

# Long-Term Effects of Head Start on Low-Income Children

JENS LUDWIG<sup>a</sup> AND DEBORAH A. PHILLIPS<sup>b</sup>

<sup>a</sup>*Social Service Administration, Law, and Public Policy,  
University of Chicago, Chicago, Illinois 60637  
The Brookings Institution, Chicago, Illinois 60637*

<sup>b</sup>*Georgetown University, Washington, DC 20057  
Georgetown Public Policy Institute, Washington, DC 20057*

**A growing body of research suggests that the first few years of life are a particularly promising time to intervene in the lives of low-income children, although the long-term effects on children of the U.S. government's primary early childhood program—Head Start—remains the topic of debate. In this article we review what is known about Head Start and argue that the program is likely to generate benefits to participants and society as a whole that are large enough to justify the program's costs. Although in principle there could be more beneficial ways of deploying Head Start resources, the benefits of such changes remain uncertain and there is some downside risk.**

*Key words:* early childhood education; early childhood interventions; head start; poverty

## Introduction

A growing body of research in neuroscience, developmental psychology, economics, and even animal studies suggests that the earliest years of life are a particularly promising time to intervene in the lives of low-income children.<sup>1,2</sup> Intensive model early childhood programs from the 1960s and 1970s, such as Perry Preschool and Abecedarian, have been shown to affect important adult economic and other outcomes.<sup>3,4</sup> These programs also appear to generate benefits far in excess of program costs.<sup>5–7</sup> These findings would at first glance seem to argue for substantial increases in society's investment in early childhood education, although whether small model programs can be effectively taken to scale remains unclear.

Head Start—the U.S. government's major early childhood program—is the main example of such a scaled-up program and has consistently generated debate about whether it produces lasting benefits to program participants. Head Start was launched in 1965 as part of President Lyndon B. Johnson's War on Poverty and provides low-income children aged 3–5 years, and their parents, with schooling, health, nutrition, and social welfare services. The first published study ar-

guing that Head Start probably benefits to children fade out rapidly was published in 1966.<sup>8,9</sup> Although Head Start has grown over time, and currently serves nearly 1 million low-income children each year at a cost of about \$7 billion, skepticism about the value of the program persists.<sup>10,11</sup>

This article reviews what is known about the value of Head Start. The best available evidence suggests that Head Start probably passes a benefit–cost test. Although there remain some important limitations to the available evidence on Head Start, we believe the weight of the evidence points in this direction. In principle there might be ways to increase the cost-effectiveness of current Head Start funding, including changes to Head Start's design or funding alternatives, such as state prekindergarten (pre-K) programs. However, the benefits of such changes remain uncertain and they entail some downside risk.

Our report seeks to develop five main arguments that lead us to these conclusions. First, much of the debate about Head Start stems from confusion about how to judge the magnitude of program effects. We argue that the most appropriate standard for judging the program's success is benefit–cost analysis.

Second, over the past several years new evidence has been accumulating about the long-term effects of Head Start on early cohorts of program participants, as well as about the short-term program effects on more recent cohorts of children. There is now suggestive evidence that Head Start probably generated lasting benefits to

---

Address for correspondence: Jens Ludwig, University of Chicago, 969 E. 60th St., Chicago, IL 60637. Voice: 773-702-3242. [jludwig@uchicago.edu](mailto:jludwig@uchicago.edu)

program participants during the first few decades of operation and passed a benefit–cost test.<sup>12–14</sup>

These findings counter the view that only intensive (and expensive) early childhood interventions can generate long-term benefits, and they run counter to the perception that Head Start has been a failure from its inception. However, these results are not directly informative about whether today’s version of Head Start passes a benefit–cost test, since Head Start and the counterfactual developmental environments poor children would otherwise experience are both changing over time. This is a generic challenge to understanding the long-term effects of contemporaneous government programs—we can estimate long-term effects only for people who participated in the program a long time ago.

The best evidence currently available on Head Start as it operates today comes from a recent randomized experimental evaluation of Head Start’s effects measured within 1 year of random assignment, which was sponsored by the federal government and carried out by Westat.<sup>15</sup> Public discussions of the experimental results have typically focused on the effects of being assigned to the experiment’s treatment group rather than the control group, known in the program evaluation literature as the intent-to-treat (ITT) effect. These effects are presented by Westat separately for 3- and 4-year-old program participants and are usually in the direction consistent with some beneficial effect of Head Start on children’s short-term outcomes, but the findings are often not statistically significant.

Our article’s third objective is to provide some benchmarks for how large these short-term effects would need to be for us to believe that any long-term benefits generated by today’s Head Start program will be large enough to justify the program’s costs. This exercise is complicated by the fact that there is currently limited evidence about how the cognitive and noncognitive skills of young children translate into long-term life outcomes and so requires imposing several assumptions and projections out of sample. With these important caveats in mind, the evidence that is available suggests that given Head Start’s costs (around \$9,000 per child on average), the program might pass a benefit–cost test if the short-term effects on achievement test scores were as small as 0.1–0.2 standard deviations.

Our fourth goal is to demonstrate that the effects implied by the recent Head Start experiment of actually enrolling in Head Start (the “effects of treatment on the treated,” or TOT) are typically large enough to imply that the program would pass a benefit–cost test. The official report on the Head Start experiment’s first-year findings shows results separately for 3- and

4-year-olds, and many effects are not statistically significant. However, if program effects were estimated pooling 3- and 4-year-olds, most of the program effects on the main cognitive outcomes of interest would be statistically significant.

Finally, we close with some discussion about recent suggestions that have been made for increasing the cost-effectiveness of Head Start funding, including changing the program design, making the program more like some of the newer state pre-K programs in operation around the country, or even diverting some Head Start funding to these state programs. There is in our view some uncertainty about both the short-term and long-term benefits associated with these changes. There are also downside risks, particularly if one recognizes that there is some opportunity cost associated with the resources required to implement some of the proposed changes to Head Start.

