effect sizes in the modest range. Comparisons to children receiving other preschool do not.

Consequently, it may be unreasonable to expect effect sizes today that are similar to those in the past if most children are now receiving some ECE at three and four years of age. This does not mean that preschool is ineffective.

These findings have at least two implications. First, some preschool is better for children than none (even if quality differs across programs), as researchers have demonstrated in nationally representative longitudinal studies (Duncan & Magnuson, 2005; Duncan & Magnuson, 2013; Lee, Brooks-Gunn, Schnur, & Liaw, 1990; Lee, Zhai, Brooks-Gunn, Han, & Waldfogel, 2014; Yoshikawa et al., 2013). Second, although specific programs that are believed to be of high quality are likely to be better than other programs presumably of lower quality, these differences will be smaller than what was seen in the past, given that the counterfactual is different (Duncan & Magnsuon, 2013). Consequently, it may be unreasonable to expect effect sizes today that are similar to those in the past if most children are now receiving some ECE at three and four years of age. This does not mean that preschool is ineffective. It just means that traditional evaluations of treatment and control will find smaller effect sizes, since most children in the control group are receiving some sort of preschool.

Alternative evaluation approaches

The evaluation approaches discussed above are often considered superior to others, but they do have limitations, the most serious having to do with external validity, generalization, and take-up. Other approaches include using district-wide achievement test scores to examine cohorts before and after a district-wide intervention is initiated (see the example of Montgomery County discussed below). Another is to employ much more short-term, small-scale interventions to test a particular innovation before implementing it on a broad scale, or even before a traditional randomized trial to test for efficacy. An example of this approach has been outlined by Fisher et al. (2016) and Shonkoff & Fisher (2013).

Yet another approach is to forgo assessment of children altogether and, instead, focus on documenting changes made on quality indicators (see Burchinal & Farran and Pianta, this volume). Of course, such an approach is based on a strong premise—that quality is associated with child outcomes and that increasing the former improves the latter. (For example, if child outcomes are enhanced only when a certain level of quality is obtained [threshold effect] or if only children who initially experience a very low-quality program are affected [baseline effect], then just documenting quality increases cannot be assumed to result in more school readiness). Indeed, the ECE evaluation field is still struggling with the question of how much and what types of quality improvement actually make a difference.

DESCRIBING EFFECT SIZES

A definition of effect sizes

Evaluations typically report findings in terms of effect sizes as a function of the standard deviation; evaluation research often defines large effects as two-fifths to one-half of a standard deviation (with an assessment normed to have a mean of 100 and a standard deviation of 15, the treatment group would have a 6- to 8-point advantage at the end of an intervention compared to the control group) (Barnett, 2008).

Effect sizes in everyday language

It is sometimes difficult for the public, policymakers, and educators to understand what an effect size means. For example, does an effect size of .40 on early indicators of achievement for low-income students mean they'll do better in elementary school, and how much better compared to high-income students? The same question, of course, could be asked for dual language learners or for minority students. Two approaches can help translate effect sizes into more concrete indicators. The first is to explain what might be seen in a classroom where low-income students' performance was one standard deviation below that of high-income students. As a heuristic, we are using the difference between students whose family incomes are in the bottom 10% and students whose family incomes are in the top 10% of the income distribution (Reardon, 2011). The following discussion is taken from Rock and Stenner (2005); they were comparing black and white students, not low-income and high-income students, but the general principle is the same. Based on a normal distribution (68% of scores will be within one standard deviation of the mean score, the difference between the peak of the distributions is one standard deviation, and the distributions for both groups are "normal"), the following estimates can be made:

> *First, randomly selecting one black child and one white child and comparing their scores will show the white child exceeding the black child 76% of the time and the black child exceeding the white child 24% of the time. Second, 84% of white children will perform better than the average black child, while 16% of black children will perform better than the average white child. Third, if a class that is evenly divided by race is divided into two equal-sized groups based on ability, then black students will compose roughly 70%, and whites 30%, of the students in the lower performing group. Fourth, if a school district chooses only the top-scoring 5% of students for "gifted" courses, such classes will have thirteen times more whites than blacks. Fifth, assume that a school district's student body mimics the national racial distribution (17% black, 83% white and other). The district chooses the lowest-scoring 5% of all students for a special needs program. Although 17% of the district's children are black, 72% of the special needs students will be black (pp. 26-27).*

If low-income students benefit from an ECE program with an effect size of one-half of a standard deviation, then the difference between low- and high-income children would be reduced by one-half (assuming that the low-income students, rather than both groups, received the treatment, or that the low-income students responded twice as much to the treatment as did the high-income students). The corresponding changes in the differences between the hypothetical students in an example like the one above would be very large. Even benefits of one-third of a standard deviation would be considered large.

Estimation of adult outcomes

Another approach is to take an effect size at the end of a preschool intervention and estimate the increase in the number of children graduating from high school or college or predict kindergarten achievement scores to high school achievement scores. Then the adolescent outcomes become the predictors for adult success (i.e., lifetime earnings). Brooks-Gunn, Magnuson, and Waldfogel (2009) used this estimation approach to see to what degree different effect sizes from preschool interventions are associated with gains in lifetime earnings. Card and Krueger (1996) used a similar procedure to estimate the long-term effects of reductions in elementary school class sizes, and Heckman et al. (2009) have done estimates using actual earnings data from the Perry Preschool Project. These estimates do not look at reducing the gap between groups of students, as the Rock and Stenner (2005) estimates do.