### Benefits of Benefit–Cost Analysis

The argument that we should judge the magnitude of Head Start’s effects by how the dollar value of these benefits compares to the cost of the program will not seem like a new idea to economists and policy analysts. Yet much of the public debate about the value of Head Start reflects some basic confusion on this point.

One benchmark that has been used to gauge the size of Head Start’s effects is relative to the scale of the social problem that is being addressed. For example, Besharov reviews the Westat report and argues “these small gains will not do much to close the achievement gap between poor children (particularly minority children) and the general population. We should expect more . . . ”<sup>11</sup>

But the right standard of success for a public program is not the elimination of a social problem. Consider that mortality rates from lung cancer in the United States in 2003 remain high—71.9 deaths per 100,000 people for males and 41.2 deaths per 100,000 for females.<sup>16</sup> The fact that thousands of Americans continue to die each year from lung cancer does not mean that the large decline in tobacco smoking observed during the last half of the 20th century should be considered a public health “failure,” particularly because diverting smokers from smoking appears to make them happier as well as healthier.<sup>17,18</sup>

Psychologists and education researchers often use the typology offered by Cohen, who argues that an “effect size” (that is, program effect expressed as a share of a control group standard deviation) of 0.2 should be considered “small,” whereas effect sizes of

0.5 should be considered “medium,” and those of 0.8 or more, “large.”<sup>19</sup> Cohen’s typology is based on the empirical distribution of effect sizes observed in the body of education research available at the time of his study.<sup>20</sup> Cohen’s typology for characterizing effect sizes is the convention adopted by Westat in its report on the short-term results of the recent randomized Head Start experiment. The Westat report calls effect sizes below 0.2 standard deviations “small,” whereas those of 0.2–0.5 standard deviations are “medium” and those over 0.5 are “large” (see p. ii, footnote 1, in Ref. 15).

Yet any assessment of what a program accomplishes should take into account not only the program’s benefits but also its costs, which requires converting both into some common metric—that is, benefit–cost analysis. A program that improved test scores by 0.8 standard deviations—“large” in Cohen’s scheme—but cost a total of \$10 trillion per year would be difficult to support, because undertaking such an early childhood intervention would absorb most of the nation’s gross domestic product with little left to house, clothe, feed, and protect the nation’s child (and adult) population.<sup>19</sup> At the other extreme, a program that generated effects on the order of 0.2 standard deviations but cost only a nickel per child per year would be difficult to oppose. We should expect social programs to generate net benefits, not miraculous benefits. For an excellent discussion of these points, see Duncan and Magnuson (forthcoming), “Penny size and effect size foolish,” *Child Development Perspectives*. Harris presents a cost-effectiveness framework for judging program effects and suggests that any intervention that generates increased test scores of 0.025 standard deviations per child per \$1,000 spending should be considered “large” (Harris, Douglas N. [2007], “New benchmarks for interpreting effect sizes: Combining effects with costs,” Working Paper, University of Wisconsin at Madison). The implication is that Head Start effects of 0.175 standard deviations would be “large” under this framework, roughly consistent with our own benchmarks.

### **Evidence on Head Start’s Long-Term Effects**

Whereas researchers have been studying Head Start for more than 40 years, only recently have social scientists made much headway in identifying the causal effects of the program on participating children. There is now at least suggestive evidence on Head Start’s long-term effects that the program passed a benefit–cost test over the first few decades of operation, but the program is changing over time and so these effects might

not be relevant to today’s Head Start. Short-term effects from the recent randomized experimental study of Head Start by themselves are not directly informative about whether the program’s long-term benefits justify program costs.

Long-term effects of Head Start can be identified only for those children who participated in the program a long time ago. The main challenge in identifying the long-term effects of Head Start on earlier cohorts of children comes from trying to figure out what the outcomes of Head Start participants would have been had they not enrolled in the program. Simply comparing the long-term outcomes of children who did participate with those who did not may provide misleading answers to the key causal question of interest. If, for example, relatively more disadvantaged families participate in Head Start, then simple comparisons of Head Start recipients to other children may understate the program’s effectiveness if researchers cannot adequately measure all aspects of family disadvantage. The opposite bias may result if instead the more motivated and effective parents are the ones who can get their children into (or are selected by program administrators for) scarce Head Start slots.

Economists Garces, Thomas, and Currie evaluate Head Start by comparing the experiences of siblings who did and did not participate in the program.<sup>13</sup> The analytic sample consists of children who would have participated in Head Start in 1980 or earlier. These sorts of within-family, across-sibling comparisons help to eliminate the confounding influence of unmeasured family attributes that are common to all children within the home (but not, of course, unshared family inputs).

The research design that Garces and colleagues used represents a substantial improvement over previous research, although there necessarily remains some uncertainty about why some children within a family but not others participate in Head Start and whether whatever is responsible for this within-family variation in program enrollment might also be relevant for children’s outcomes. For example, sibling comparisons might overstate (or understate) Head Start’s effects if parents enroll their more (or less) able children to participate in the program.

The Garces study might also understate Head Start’s effects if there are positive spillover effects of participating in the program on other members of the family, because here the control group for the analysis (i.e., siblings who do not enroll in Head Start themselves) will be partially treated (i.e., benefit to some degree from having a sibling participate in Head Start). Also, their study relies on retrospective self-reports of Head Start participation by people who have reached

adulthood, which some people may misremember or misreport. If this measurement error is uncorrelated with respondents' characteristics and potential outcomes, then misreporting will attenuate their estimated Head Start effects to some degree (i.e., biased toward zero).

With these caveats in mind, Garces, Thomas and Currie report that non-Hispanic, white children who were in Head Start are about 22 percentage points more likely to complete high school than their siblings who were in some other form of preschool, and about 19 percentage points more likely to attend some college. These effect estimates are equal to around one-quarter and one-half of the "control mean." For African Americans, the estimated Head Start effect on schooling attainment is small and not statistically significant, but for this group Head Start relative to other preschool experience is estimated to reduce the chances of being arrested and charged with a crime by around 12 percentage points, which, as with the schooling effect for whites, is a large effect.