Differential effectiveness for poor and nonpoor children

The example used here is based on one of the goals of ECE, which is to improve school readiness for disadvantaged children (whose parents are poor, have little education, are immigrants, do not speak English well, or are from minority backgrounds), targeting health and emotional, literacy, and cognitive skills. Some ECE programs are taking a different approach, targeting an entire school district. If all four-year-olds receive quality ECE, the differences between advantaged and disadvantaged students are likely to be smaller, unless large differential benefits are seen among groups. That is, both advantaged and disadvantaged children will benefit (a rising tide lifts all boats). Remember that until very recently, ECE program evaluations have concentrated on groups likely to have lower rates of school readiness. Universal services may need to be evaluated differently, or at the very least, the possibility of not attenuating differences between groups needs to be explicitly addressed, and it is important to examine the specific mechanisms that lead to such differences. For example, in the Boston program, effects differ based on subgroup status—the program had higher effect sizes for low-income children than for higher-income children for numeracy, inhibitory control, and attention shifting (Weiland & Yoshikawa, 2013).

If all four-year-olds receive quality ECE, the differences between advantaged and disadvantaged students are likely to be smaller, unless large differential benefits are seen among groups.

Benefit-cost analyses

As another example, the varying estimates from benefit-cost analyses are confusing. Such estimates, of course, are based on myriad decisions (Steuerle & Jackson, 2016) on both the cost and the benefit sides of the equation. For example, benefit-cost estimates for the Perry Preschool Project range from 17:1 to 3:1, a huge range (and for the gender-linked estimates, the comparisons involve about 40 treatment and 40 control group boys and involve lower crime rates for the boys in the treatment group, meaning that the large benefit-to-cost ratios are based on about four fewer boys in the treatment group having been involved in a serious crime than in the control group) (Barnett, 1985; Belfield, Nores, Barnett, & Schweinhart, 2006; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010). And the chances that a preschool program today will result in even a 3:1 savings (let alone a 17:1 savings) are likely to be small, given the counterfactual. It may be time for those of us in ECE to manage expectations by making it clear that benefit-cost ratios are likely to be no greater than 2:1.

CHOOSING DEVELOPMENTAL OUTCOMES

What outcomes are preschool programs expected to influence, given that the goal is usually enhanced readiness for kindergarten and elementary school? School readiness is typically considered to encompass all facets of children's development—language and cognition, social and emotional development, physical growth and health, approaches to learning and persistence, enthusiasm, and motivation. Today, EF would be added as a separate facet, given its links to emotional and cognitive development (Raver & Blair, 2016). Educators and developmental psychologists may parse the domains a bit differently, yet they agree in looking at what they call the "whole child," rather than at just academic achievement. At the same time, most preschool programs privilege some domains over others, with language and cognitive development—as reflected in achievement test scores—being the most desirable (and measured) outcome. Whether the implicit move away from the whole child approach is merited, given what we know about development and learning, is an open question. Most practitioners and evaluators are calling for less reliance on achievement test scores, and efforts to measure other domains continue. The relatively recent addition of EF outcomes to evaluations is a good example, as EF is thought to be central for learning and achievement. At the same time, emphasis on physical development and health is waning. We suspect that such changes are driven in part by the increase in state and local pre-K programs that are primarily administered through or with education departments rather than health and human services departments. Most educators subscribe to the belief that children need to be healthy to learn most effectively, but most programs don't emphasize health, per se.

All educational programs focus on language and cognition. Often these are defined in terms of achievement rather than developmental outcomes. Most curricula and teacher training emphasize literacy, numeracy, and science skills appropriate for each age group. Generic curricula are most often used in preschool programs, especially in Head Start (for example, Creative Curriculum). Preschool literacy curricula have the most extensive history (although, perhaps surprisingly, they have not been subject to evaluation; see Snow & Matthews, 2016). Specific numeracy

curricula have been developed and evaluated extensively (Building Blocks being the notable example), but science curricula for preschool children are less refined (Clements & Sarama, 2016). New approaches to enhancing EF are also being evaluated (Raver & Blair, 2016). In addition, some emerging curricula integrate learning across domains.

Head Start explicitly includes physical skills and health outcomes among its goals, whereas state and local pre-K are less likely to do so. Head Start funds services like occupational and physical therapy, and it offers health checkups and referrals for various forms of health care (like dental care). Such services are not the purview of schools, so they are less likely to be funded in state pre-K programs (although children who qualify for IDEA presumably would be referred for occupational and physical therapy, if indicated). When Head Start began, children from low-income families were unlikely to get regular health care; one of Head Start's successes in the past century was ensuring that high proportions of children obtained such services. Today, with more children covered by Medicaid and CHIP, differences in receiving health care between children in Head Start and not in Head Start are quite small (the exception being children whose parents are immigrants, who are less likely to receive health care than children whose parents are not immigrants). However, Head Start today does make a large difference in dental care, which many low-income children don't receive. Links to diagnostic and screening services may also increase the likelihood of receiving special education services through IDEA. Given Head Start's mandate to set aside slots for children with special needs, it is likely that Head Start serves proportionately more such children than do state and local pre-K (Reid, Melvin, Kagan, & Brooks-Gunn, 2019).

Interestingly, evaluations of ECE programs almost always include indicators of disabilities and individualized education plans. But we know little about whether teachers who have special needs children in their classrooms have received appropriate training or whether they provide specific or modified instruction for these students (Hebbeler & Spiker, 2016), let alone the additional services that children are receiving through IDEA. Few evaluations assess activities of daily living, a common measure in health surveys. Nor do they measure common health problems, such as asthma, which if not controlled is linked to school absence (Currie, 2005). Evaluations also measure more general indicators of health, such as weight for height (the concern being overweight and obesity, not underweight), nutritional intake (usually general measures), and exercise patterns. Whether programs actually emphasize such health behaviors is not known (Head Start does so, although very little is known about how much attention any individual program gives to health) (Lee, Zhai, Han, Brooks-Gunn, & Waldfogel, 2013).

Evaluations also often assess emotional development, most often in terms of aggression and inattention, as it is believed that disruptive behaviors impair the learning of individual children and in the classroom as a whole (Georges, Brooks-Gunn, & Malone, 2012; Duncan et al., 2007). We know less about how teachers actually manage such behaviors (and about how they are trained to do so) than about how teachers provide instruction in literacy and numeracy (Raver & Blair, 2016; Raver, Jones, Li-Grining, Metzger, Smallwood, & Sardin, 2009). Even so, reducing aggressive and inattentive behaviors is seen as an outcome of ECE programs. Likewise, what educators call "approaches to

learning," or what psychologists term "motivation, enthusiasm, and persistence," are often measured. As in the case of aggression, what teachers actually do to enhance motivation has not been studied very well.

One takeaway from this brief discussion of preschool outcomes is that links are often tenuous between expectations for children's success or preparation for elementary school and what is known about curricula, teacher training, and even teacher behavior in the classroom. The notable exception is for literacy and numeracy achievement (Clements, & Sarama, 2016; Snow & Matthews, 2016). If we expect an ECE program to reduce aggression and inattention, enhance motivation and enthusiasm, promote healthy eating, increase EF, or decrease school absences due to illness, we will need to specify (and implement) classroom practices that explicitly target these outcomes.

If we expect an ECE program to reduce aggression and inattention, enhance motivation and enthusiasm, promote healthy eating, increase EF, or decrease school absences due to illness, we will need to specify (and implement) classroom practices that explicitly target these outcomes.

EXAMINING ECE EFFECT SIZES

Preschool's efficacy has been examined in over 120 evaluations (Brooks-Gunn, Markman-Pithers, & Rouse, 2016; Camilli, Vargas, Ryan, & Barnett, 2010; Yoshikawa, Weiland, & Brooks-Gunn, 2016). In general, evaluations report significant effects for four-year-olds. Recent evaluations show that preschool has positive effects in the short term on language, literacy, and math skills, with higher-quality programs showing the biggest effects (Yoshikawa, Weiland, & Brooks-Gunn, 2016). Some evidence suggests that preschool may have positive effects on socioemotional behaviors (e.g., decreased aggressive behavior), although the research in this area is not as definitive (Yoshikawa, Weiland, & Brooks-Gunn, 2016). But the range of effects is large. Even the early programs from the 1960s and 70s exhibited a range, although we usually emphasize the successful programs from that era (Brooks-Gunn & Hearn, 1982; Stipek, Franke, Clements, Farran, & Coburn, 2017). This state of affairs continues today; as examples, we have only to look at the Head Start Impact Study results (small effects at the end of the program with few effects sustained into elementary school) (U.S. Department of Health and Human Services, 2010) and the Tulsa Head Start results (large and sustained effects seen through middle school) (Gormley & Gayer, 2005; Phillips, Gormley, & Anderson, 2016). How do we interpret such disparate findings? Other authors in this volume focus on program quality and implementation (the two are difficult to separate), curricula, and teacher training and oversight. The composition of students in a classroom also matters (via a process economists often call heterogeneity of effects). Some groups, such as students with developmental disabilities and dual language learners, have not received enough attention regarding effective ways of teaching and including them in classrooms (Barrow & Markman-Pithers, 2016; Hebbeler & Spiker, 2016).

The evaluation literature is replete with examples of differential effectiveness across subgroups within a center, across types of centers, and even across centers under the same auspices. Such variation makes it difficult to say

what expectations may be reasonable for outcomes in different programs. We provide a few examples, making comparisons within and across centers.

Comparisons within centers

Within centers, comparisons have examined which subgroups benefit the most from ECE programs. Yoshikawa et al. (2013) looked at the effects of ECE programs on four sometimes overlapping subgroups: 1) poor and nonpoor children; 2) black, white, and Hispanic children; 3) dual language learners and children of immigrants; and 4) children with special needs/disabilities. Gaps in school readiness based on income and race/ethnicity appear as early as age two, when children from nonpoor families and white children perform better on measures of literacy and cognitive skills (Brooks-Gunn, Markman-Pithers, & Rouse, 2016; Garces, Thomas, & Currie, 2002; Snow & Matthews, 2016). Preschool enrollment is lower for minority children and children from low-income families than for white children and children from higher-income families, possibly contributing to this gap (Brooks-Gunn, Smith, Klebanov, Duncan, & Lee, 2003; Reardon & Portilla, 2016; Yoshikawa, Weiland, & Brooks-Gunn, 2016). However, preschool's positive effects in terms of literacy, math, and social-emotional skills may be most effective for children living in or near poverty (Yoshikawa et al., 2013).

Most early evaluations have examined ECE programs' effects on black children but not on other minority groups, and therefore they can't give full insight into differential ECE program effects based on race/ethnicity (Bassok, 2010; Yoshikawa et al., 2016). In response, recent evaluations of programs like Head Start, Tulsa Pre-K, and Boston Pre-K have made comparisons across racial groups. These programs showed positive effects for children of all racial/ethnic backgrounds, but the effects were highest for Hispanics at age three in Head Start and in both the Tulsa and Boston studies (Yoshikawa et al., 2016). Some studies found especially strong effects for minority children from low-income families (Love, Chazan-Cohen, Raikes, & Brooks-Gunn, 2013). Other studies found no racial differences for children living below the poverty line, but more benefits for black students than for whites or Hispanics among the nonpoor (Bassok, 2010).