Ludwig and Miller use a different research design to overcome the selection bias problems in evaluating the long-term effects of Head Start and generate qualitatively similar findings for schooling attainment, although unlike Garces *et al.* they find evidence for effects for blacks as well as whites.<sup>14</sup> Their design exploits a discontinuity in Head Start funding across counties generated by how the program was launched in 1965. Specifically, the Office of Economic Opportunity provided technical grant-writing assistance for Head Start funding to the 300 counties with the highest 1960 poverty rates in the country, but not to other counties. The result is that Head Start participation and funding rates are 50%–100% higher in the counties with poverty rates that just barely put them into the group of the 300 poorest counties than in those counties with poverty rates just below this threshold. So long as other determinants of children's outcomes vary smoothly by the 1960 poverty rate across these counties, any discontinuities (or "jumps") in outcomes for those children who grew up in counties just above versus below the county poverty rate cutoff for grant-writing assistance can be attributed to the effects of the extra Head Start funding.

Using this regression discontinuity design, Ludwig and Miller find that a 50%–100% increase in Head Start funding is associated with a decline in mortality from causes of death that could be affected by the program of 33%–50% of the control mean, as well as suggestive evidence for an increase in schooling attainment of about one-half year, and an increase in the likelihood of attending some college of about 15%.

Importantly, the estimated effects of extra Head Start funding on educational attainment are found for both blacks and whites. These estimates are calculated for children who would have participated in Head Start during the 1960s or 1970s and cannot be calculated for more recent cohorts of program participants because the Head Start funding discontinuity across counties at the heart of this research design appears to have dissipated over time.

Taken together, these effect estimates suggest that Head Start—as it operated in the 1960s through 1980s—generated benefits in excess of program costs, with a benefit–cost ratio that could be as large as the 7-to-1 figure often cited for model early childhood programs, such as Perry Preschool. Currie notes that the short-term benefits of Head Start to parents in the form of high-quality child care together with medium-term benefits from reductions in special education placements and grade retention might together offset between 40% and 60% of the program's costs.<sup>21</sup> Ludwig and Miller's estimates imply that each extra dollar of Head Start funding in a county generates benefits from reductions in child mortality and increases in schooling attainment that appear to outweigh the extra program spending.<sup>14</sup> Also, Frisvold provides some evidence that Head Start reduces childhood obesity.<sup>22</sup>

These findings run counter to the common view that only intensive and expensive early childhood interventions can generate long-term benefits. The origin of this conventional wisdom is itself not entirely clear, because there is no logical reason that lower-cost programs will necessarily have lower benefit–cost ratios than those from higher-cost programs. The effects of Head Start on children will depend on the difference in the developmental quality of the program versus the quality of the environments that low-income children would have experienced otherwise. During its early years, Head Start did not score well on commonly used indicators of early childhood program quality, such as teacher educational attainment. But for poor children in the 1960s through 1980s, the evaluation studies described above imply that the environments Head Start children would have experienced if not enrolled in the program were less developmentally productive than Head Start.

One implication of this last point is that the effects of Head Start on poor children may be changing over time in ways that are difficult to predict, and so the long-term effects of Head Start on previous cohorts of children may not represent the long-term effects of the program on today's participants. Over time the Head Start program has improved in quality, but arguably so has the alternative to Head Start for poor children

because parent educational attainments and real incomes have increased since the 1960s and state-funded preschool programs have been introduced. Which environment is improving faster in this horse race is unclear.

Fortunately, the federal government has recently sponsored the first-ever randomized experimental evaluation of Head Start, the Head Start National Impact Study (HSNIS; hereafter “the randomized Head Start experiment”), with first-year results that are now available from Westat, the evaluation subcontractor.<sup>15</sup> Starting in 2002, nearly 4,700 3- and 4-year-old children whose parents applied for Head Start were randomly assigned to a Head Start treatment group or a control group that was not offered Head Start through the experiment, but they could participate in other local preschool programs if slots were available. The 84 Head Start centers participating in the experiment were selected to be representative of all programs in operation across the country that had waiting lists.

The experiment appears to have been done well—randomization was implemented properly, and careful assessments were made of a wide variety of children’s cognitive and noncognitive outcomes; parents were also studied. However, there were some differences in response rates across groups. Puma *et al.* report that for the first data collection wave in fall 2002, child response rates were 85% for the treatment group and 72% for the control group, and for parents were 89% and 81% for the treatment and control groups, respectively. For the spring 2003 follow-up, response rates for children were 88% and 77% for the treatment and control groups, respectively, and 86% and 79%, respectively, for parents (see pp. 1–18 of Ref. 15). By randomly assigning income-eligible children to the treatment and control conditions, the Head Start experiment uncovers the effects of making Head Start available to all eligible children. If, in practice, Head Start centers focus on enrolling the most disadvantaged of the eligible children that apply, and if the effects of Head Start are more pronounced for more disadvantaged children, then the experimental effect estimates may understate the effect of Head Start on the average program participant in the nation at large.

Although the Head Start experiment is informative about several key policy questions about the program, we cannot directly measure the long-term effects of Head Start for the children who only recently participated in the program. Therefore, we must rely on indirect evidence about what the short-term effect estimates from the recent Head Start experiment might imply about the program’s long-term effects.

## Short-Term Benchmarks for Long-Term Success

Head Start, as the program currently operates, costs about \$9,000 per participating child (paid for by a combination of federal, state, and local funds). How large would Head Start’s short-term effects need to be, and in what outcome domains, for us to believe that the program’s long-term benefits justify the program expenditures? We try to answer this question in two ways: (1) examining the short-term effects that have been found for studies of other early childhood interventions where there is also evidence for long-term benefits in excess of program costs and (2) trying to assess directly the dollar value of a standard deviation increase in early childhood test scores. Both approaches require several assumptions and forecasting out of sample. But we nonetheless believe that there is a reasonable case to be made that positive effects on achievement test scores on the order of 0.1–0.2 standard deviations might be large enough to generate long-term dollar value benefits that would outweigh the program’s costs.

The findings from Garces *et al.* and from Ludwig and Miller suggest that Head Start, as the program operated in the 1960s through 1980s, appears to have generated long-term benefits that exceeded the program’s costs.<sup>13,14</sup> How large were the short-term effects of Head Start on participating children, and in what outcome domains? If the short-term effects of today’s Head Start were about as large as the short-term effects of yesterday’s program, and if the latter passes a benefit–cost test, there would be some reason to believe that the same might be true of the current program.

Using the same sibling-difference design as Garces *et al.*, Currie and Thomas studied children who would have been in Head Start in the 1980s or earlier and found that Head Start participation appears to increase scores on the Peabody Picture Vocabulary Test (PPVT) by around 0.25 standard deviations in the short term for both white and African American children.<sup>13,23</sup> These effects persist for whites but fade out within 3 or 4 years for blacks. Head Start’s effects on Peabody Individual Achievement Test math scores might be around half as large and are not statistically significant (see p. 345, footnote 10, of Ref. 24). Ludwig and Miller find that increased Head Start funding does not lead to statistically significant increases in student achievement test scores in 8th grade in either math or reading, although they cannot rule out effects smaller than around 0.2 standard deviations.<sup>14</sup> Also, they do not have adequate sample sizes to examine effects on test scores separately for blacks and whites.

Unfortunately, not much is currently known about Head Start's causal effects on short-term, *noncognitive* outcomes for earlier cohorts of program participants. Currie and Thomas do find some evidence that Head Start might reduce grade retention for white children who participated in the program in the 1980s or earlier (see Table 14 of Ref. 23). Yet if we interpret short-term test scores as a proxy for the bundle of early skills that promote long-term outcomes, then the previous research on earlier Head Start cohorts indicates that short-term effects of around 0.25 standard deviations for vocabulary and perhaps 0.1 for math might be large enough to generate long-term benefits in excess of program costs.

We can also look at the short-term versus long-term effects of the widely cited Perry Preschool program, which provided poor 3- and 4-year-old children with 2 years of services at a total per-child cost of about twice that of Head Start. Currie cites Perry costs of \$12,884 per child in 1999 dollars.<sup>21</sup> At the end of the second year of services, Perry had increased PPVT vocabulary scores by around 0.91 standard deviations and scores on a test of nonverbal intellectual performance (the Leiter International Performance Test) by around 0.77 standard deviations (see p. 61 of Ref. 3). By age 9, the effect on vocabulary scores had faded out entirely, whereas around half of the original effect on nonverbal performance had dissipated. By age 14, effects on reading and math scores are just over 0.3 standard deviations. Despite this partial fade-out of test score effects, Perry Preschool shows large long-term effects on schooling, crime, and other outcomes measured through age 40.<sup>3</sup> The dollar value of Perry Preschool's long-term benefits (in present dollars) ranges from nearly \$100,000 calculated using a 7% discount rate to nearly \$270,000 with a 3% discount rate (see pp. 180–181 of Ref. 5).

Suppose that short-term test score effects are proportional to the dollar value of long-term program benefits. Here, even if we used a conservative 7% discount rate, Head Start's short-term effects would need to be at most around 9% as large (\$9,000/\$100,000) as those of Perry Preschool (i.e., around 0.1 and 0.06 standard deviations for vocabulary and nonverbal performance, respectively) to generate benefits that are large enough to outweigh Head Start's costs of around \$9,000 per child. If we use a 3% discount rate instead, the necessary short-term effects may be more on the order of 0.04 and 0.03 standard deviations, respectively.

Of course, perhaps long-term gains are not strictly proportional to short-term effects. For example, maybe some minimum short-term effect is necessary to generate lasting cognitive or noncognitive benefits. Or

perhaps the behavioral consequences of achievement effects on the low-IQ sample of Michigan children in Perry Preschool are different from those arising from similar-sized effects on a more representative Head Start population. But, at a minimum, the Perry Preschool data raise the possibility that "small" short-term effects are sufficient for a program with the costs of Head Start to pass a benefit–cost test.

Another way to think about how large Head Start's short-term effects would need to be for the program to pass a benefit–cost test is to measure directly the value of a 1–standard deviation increase in early childhood test scores. Because few studies have monitored people from early childhood all the way through adulthood, this exercise is necessarily subject to some uncertainty. But the available evidence indicates that short-term effect sizes of 0.15–0.2 might be more than enough for Head Start to pass a benefit–cost test, consistent with the evidence from the previous section.<sup>24</sup> So although there is to date no entirely satisfactory way of determining how early test score effects relate to longer life outcomes, the two imperfect approaches used here both indicate that short-term effects that would be considered small by the usual standards of education research could generate long-term benefits that would at least equal Head Start's cost per participant.

### How Large are Head Start's Current Short-Term Effects?

The best available evidence on current Head Start's effects on children comes from the Head Start National Impact Study carried out by Westat for the U.S. Department of Health and Human Services, which we will refer to for convenience as "the randomized Head Start experiment." The results of this experiment have been characterized as "disappointingly small," although other assessments are more positive (see quote at p. 1 of Ref. 11).<sup>25</sup> In any case, much of the public discussion of these findings appears to confuse the ITT effects emphasized in Westat's report on the experimental results with the effects of Head Start participation *per se* (i.e., the TOT). The short-term effects of Head Start participation are usually equal to or greater than the standard deviation benchmark of 0.1 or 0.2 that is necessary for Head Start to pass a benefit–cost test.

One common source of confusion about the recent randomized Head Start experiment stems from the fact that the main results—particularly those in the executive summary to the several-hundred-page report—are not *intended* to reflect the effects of actual Head Start

participation. The executive summary and most of the tables in the body of the report itself focus on the causal effects of offering children the *chance* to participate in Head Start by assigning them to the Head Start experimental group—that is, the ITT effect. These results are often discussed as if they represent the effects of Head Start participation. They do not.

In practice, not everyone who is offered the chance to participate in Head Start will actually enroll—parents, for example, might decide that Head Start will not meet their own or their children's needs, or better alternative opportunities might present themselves. If some people assigned to the experimental treatment group do not participate in the program, and, relatedly, if some people assigned to the control group enroll in Head Start on their own, then the effects of Head Start participation (the TOT) can be different, sometimes markedly so, from the effects of treatment-group assignment.