Although research on ECE's effects on ELLs and children of immigrants is somewhat limited, some evidence suggests that ECE has positive effects on language development and cognitive skills for ELLs (Barrow & Markman-Pithers, 2016; Yoshikawa et al., 2016). Policy scholars debate whether language instruction should be conducted solely in English, or in a combination of English and children's first language, but Barrow and Markman-Pithers, (2016) find that the general quality of ECE programs may be more important than the language of instruction. Still, some evaluations show that dual language instruction does not hurt children's ability to learn English and may encourage bilingualism and even achievement overall (Hoff, 2012; Yoshikawa et al., 2016).

We also have few evaluations (especially randomized controlled trials) of ECE's impacts on children with disabilities (Hebbeler & Spiker, 2016). Head Start has shown positive effects on math and social-emotional skills for children

with disabilities, and Tulsa showed positive effects on their literacy skills (Gormley & Gayer, 2005; Yoshikawa et al., 2016). Some effective interventions include programs emphasizing language development and social-emotional development, which have been shown to be effective in promoting language/literacy skills and social skills, respectively (Raver & Blair, 2016). Additionally, specialized curricula and instructional strategies for children with disabilities have been shown to improve children's oral language, literacy, motor, and social skills (Hebbeler & Spiker, 2016). However, we need more evaluations of ECE's effectiveness for children with disabilities.

Comparisons across Head Start centers

Comparisons are also being made across Head Start centers. Head Start has clear and extensive standards, which might lead us to expect that variation in impacts from site to site might be small. Yet such differences exist. In one analysis, inter-center variation was found for language and literacy but not for mathematics (U.S. Department of Health and Human Services, 2010). One possible explanation is that Head Start teachers generally are not doing much in the way of math instruction (see Clements, Sarama, & Germeroth, 2016, for evidence that in general, pre-K teachers are not spending much time on math and that when they do, they focus on simple math skills). Low math skills among students across the board would be evidence that such an explanation is correct. Head Start teachers do focus on language and literacy; the differences in outcomes suggest that some teachers are more effective than others. However, we need to know more about what exactly teachers are doing in literacy instruction (Snow & Matthews, 2016).

Another (nonexperimental) analysis from the Head Start Impact Study suggests that full-day programs had larger effects than half-day programs, which is not surprising (Yoshikawa, Weiland, & Brooks-Gunn, 2016; Yoshikawa et al., 2013). What is perhaps surprising is that teacher education (BA), teacher training (teaching license), and student-teacher ratios were not associated with inter-center program impacts (Yoshikawa, Weiland, & Brooks-Gunn, 2016). Still, a new analysis by Morris et al. (2018) suggests that Head Start's positive impacts are more variable than impacts shown in previous analyses, such as the U.S. Department of Health and Human Services' Head Start Impact Study from 2010. Morris et al. (2018) found that the effect sizes of Head Start on enrollment and exposure to high-quality care varied by site, with standard deviations of 21.4 percentage points (any center care), 22.3 percentage points (Head Start care), and 28.4 percentage points (nonrelative care). This variation may be due to differences in state regulations and implementation, as well as to variation in child characteristics (e.g., pretest scores and dual language learners).

Comparisons across types of centers

Generally, children attending either Head Start, pre-K, or other center-based care performed better on academic-skill assessments than children in parental or relative care (Zhai, Waldfogel, & Brooks-Gunn, 2013), and recent studies have begun examining differences in the effects of different types of center-based programs. Children in Head Start performed better on reading and math assessments than children in parental care, pre-K, or other center-based care (ECLS-B data; Lee, Zhai, Brooks-Gunn, Han, & Waldfogel, 2014; Lee, Zhai, Han, Brooks-Gunn, & Waldfogel, 2013).

Additionally, children spent more hours in Head Start, on average, than children spent at non-Head Start centers (Lee, Zhai, Brooks-Gunn, Han, & Waldfogel, 2014). This increased exposure could be one of the mechanisms behind the finding that three- and four-year-olds attending Head Start fared better in classroom literacy and math instructional activities than children in non-Head Start centers (U.S. Department of Health and Human Services, 2010).

But other analyses conducted with data from the Fragile Families and Child Wellbeing Study showed that Head Start attendance was not significantly associated with cognitive gains when compared to attending pre-K or other center-based care (Zhai, Brooks-Gunn, & Waldfogel, 2011). Similarly, an analysis of Head Start Impact Study data found more substantial differences between children attending Head Start and children in parental or relative/nonrelative care than between children attending Head Start and children attending other center-based care (Zhai, Brooks-Gunn, & Waldfogel, 2014).

Understanding the reduction in effect sizes in elementary school

Evaluations show that ECE programs have positive short-term effects. But multiple studies show that these effects fadeout (or decrease) by the third grade, with a decline of up to .03 per year in effect sizes for cognitive and test score outcomes (Camilli, Vargas, Ryan, & Barnett, 2010). Fadeout is also called the "convergence" or the "catch-up effect," as the gap in achievement between children who attended (and benefited from) ECE programs and children who did not attend such programs decreases as the children get older (Yoshikawa et al., (2013). Eventually, children without any ECE perform as well as children who received ECE. However, receiving ECE is positively related to other long-term outcomes, such as higher earnings and a lower likelihood of criminal activity (Duncan & Magnuson, 2013; Karoly, 2016).