In the Head Start experimental data, around 86% of 4-year-olds assigned to the experimental treatment group enrolled in Head Start, whereas 18% of 4-year-olds assigned to the control group wound up in Head Start on their own. (See pp. 3–7 of Ref. 15. The figures for 3-year-olds assigned to the treatment and control groups are 89% and 21%, respectively.) The body of the report does mention that the ITT estimates will understate the effects of actually participating in Head Start. But the report's description of how it tries to convert the ITT estimates into something like an estimate for the effect of Head Start participation is confusing, and the actual approach they use might be misleading.<sup>24</sup> In any case, these results are relegated to one of the appendices and perhaps therefore have been largely ignored in public discussions compared to the ITT estimates included in the executive summary.

Why focus on the effects of actually participating in Head Start rather than the ITT estimates? The effect sizes for the Head Start experiment's ITT estimates are often compared to estimates from Perry Preschool and Carolina Abecedarian and the results of more recent evaluations of universal state pre-K programs, all of which estimate the effects of actually participating in these other programs. This sort of uneven comparison (TOT to ITT) comparison will understate the relative effectiveness of Head Start.

A more important reason for focusing on estimates for the effects of actually participating in Head Start (TOT) is to avoid confusion in conducting a benefit–cost analysis of Head Start. In public discussions about Head Start's costs, the focus is always on the costs per actual enrollee. The benefit measure that should be compared with this cost is then the dollar value of the

benefits per enrollee—that is, the dollar value of the gains from actually participating in Head Start.

In TABLE 1 we show the ITT effects on each of the cognitive outcome domains reported in the executive summary of Westat's report for the first-year findings of the Head Start experiment.<sup>15</sup> Although the published Westat report did not show standard errors for impact estimates, Ronna Cook at Westat has generously made these available to us. In TABLE 1 we present point estimates and standard errors that are converted into effect size terms (i.e., expressed as a share of the control group standard deviation for that outcome measure).

TABLE 1 also presents our own estimates for the effects of actually participating in Head Start (the effects of TOT), derived using the approach proposed by Bloom, which involves scaling the difference in mean outcomes between children assigned to the experimental versus control group by the difference across groups in Head Start enrollment rates.<sup>26</sup> (This approach is numerically equivalent to using two-stage least squares with experimental group assignment as an instrument in a model without other covariates.) In the Head Start experiment, the difference in Head Start participation rates between the treatment and control groups is around 68 percentage points and so, using the Bloom procedure, we would estimate that the effects of Head Start enrollment on children are about 1.5 times as large as the ITT effects that are commonly misinterpreted to represent the effects of Head Start participation. These results are best interpreted as providing a range within which the “true” effects of Head Start probably fall. If the average Head Start program quality is somewhat higher for the treatment than control groups, then our Bloom-style estimates for the TOT might be biased somewhat upward.

TABLE 1 shows that, at least for cognitive skills, all of the Head Start effect estimates point in the direction consistent with beneficial program effects, although many of these point estimates are not statistically significant and, in general, the point estimates are larger (both absolutely and in relation to their standard errors) for 3-year-olds than 4-year-olds. For rhetorical convenience we focus on the TOT estimates because we believe that they are likely to be much closer approximations of the true effect of Head Start participation per se than are the ITT estimates. Nevertheless, the true effect is probably somewhere between the ITT and TOT estimates.

For vocabulary, prereading, and prewriting skills, Head Start's effects (TOT) range from 0.15 to 0.35 standard deviations, whereas for 4-year-olds the effects are one-third to one-half as large as for 3-year-olds on the PPVT and smaller for prereading and prewriting.

**TABLE 1. Intent-to-treat (ITT) effect sizes from the National Head Start Impact Study and estimated effects of treatment on the treated (TOT)**

Outcome	3-year-olds		4-year-olds	
	ITT (SD)	TOT	ITT (SD)	TOT
Woodcock–Johnson letter identification	.235* (.074)	.346* (.109)	.215* (.099)	.319* (.147)
Letter naming	.196* (.080)	.288* (.117)	.243* (.085)	.359* (.126)
McCarthy draw-a-design	.134* (.051)	.197* (.075)	.111 (.067)	.164 (.100)
Woodcock–Johnson spelling	.090 (.066)	.132 (.096)	.161* (.065)	.239* (.097)
PPVT vocabulary	.098* (.043)	.144* (.064)	.108 (.071)	.159 (.107)
Color naming	.340* (.066)	.499* (.097)	.293* (.075)	.435* (.112)
Parent-reported literacy skills	.025 (.062)	.036 (.091)	–.058 (.052)	–.086 (.077)
Oral comprehension	.124 (.083)	.182 (.122)	.100 (.070)	.147 (.103)

First and third columns reproduce ITT effect estimates for all cognitive outcomes reported in Westat's Executive Summary of the first-year findings report from the National Head Start Impact Study, reported as effect sizes, that is, program effects divided by the control group standard deviation (Puma *et al.*, 2005). Standard errors, shown in parentheses, are also in effect size terms; these were not included in the Westat report but were generously shared with us by Ronna Cook of Westat. Second and fourth columns are our own estimates for the effects of TOT, derived using the approach of Bloom (1984), which divides the ITT point estimates and standard errors by the treatment–control difference in Head Start enrollment rates. For 3-year-olds, the adjustment is to divide ITT by  $(.894 - .213) = .681$ ; for 4-year-olds, the adjustment is to divide ITT by  $(.856 - .181) = .675$  (see Exhibit 3.3, Puma *et al.*, 2005, p. 3–7). \* = Statistically significant at the 5% cutoff.

Parent-reported literacy skills show much more pronounced Head Start effects, equal to 0.5 and 0.4 standard deviations for 3- and 4-year-olds, respectively. There are reasons to believe that the results from direct student assessments in this outcome may be more reliable than those from parent reports.<sup>27</sup> Head Start's effects on early math scores (measured by the Woodcock–Johnson applied problems test) are equal to 0.18 and 0.15 standard deviations for 3- and 4-year-olds, respectively, and are not statistically significant.