Across almost all experiments, effect sizes from ECE evaluations fall by one-half, on average, between the end of the program and the middle of elementary school. At the moment, this evidence is based almost exclusively on achievement test scores, although a few evaluations have reported a similar decline for aggressive behaviors and approaches to learning. Consequently, a reasonable expectation is that unless changes are made to K–3 education, sustained effects will be one-half the size of short-term effects.

The possible reasons for this decline include:

1. Children who did not receive ECE use kindergarten and first grade to catch up to their peers, mastering comparable skills later than children who received ECE (Duncan & Magnuson, 2013).
2. Early elementary school teachers may emphasize skills that children do not have (i.e., they direct teaching toward students with lower skills, including those who may not have had any preschool education) (Duncan & Magnuson, 2013).

3. Differences in curricular content and quality of instruction may also contribute to the fadeout of ECE's positive effects. Another potential cause is a lack of integration between preschool and elementary school curricula. Continuity between ECE and elementary school curricula is important for sustaining effects over time; some interventions, including one in Maryland's Montgomery Country, have implemented such continuity (Brooks-Gunn, Markman-Pithers, & Rouse, 2016). When curricula are integrated, skills developed in ECE can be practiced and reinforced in elementary school.
4. In terms of instructional style, elementary schools may emphasize individualized learning less than preschools do. Preschool classes have lower adult-child ratios than elementary schools; preschool classes are often limited to 20 students, while elementary school classes often have 26 to 30 students (Pianta, Downer, & Hamre, 2016). In one study in Tennessee, smaller classes in elementary school were associated with better cognitive outcomes (Mosteller, 1995; Heckman, 2006). Individualized instruction has been shown to be most effective for learning outcomes (Clements & Sarama, 2016; Hebbeler & Spiker, 2016), and increased class sizes hinder teachers' ability to provide high-quality interactions with children. Moreover, low-income students are likely to attend elementary schools with larger class sizes, which are associated with lower achievement in general and may dilute preschool gains.
5. Instructional quality may also vary more in elementary school than in preschool, or quality may matter more for learning in elementary school (Pianta, Downer, & Hamre, 2016). For example, students from low-income backgrounds and students from racial/ethnic minority groups—for whom ECE was developed and who tend to benefit most from ECE—often receive low-quality instruction in elementary school (Burchinal, Howes, Pianta, Bryant, Early, Clifford, & Barbarin, 2008; Mashburn, Pianta, Hamre, Downer, Barbarin, Bryant, & Howes, 2008; Moiduddin, Aikens, Tarullo, West, & Xue, 2012). Students attending low-quality elementary schools cannot build on or sustain gains made in preschool, and the positive effects of preschool become less apparent. Low-income students who attended preschool may also be more likely to attend schools in communities where after-school programs, an extended school year, and other enrichment activities are not offered, making it difficult to sustain effects.
6. Elementary schools also tend to provide less support to parents than preschool programs do—especially ECE programs that primarily serve low-income families. For example, Early Head Start offers home visiting, referrals for health care, and parent education. Similarly, the Tulsa program offers parent education, health and vision screenings, and child care services. Such comprehensive supports have been shown to improve cognitive, academic, and health outcomes for children, but elementary schools don't often offer them (Phillips, Gormley, & Anderson, 2016). More comprehensive services for parents and families during elementary school might help sustain ECE's positive effects (Reynolds, Magnuson, & Ou, 2010), although little is known about the efficacy of such efforts (Magnuson & Schindler, 2016).

PROMOTING SUCCESS: A MULTILEVEL MODEL

Multiple changes in pre-K to 3: A model for ensuring success?

Almost all ECE evaluations have assessed individual children, typically those who received an intervention and those who did not via random assignment, waitlist, or eligible age for entrance into preschool. But some have used administrative data as well. One interesting approach is to analyze school- or district-wide data from standardized testing to look at differences in achievement levels. In this way, comparisons can be made across time to see whether an intervention implemented at the school or district level has increased mean scores or competency levels. Such a design is a variant of regression discontinuity.

Such a cohort comparison was used effectively in the county-level effort in Montgomery County, MD (Marietta, 2010). The school district staff, after examining the proportion of the district's high school seniors who were ready for college, set a goal of having 80% of a graduating class college-ready. Working backward, they defined their goals for classes of pre-K to third-grade children. They aimed to increase the percentage of children reading at grade level in the early grades. They then made a list of possible reforms that based on research were likely to prepare their young students to eventually be ready for college.

> *The reforms were extensive, underlying the fact that no single change is likely to have large effects. The county applied most of the changes recommended by early childhood educators. These included:*
>
> 1. pre-K for all four-year-olds,
> 2. full-day pre-K,
> 3. full-day kindergarten,
> 4. after-school programs,
> 5. summer programs,
> 6. curricula aligned across the early grades,
> 7. student-teacher ratios of only 15 to 1 from pre-K to third grade,
> 8. pre-K teachers having a BA and being certified in ECE,
> 9. earnings of pre-K teachers at parity with teachers in kindergarten to third grade,
> 10. English as a second language courses for parents, and
> 11. welcome packets and curricular guidebooks for parents of entering kindergartners.

This intensive and extensive set of reforms doubled the percentage of children reading at grade level by third grade, and this proportion was sustained through the later elementary school years (Marietta, 2010).

Most system-wide initiatives have not taken Montgomery County's approach to evaluation. And such initiatives have not coupled school-level data with individual-child data. Putting the two together might, for example, let us discover which subgroups of children are most likely to see an increase in the share of students reading at grade level, or which set of services are most likely to produce a higher proportion of competent readers.

CONCLUSION

Programs that report sustained effects in elementary school and beyond typically have large effects at the end of an ECE program. Is it critical to have effect sizes of about one-sixth to one-third of a standard deviation at the end of a program to have any chance of seeing sustained effects? The evidence to date suggests that the answer is yes, absent changes in elementary school.

Therefore, we should try to amplify effect sizes in ECE programs, in the hope of improving both short- and long-term outcomes for children. Multiple steps could be taken to increase effect sizes in preschool. First, it is important to increase the dosage and duration of preschool. To increase dosage and duration, it is recommended that students attend preschool more days per year, and even that children attend two consecutive years of preschool. Additionally, there is some evidence that full-day programs are more effective than partial-day programs. Second, it is important to develop and implement more targeted and integrated curricula in preschool. Curricula should be developmentally appropriate and should aim to help children develop essential cognitive and social-emotional skills, as well as ensure that children have the necessary academic skills for elementary school. Moreover, preschool curricula and elementary school curricula should be integrated in an attempt to ensure continuity between the two programs. Third, to ensure the effective implementation of the targeted and integrated curricula, teachers need to be better trained. Fourth, programs should focus on teacher support and scaffolding of skills. Adequate support allows teachers to use structured, individualized teaching models that help children progressively build skills. Last, programs should target poor, minority, and immigrant children to narrow some of the early gaps in math and language literacies.

Although early childhood education programs like Tulsa's Head Start and Boston's Pre-K initiatives provide encouraging support for further investment in early childhood education, we should be specific in determining where and when to invest. Numerous studies have illustrated that it's important to increase young children's exposure to ECE while also working to ensure that quality is consistent across sites and types of programs. Further, the connection (in terms of curriculum, outcomes, and quality) between ECE and K–12 education should be strengthened to promote the maintenance of ECE gains. Policymakers should aim to use the lessons from previous evaluations to improve ECE programs in hopes of reducing achievement gaps and preparing young children for elementary school and beyond.

References

Barrow, L., & Markman-Pithers, L. (2016). Supporting young English learners in the United States. *Future of Children*, 26(2), 159-183.

Barnett, W. S. (1985). Benefit-cost analysis of the Perry Preschool Program and its policy implications. *Educational Evaluation and Policy Analysis, 7*(4), 333-342.

Barnett, W. S. (2008). *Preschool education and its lasting effects: Research and policy implications*. Boulder and Tempe, CO: Education and the Public Interest Center & Education Policy Research Unit.

Bassok, D. (2010). Do black and Hispanic children benefit more from preschool? Understanding differences in preschool effects across racial groups. *Child Development, 81*(6), 1828-1845.

Belfield, C., Nores, M., Barnett, W.S, & Schweinhart, L. (2006). The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Follow-Up. *Journal of Human Resources 41*(1), 162-190.

Blau, D. (2003). Childcare subsidy programs. In. R. A. Moffitt (Ed.), *Means-tested transfer programs in the United States* (pp. 443-516). Chicago: University of Chicago Press.

Brooks-Gunn, J. (2003). Do you believe in magic? What we can expect from early childhood intervention programs. *Social Policy Report, Society for Research in Child Development, 17*, 3-15.

Brooks-Gunn, J., & Hearn, R. (1982). Early intervention and developmental dysfunction: Implications for pediatrics. *Advances in Pediatrics, 29*, 497-527.

Brooks-Gunn, J., Magnuson, K., & Waldfogel, J. (2009). Long-run economic effects of early childhood programs on adult earnings. *Partnership for America's Economic Success Issue Report #12*, 1-34. Retrieved from http://www.readynation.org/docs/researchproject_adultearnings_200903_brief.pdf

Brooks-Gunn, J., & Markman, L. (2005). The contribution of parenting to ethnic and racial gaps in school readiness. *Future of Children, 15*, 139-168.

Brooks-Gunn, J., Markman-Pithers, L., & Rouse, C. E. (Eds.). (2016). Starting early: Education from prekindergarten to third grade (special issue). *Future of Children, 26*(2).

Brooks-Gunn, J., Smith, J., Klebanov, P. K., Duncan, G. J., & Lee, K. (2003). The black-white test score gap in young children: Contributions of test and family characteristics. *Applied Developmental Science, 7*, 239-252.

Bureau of Labor Statistics. (2016a). *Facts over time*. U.S. Department of Labor. Retrieved from https://www.dol.gov/wb/stats/facts_over_time.htm

Bureau of Labor Statistics. (2016b). *Employment characteristics of familes-2015*. U.S. Department of Labor. Retrieved from https://www.bls.gov/news.release/pdf/famee.pdf

Burchinal, M. R., Howes, C. Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Applied Developmental Science, 12*, 140-153.

Burchinal, M., Nelson, L., Carlson, M., & Brooks-Gunn, J. (2008). Neighborhood characteristics, and child care type and quality. *Early Education and Development, 19*, 702-725.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W.S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*(3), 579-620.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science, 6*(1), 42-57.

Cannon, J. S., Kilburn, M. R., Karoly, L. A., Mattox, T., Muchow, A. N., & Buenaventura, M. (2017). *Investing early: Taking stock of outcomes and economic returns from early childhood program*. Santa Monica, CA: RAND.