One concern comes from questions about the ability of the available assessments to detect reliable effects of this size in young children. One criterion we have for cognitive or noncognitive assessments is that they are reliable—that is, they generate similar results when applied on different occasions. Reliability scores for achievement tests administered to adolescents are usually on the order of 0.8–0.9.<sup>28</sup> Westat shared with us the reliability scores for the cognitive outcomes used in the Head Start experiment, and these are typically on the same order but sometimes a bit lower. They are also lower for measures of noncognitive skills than for cognitive outcomes.<sup>24</sup> If the limitations of available assessments simply introduce random noise into children's outcome scores, then the dependent variables in

the Head Start experimental analysis will suffer from classical measurement error, causing less precise estimation of Head Start effects (i.e., larger standard errors).

This last point is related to the larger concern that many of the Head Start effects estimated in this experiment are not statistically significant. The Head Start experiment enrolled nearly 4,700 children, which is large by the standards of many social program evaluations but tiny compared with many randomized clinical trials in medicine. The standard errors around the resulting point estimates are therefore subject to some nontrivial sampling uncertainty. This uncertainty is compounded by the fact that Westat's report on the Head Start experiment further splits the sample by showing results separately for 3- and 4-year-olds, particularly in the main table of results shown in the executive summary. Although this splitting of the sample makes sense for developmental reasons (program effects may differ by age), it further reduces statistical power.

Another analytic approach would have been to pool the 3- and 4-year-old samples in the Head Start experiment. Although the Westat report does not present these analyses, with data on the separate

effect estimates, sample sizes, and standard deviations for the 3-year-old and 4-year-old samples we can approximate what the effect estimates and standard errors would be if Westat had pooled the two age groups for analysis. Our calculations will only allow us to calculate standard errors that do not benefit from the improved precision afforded by adjusting for baseline covariates, and so our pooled standard errors are, if anything, conservative. Our calculations suggest that for a pooled sample of 3- and 4-year-olds, the Head Start effect estimate would be statistically significant for every cognitive outcome domain shown in TABLE 1 except oral comprehension.

### Head Start Alternatives

That the current incarnation of Head Start appears to pass a benefit–cost test does not rule out the possibility that there could be even more cost-effective ways of deploying Head Start resources. One possibility that has figured prominently in debates about Head Start is to make the program more academically oriented, rather than to preserve the broad range of academic, health, nutrition, and social services that Head Start has provided to disadvantaged children since the program’s inception. The assumption is that focusing a greater share of children’s time in the program on academic instruction will generate stronger achievement outcomes. Some observers point to larger effect estimates that have been reported from recent studies of new universal state pre-K programs, which are more narrowly focused on instructional activities. They suggest that we should make Head Start operate more like those programs, particularly with respect to the state pre-K requirements that teachers have 4-year college degrees, or even divert funding from Head Start to the state programs. These proposals hold some intuitive appeal. However, the benefits associated with these changes in practice are uncertain; plus, there is some downside risk, and so the expected value of these proposed changes to Head Start remain unclear now.

The recent Head Start experimental evaluation provides rigorous information about the short-term effects of Head Start as it has operated since the program began, as a comprehensive program focused on nutrition, physical and mental health, parenting and social services, as well as education. The studies of Garces, Thomas, and Currie; of Ludwig and Miller; and of Currie and Thomas provide at least suggestive evidence for the long-term effects of Head Start as the program was originally designed.<sup>13,14,23</sup> To date, there

is no evaluation evidence available about what would be achieved for current recipients by a different, new version of Head Start that was more academically oriented.

Several recent studies of universal state pre-K programs suggest impressively large effect estimates. Gormley *et al.* evaluate the effects of Tulsa, Oklahoma’s pre-K program and report TOT estimates equal to 0.8 standard deviations for the Woodcock-Johnson–Revised (WJ-R) letter–word identification test (more than twice as large as those found in the recent Head Start experiment), with effect sizes of 0.65 for the WJ-R spelling test (almost three times as large as those reported for 4-year-olds in the Head Start experiment) and of 0.38 for the WJ-R applied problems math test (more than twice as large as for 4-year-olds in the Head Start experiment), all of which are statistically significant.<sup>29</sup> Barnett *et al.* examine pre-K programs in five separate states and report effect sizes of 0.26 for the PPVT and 0.28 for the WJ-R applied problems test, both of which are statistically significant.<sup>30</sup>

What explains the difference in effect estimates between these state pre-K programs and Head Start? One candidate explanation is that the pre-K programs that have been evaluated to date require all teachers to hold 4-year college degrees, whereas Head Start does not impose that requirement. Teachers in these state pre-K programs will presumably also have higher salaries than those of Head Start teachers, given the difference in average educational attainment. But most of the state pre-K programs report average per-student costs below those of Head Start, perhaps because they use higher class sizes (e.g., a student-to-teacher ratio of 10 to 1 in the Tulsa program compared with around 6 or 7 to 1 in Head Start) or potentially because of differences in how cost estimates for the two types of programs account for fixed costs. For example, Gormley and Gayer report per-pupil costs for 2005 for Tulsa pre-K of \$3,500–\$6,000 for the full-day version of the program, which is less than the \$7,000 figure for Head Start that represents an average of half-day and full-day students.<sup>31</sup>

Another candidate explanation comes from the fact that the state pre-K programs that have been recently evaluated are universal, whereas Head Start is targeted mostly at very low-income children. If there are positive spillover effects from attending school with more affluent or higher-achieving children, then “peer effects” could account for part of the difference in effects between pre-K and Head Start.