Card, D., & Krueger, A. (1996). School resources and student outcomes: An overview of the literature and new evidence from North and South Carolina. *Journal of Economic Perspectives 10*(4), 31-50.

Chaudry, A., Morrissey, T., Weiland, C., & Yoshikawa, H. (2017). *Cradle to kindergarten: A new plan to combat inequality*. New York: Russell Sage Foundation.

Clements, D. H., & Sarama, J. (2016). Math, science, and technology in the early grades. *Future of Children, 26*(2), 75-94.

Clements, D. H., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly, 36*(3), 79-90.

Currie, J. (2005). Health disparities and gaps in school readiness. *Future of Children, 15*(1), 117-138.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *The American Economic Review, 85*(3), 341-364.

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *Future of Children*, 15(1), 35-54.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*(2), 109-132.

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

Duggan, M., Kearney, M. S., & Rennane S. (2015). The Supplemental Security Income program. In R. A. Moffitt (Ed.), *Economics of Means-Tested Transfer Programs in the United States, Volume II* (pp. 235-297). Chicago: University of Chicago Press.

Elango, S., Garcia, J. L., & Heckman, J. J., & Hojman, A. (2015). Early childhood education. Working paper, National Bureau of Economic Research, Cambridge, MA. https://www.nber.org/papers/w21766

Fisher, P. A., Frenkel, T. I., Noll, L. K., Berry, M., & Yockelson, M. (2016). Promoting healthy child development via a two-generation translational neuroscience framework: The Filming Interactions to Nurture Development video coaching program. *Child Development Perspectives, 10*(4), 251-256.

Friedman-Krauss, A. H., Barnett, W. S., Weisenfeld, G. G., Kasmin, R., DiCrecchio, N., & Horowitz, M. (2017). *The state of preschool 2017: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

Fryer, R. G., & Levitt, S. D. (2013). Testing for racial differences in the mental ability of young children. *American Economic Review, 103*(2), 981-1005.

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review,* 92(4), 999-1012.

Georges, A., Malone, L., & Brooks-Gunn, J. (2012). Link between young children's behavior and achievement: The role of social class and classroom composition. *American Behavioral Scientist, 56*, 961-990.

Gormley, W. T., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's pre-K program. *Journal of Human Resources, 40*(3), 533-558.

Han, W-J., Hetzner, N. P., & Brooks-Gunn, J. (2019). Employment and parenting. In M. Bornstein (Ed.), *The Handbook of Parenting* (3rd ed.). New York: Routledge.

Hebbeler, K., & Spiker, D. (2016). Supporting young children with disabilities. *Future of Children, 26*(2), 185-205.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312*(5782), 1900-1902.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the High/Scope Perry Preschool Program. *Journal of Public Economics, 94*(1-2), 114-128.

Hill, C. J., Gormley, W. T., & Adelstein, S. (2015). Do the short-term effects of a high-quality preschool program persist? *Early Childhood Research Quarterly, 32*, 60-79.

Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *Journal of Policy Analysis and Management, 21,* 601-627.

Hill, J., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39*, 730-744.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review, 26*, 55-88.

Hoff, E. (2012). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49*(1), 4-14.

Johnson, A., Martin, A., & Brooks-Gunn, J. (2013). Child-care subsidies and school readiness in kindergarten. *Child Development, 84*, 1806-1822.

Jones, J. (2014, September 8). In U.S., 70% favor federal funds to expand Pre-K education. *Gallup*, Retrieved from https://news.gallup.com/poll/175646/favor-federal-funds-expand-pre-education.aspx

Kalil, A., Ziol-Guest, K. M., Ryan, R. M., & Markowitz, A. J. (2016). Changes in income-based gaps in parent activities with young children from 1988 to 2012. *AERA Open, 2*(3), 1-17.

Karoly, L. A. (2016). The economic returns to early childhood education. *Future of Children, 26*(2), 37-55.

Kilburn, M. R., & Karoly, L. A. (2008). *The economics of early childhood policy: What the dismal science has to say about investing in children*. Santa Monica, CA: RAND.

Klebanov, P. K., Brooks-Gunn, J., McCarton, C., & McCormick, M. C. (1998). The contribution of neighborhood and family income to developmental test scores over the first three years of life. *Child Development*, 69, 1420-1436.

Lee, V., Brooks-Gunn, J., & Schnur, E. (1988). Does Head Start work? A 1-year follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Developmental Psychology, 24*, 210-222.

Lee, R., Brooks-Gunn, J., Han, W-J., Waldfogel, J., & Zhai, F. (2014). Is participation in Head Start associated with less maternal spanking for boys and girls? *Children and Youth Services Review, 46*, 55-63.

Lee, V., Brooks-Gunn, J., Schnur, E., & Liaw, T. (1990). Are Head Start effects sustained? A longitudinal follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Child Development, 61*, 495-507.

Lee, R., Han, W-J., Waldfogel, J., & Brooks-Gunn, J. (2018). Preschool attendance and school readiness for children of immigrant mothers in the United States. *Journal of Early Childhood Research, 16*(2), 190-209.

Lee, R., Zhai, F., Han, W-J., Waldfogel, J., & Brooks-Gunn, J. (2014). Head Start participation and school readiness: Evidence from the Early Childhood Longitudinal Study-Birth Cohort. *Developmental Psychology, 50*, 202-215.

Lee, R., Zhai, F., Han, W., Brooks-Gunn, J., & Waldfogel, J. (2013). Head Start and children's nutrition, weight, and health care receipt. *Early Childhood Research Quarterly, 28*, 723-733.

Love, J. M., Chazan-Cohen, R., Raikes, H., Brooks-Gunn, J. (2013) What makes a difference? Early Head Start Evaluation findings in a developmental context. *Monographs of the Society for Research in Child Development, 78*.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics, 122*(1), 159-208.

Magnuson, K., & Schindler, H. (2016). Parent programs in pre-K through third grade. *Future of Children, 26*(2), 207-221.

Marietta, G. (2010). *Lessons for PreK-3rd from Montgomery County Public Schools*. New York, NY: Foundation for Child Development.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., & ... Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732-749.

Matthews, H., Schulman, K., Vogtman, J., Johnson-Staub, C., & Blank, H. (2015). *Implementing the Child Care and Development Block Grant reauthorization: A guide for states*. Washington, D.C.: The Center for Law and Social Policy and The National Women's Law Center.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *Future of Children, 5*(2), 113-127.

Moiduddin, E., Aikens, N., Tarullo, L., West, J., & Xue, Y. (2012). *Child outcomes and classroom quality in FACES 2009*. OPRE Report 2012-37a. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Noble, K. G., Houston, S. M., Brito, N. H., Bartsch, J., Kan, E., Kuperman, J. M., ... & Sowell, E. R. (2015). Family income, parental education, and brain structure in children and adolescents. *Nature Neuroscience, 18*, 773-778.

Phillips, D., Gormley, W., & Anderson, S. (2016). The effects of Tulsa's CAP Head Start Program on middle-school academic outcomes and progress. *Developmental Psychology, 52*(8), 1247-1261.

Pianta, R. C., Barnett, W. S., Burchinal, M., & Thornburg, K. R. (2009). The effects of preschool education: What we know, how public policy is or is not aligned with the evidence base and what we need to know. *Psychological Science, 10*(2), 49-88.

Pianta, R., Downer, J., & Hamre, B. (2016). Quality in early education classrooms: Definitions, gaps, and systems. *Future of Children, 26*(2), 119-137.

Rathbun, A., Zhang, A., & Snyder, T. D. (2016). *Primary early care and education arrangements and achievement at kindergarten entry*. (NCES 2016-070). Washington, D.C.: National Center for Education Statistics.

Raver, C. C., & Blair, C. (2016). Neuroscientific insights: Attention, working memory, and inhibitory control. *Future of Children, 26*(2), 95-118.

Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Smallwood, K., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly, 63*(3), 253-255.

Reardon, S. (2011). The widening academic achievement gap between the rich and poor: New evidence and possible explanations. In R. Murnane & G. Duncan, (Eds.), *Whither opportunity? Rising inequality and the uncertain life chances of low-income children* (pp. 91-116). New York: Russell Sage Foundation Press.

Reardon, S., & Portilla, X. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open,* 2(3), 1-18.

Reid, J. L., Melvin, S. A., Kagan, S. L., & Brooks-Gunn, J. (2019). Building a unified system for universal Pre-K: The case of New York City. *Children and Youth Services Review, 100*, 191-205.

Reynolds, A. J., Magnuson, K. A., & Ou, S. (2010). Preschool-to-third grade programs and practices: A review of research. *Children and Youth Services Review, 32*(8), 1121-1131.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). 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. *Journal of the American Medical Association, 285*(18), 2339-2346.

Rock, D. A., & Stenner, A. J. (2005). Assessment issues in the testing of children at school entry. *Future of Children, 15*(1), 15-34.

Shonkoff, J. P., & Fisher, P. A. (2013). Rethinking evidence-based practice and two-generation programs to create the future of early childhood policy. *Developmental Psychopathology, 25*(4), 1635-1653.

Snow, C., & Matthews, T. (2016). Reading and language in the early grades. *Future of Children, 26*(2), *57-74*. ?

Steuerle, E., & Jackson, L. M. (Eds.). (2016). *Advancing the power of economic evidence to inform investments in children, youth, and families*. Washington, D.C.: The National Academies Press.

Stipek, D., Franke, M., Clements, D., Farran, D., & Coburn, C. (2017). PK-3: What does it mean for instruction? *Social Policy Report, 30*(2), 1-22.

U.S. Department of Health and Human Services. (2010). *Head Start impact study: Final report, January 2010*. Washington, DC: Administration for Children and Families, Office of Planning, Research and Evaluation.

Votruba-Drzal, E. (2003). Income changes and cognitive stimulation in young children's home learning environments. *Journal of Marriage and Family, 65*(2), 341-355.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*(6), 2112-2130.

Yazejian, N., Bryant, D. M., Hans, S., Horm, D., St. Clair, L., File, N., & Burchinal, M. (2017). Child and parenting outcomes after 1 year of Educare. *Child Development, 88*(5), 23-39.

Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When does preschool matter? *Future of Children, 26*(2), 21-36.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M., Espinosa, L., Gormley, W., Ludwig, J.O., Magnuson, K.A., Phillips, D.A., & Zaslow, M.J. (2013). *Investing in our future: The evidence base for early childhood education*. New York: Foundation for Child Development and Ann Arbor, MI: Society for Research in Child Development.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2011). Head Start and urban children's school readiness: A birth cohort study in 18 cities. *Developmental Psychology, 47*, 134-152.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology, 50*(12), 2572-2586.

Zhai, F., Waldfogel, J., & Brooks-Gunn, J. (2013). Head Start, prekindergarten, and academic school readiness: A comparison among regions in the United States. *Journal of Social Service Research, 39*, 345-364.