A third candidate explanation for the difference in effect estimates for state universal pre-K programs and Head Start is the possibility of bias within the recent

evaluations of state pre-K programs. Although these recent state pre-K studies are major improvements over anything that has been done to examine such programs in the past, they are nonetheless all derived using a research design that may be susceptible to bias of unknown sign and magnitude. Specifically, these recent studies all use a regression discontinuity design that compares fall semester tests for kindergarten children who participated in pre-K the previous year and have birth dates close to the cutoff for having enrolled last year with fall tests of children who are just starting pre-K by virtue of having birth dates that just barely excluded them from participating the previous year. One identifying assumption here is that the selection process of children into pre-K is “smooth” around the birthday enrollment cutoff, but this need not be the case because there is a discrete change at the birthday threshold in terms of the choice set that families face in making this decision. Suppose that among the children whose birthdays just barely excluded them from enrolling in pre-K during the previous year, those with the most motivated parents were sent the previous year to private programs that are analogous to the public pre-K program and are then enrolled in private kindergarten programs in the fall semester that the pre-K study outcome measures are collected. This type of selection would reduce the share of more motivated parents among the control group in the pre-K studies and lead them to overstate the benefits of pre-K participation.

Moreover, the pre-K evaluations that have been done to date focus on those states that are leaders in this area. The experiences of pre-K programs in these states may or may not reflect the average pre-K effect we would observe if we made a wholesale shift of resources from Head Start to pre-K.

The critical policy question is whether such a shift would create the possibility of greater benefits or of harm. For now, this is an unanswerable question. The recent Head Start experimental evaluation, as well as the ongoing evaluation of Early Head Start, have pointed in the direction of beneficial effects on both cognitive and noncognitive outcome domains (e.g., social, emotional, and health outcomes), even if not all of the effect estimates are statistically significant. Previous studies have also found beneficial Head Start effects on health outcomes and on crime reduction.<sup>13,14,22</sup> Changing Head Start’s design to make the program more academic, or to look more like existing universal state pre-K programs, or even to shift Head Start funding to state programs that sometimes rely on mixed delivery systems, could generate improved academic outcomes, but the possible effects on these other important

domains of development remain unknown. Although evaluations of high-quality, intensive early childhood interventions have found positive short-term and long-term effects on social–emotional outcomes, studies focusing on community-based child care have found some unfavorable social outcomes with greater participation, especially in center-based care.<sup>32,33</sup> Studies of state-funded universal pre-K programs have not yet reported findings for social–emotional outcomes. Policy actions that would shift or withdraw resources from Head Start are therefore risky.

A different sort of risk arises from the recognition that the resources required to implement some of the changes that have been proposed for Head Start have some opportunity cost. New funding devoted to any given program change in Head Start could in principle have been devoted to other uses, including efforts to make different improvements to the Head Start program or to increase spending for other social programs that serve low-income children and families. Under the assumption of declining marginal benefits from expanding government programs, the usual efficiency standard in public economics is to invest up to the point where the marginal dollar invested generates exactly \$1 more in program benefits. By this standard, there is an efficiency argument to be made for substantially expanding existing investments in Head Start and in early childhood education more broadly.

However, the current challenge facing policy makers and analysts is to select the most effective way of allocating any new investments. Efforts to identify the active ingredients of successful programs, including Head Start, are in their infancy. Plausible candidates include the use of college-educated teachers who are paid on the usual public school salary scale, focused professional development, smaller class sizes, full-day exposure to proven curricula and instructional strategies, identification and provision of extra help for students who lag, and stepped-up support and leadership from program and school administrators.<sup>29,34,35</sup> Several of these approaches are being implemented right now. Evaluating the effects of these policy innovations should be a top priority so that future decision makers will be able to make more informed judgments about how best to expand and improve early childhood programs in the United States.

## Conclusions

There is now a body of evidence that at least suggests that Head Start generates long-term benefits and passes a benefit–cost test for children who participated

during the first few decades of the program. For the current version of Head Start, we have rigorous evidence of short-term effects from a recent experimental evaluation but no direct data on long-term effects because experimental subjects have just recently finished participating in the program.

However, there are reasons to believe that, with a cost of \$9,000 per child, Head Start does not need to yield large short-term test score effects to pass a benefit–cost test. Effect sizes of 0.1 or 0.2 might be enough, and effects even smaller than this, perhaps much smaller, might be sufficient. The estimated effects of Head Start enrollment on children—the effects of TOT—implied by the recent experimental study of the program typically exceed this threshold. Many point estimates are not statistically significant when the results are presented separately for 3- and 4-year-old participants, but pooling the two age groups yields effects that are significant for almost all of the main cognitive outcomes of interest.

We certainly do not mean to claim that Head Start is a perfect program that cannot be improved. It is possible that modifying the program in some of the ways that have been discussed in recent years, such as increasing the program's academic focus to better target those skills that predict later literacy,<sup>33</sup> or requiring teachers to hold a 4-year college degree, could make the program more effective or even more cost-effective. But, regarding efforts to more fundamentally change the structure of the program by, for example, shifting its emphasis from comprehensive services to literacy education, there is some uncertainty about the benefits that would be achieved. There is also some downside risk associated with these proposals, particularly when one recognizes that the resources required to implement them entail some opportunity cost.

In sum, the available evidence suggests to us that the Head Start program as it currently operates probably passes a benefit–cost test. Changing the program in major ways that have figured prominently in recent policy discussions may not make the program any better—and could make things worse.

### Acknowledgments

We are grateful for the support provided by the Buffett Early Childhood Fund and the McCormick Tribune Foundation to the National Forum on Early Childhood Program Evaluation through the National Scientific Council on the Developing Child.

Thanks to Kathryn Clabby and Matt Sciandra for outstanding research assistance. Helpful comments were provided by Steve Barnett, Jeanne Brooks-Gunn,

Philip Cook, Thomas Cook, Janet Currie, William Dickens, Greg Duncan, Dave Frisvold, Katherine Magnuson, Gillian Najarian, Matthew Neidell, Helen Raikes, Lonnie Sherrod, Jack Shonkoff, Hiro Yoshikawa, and Marty Zaslow. Special thanks to Ronna Cook at Westat for making available additional information about the first-year randomized Head Start evaluation. Portions of this article draw on a social policy report published by the Society for Research in Child Development (Ludwig and Phillips, 2007a). Any errors and all opinions are of course our own.

### References

1. SHONKOFF, J. & D.A. PHILLIPS. 2000. *From Neurons to Neighborhoods*. National Academies Press. Washington, DC.
2. KNUDSEN, E.I., J.J. HECKMAN, J.L. CAMERON & J.P. SHONKOFF. 2006. Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences* **103**: 10155–10162.
3. SCHWEINHART, L.J., J. MONTIE, Z. XIANG, *et al.* 2005. *Life-time Effects: The High/Scope Perry Preschool Study Through Age 40*. High/Scope Press. Ypsilanti, Michigan.
4. CAMPBELL, F.A., C.T. RAMEY, E. PUNGELLO, *et al.* 2002. Early childhood education: young adult outcomes from the Abecedarian Project. *Applied Development Science* **6**: 42–57.
5. BELFIELD, C.R., M. NORES, W.S. BARNETT & L.J. SCHWEINHART. 2006. The High/Scope Perry Preschool Program: cost-benefit analysis using data from the age-40 followup. *Journal of Human Resources*. **XLI**: 162–190.
6. MASSE, L.N. & W.S. BARNETT. 2007. Comparative benefit-cost analysis of the Abecedarian Program and its policy implications. *Economics of Education Review* **26**: 1–144.
7. LUDWIG, J. & I. SAWHILL. 2007. *Success by Ten: Intervening Early, Often and Effectively in the Education of Young Children*. Washington, DC: Brookings Institution, Hamilton Project Discussion Paper 2007-02.
8. WESTINGHOUSE LEARNING CORPORATION. 1969. *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development*. Executive Summary. Ohio University Report to the Office of Economic Opportunity. Washington, DC: Clearinghouse for Federal Scientific and Technical Information, June 1969.
9. VINOVSIS, M.A. 2005. *The Birth of Head Start: Preschool Education Policies in the Kennedy and Johnson Administrations*. University of Chicago Press. Chicago.
10. OLSEN, D. 2000. *It's Time to Stop Head Start*. Human Events, September 1, 2000.
11. BESHAROV, D.J. 2005. *Head Start's Broken Promise*. American Enterprise Institute, On the Issues.
12. CURRIE, J. & D. THOMAS. 1995. Does Head Start make a difference? *American Economic Review* **85**: 341–364.

13. GARCÉS, E., D. THOMAS & J. CURRIE. 2002. Longer term effects of Head Start. *American Economic Review* **92**: 999–1012.
14. LUDWIG, J. & D.L. MILLER. 2007. Does Head Start improve children's life chances? Evidence from a regression-discontinuity design. *Quarterly Journal of Economics* **122**: 159–208.
15. PUMA, M., S. BELL, R. COOK, *et al.* 2005. Head Start Impact Study: First Year Findings. Westat. Report Prepared for the U.S. Department of Health and Human Services.
16. THUN, M.J. & A. JERMAL. 2006. How much of the decrease in cancer death rates in the United States is attributable to reductions in tobacco smoking? *Tobacco Control* **15**: 345–347.
17. GRUBER, J. & B. KOSZEGI. 2002. A theory of government regulation of addictive bads: optimal levels and tax incidence for cigarette excise taxation. NBER Working Paper 8777. Cambridge, MA.
18. GRUBER, J. & S. MULLAINATHAN. 2002. Do cigarette taxes make smokers happier? NBER Working Paper 8872. Cambridge, MA.
19. COHEN, J. 1977. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press. New York.
20. BLOOM, H.S. 2005. Randomizing groups to evaluate place-based programs. *In Learning More from Social Experiments: Evolving Analytic Approaches*. Howard S. Boom, Ed. Russell Sage Foundation. New York.
21. CURRIE, J. 2001. Early childhood education programs. *Journal of Economic Perspectives* **15**: 213–238.
22. FRISVOLD, D. 2007. Head Start participation and childhood obesity. Paper presented at the Allied Social Science Association Meetings, January 2007 Chicago.
23. CURRIE, J. & D. THOMAS. 1995. Does Head Start make a difference? *American Economic Review* **85**: 341–364.
24. LUDWIG, J. & D.A. PHILLIPS. 2007. The benefits and costs of Head Start. Cambridge, MA: National Bureau of Economic Research Working Paper 12973.
25. YOSHIKAWA, H. 2005. Placing the first-year findings of the National Head Start Impact Study in context. Brief prepared for the Society for Research in Child Development. <http://srcd.org/documents/policy/Impactstudy.pdf>.
26. BLOOM, H.S. 1984. Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **8**: 225–246.
27. ROCK, D.A. & A.J. STENNER. 2005. Assessment issues in the testing of children at school entry. *The Future of Children* **15**: 15–34.
28. MURNANE, R.J., J.B. WILLETT & F. LEVY. 1995. The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* **77**: 251–266.
29. GORMLEY, W.T., T. GAYER, D. PHILLIPS & B. DAWSON. 2005. The effects of universal pre-K on cognitive development. Working Paper, Georgetown University, Center for Research on Children in the United States.
30. BARNETT, W.S., C. LAMY & K. JUNG. 2005. The effects of state prekindergarten programs on young children's school readiness in five states. Rutgers University, National Institute for Early Education Research.
31. Gormley, W. & T. Gayer. 2005. Promoting school readiness in Oklahoma: an evaluation of Tulsa's pre-K program. *Journal of Human Resources* **XL(3)**: 533–558.
32. MAGNUSON, K.A., C.J. RUHM & J. WALDFOGEL. 2004. Does prekindergarten improve school preparation and performance? NBER Working Paper 10452. Cambridge, MA.
33. ZASLOW, M. 2006. Issues for the Learning Community From the First Year Results of the Head Start Impact Study. Plenary Presentation to the Head Start Eighth National Research Meeting, June 27, 2006.
34. SAWHILL, I. 2006. Opportunity in America: the role of education. Brookings Institution. Washington, DC.
35. CURRIE, J. & M. NEIDELL. 2007. Getting inside the 'black box' of Head Start quality: what matters and what doesn't? *Economics of Education Review* **26**: 83–99.