NBER WORKING PAPER SERIES

THE BENEFITS AND COSTS OF HEAD START

Jens Ludwig
Deborah A. Phillips

Working Paper 12973
http://www.nber.org/papers/w12973

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2007

The authors 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 for outstanding research assistance. Very helpful comments were provided by Philip Cook, Janet Currie, William Dickens, Greg Duncan, Dave Frisvold, Katherine Magnuson, Gillian Najarian, Matthew Neidell, Helen Raikes, 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. Any errors and all opinions are of course our own. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

The Benefits and Costs of Head Start
Jens Ludwig and Deborah A. Phillips
NBER Working Paper No. 12973
March 2007
JEL No. H43,I2,I3

ABSTRACT

In this essay 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. Our conclusions differ importantly from those offered in some previous reviews because we use a more appropriate standard to judge the success of Head Start (namely, benefit-cost analysis), draw on new accumulating evidence for Head Start's long-term effects on early cohorts of program participants, and discuss why common interpretations of a recent randomized experimental evaluation of Head Start's short-term impacts may be overly pessimistic. While 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.

Jens Ludwig
Georgetown University
Public Policy Institute
3520 Prospect Street, NW
Washington, DC 20007
and NBER
ludwigj@georgetown.edu

Deborah A. Phillips
37th and O Streets, NW
Georgetown University Department of Psychology
Washington, DC 20057
deborah.dap4@gmail.com

I. INTRODUCTION

Few social programs seem to enjoy the widespread popular support of Head Start.[1] Poor children represent a sympathetic target population for social programs [Mayer, 1997]. The powerful correlation between the socio-economic status of a child's family and their long-term life chances raises concerns about the fairness of American society as well as the social costs associated with outcomes such as school dropout or criminal behavior [Holzer et al., 2007]. And the fact that the educational deficits of poor children show up very early in the life course, even well before children start kindergarten, provides a powerful logic for intervening early [Knudsen et al., 2006].

Despite the obvious emotional and intuitive appeal of Head Start, questions about Head Start's effectiveness have dogged the program since its inception. The first study claiming that Head Start's benefits fade out was published in 1966 – just one year after the program was launched as part of President Lyndon B. Johnson's War on Poverty [Westinghouse, 1969; Vinovskis, 2005]. Skepticism about the program persists in some quarters to the present day. The perspective of the Cato Institute's Darcy Ann Olsen is nicely summarized by the title of her policy brief: *It's Time to Stop Head Start*.[2] Douglas J. Besharov's assessment is slightly more charitable, but only slightly, as suggested by the title of his policy brief: *Head Start's Broken Promise*.[3] Understanding

[1] For example in a 1995 opinion poll, two-thirds of voters in Colorado agreed with the statement "School districts should be required to provide special preschool classes to children from low income households to get them prepared for grade school" (Memorandum to the Donnell-Kay Foundation from Public Opinion Strategies, September 7, 2005; available at www.pos.org/presentations/77/preschool.pdf). The fraction of voters agreeing with this statement exceeds the fraction of Colorado votes in the 1996 presidential elections that went to Bill Clinton by nearly 22 percentage points. As another example, Head Start funding has increased dramatically under both Republican and Democratic presidential administrations (particularly the administrations of George HW Bush and Bill Clinton); see Haskins [2004].

[2] Darcy Ann Olsen, "It's Time to Stop Head Start," Human Events, September 1, 2000.

[3] Douglas J. Besharov, "Head Start's Broken Promise," American Enterprise Institute for Public Policy Research On the Issues, October 2005.

Head Start's impacts on poor children seems particularly urgent given that the program is up for re-authorization this year in the U.S. Congress.

This essay reviews what is known about the value of Head Start. Our bottom line is that the best available evidence suggests Head Start passes a benefit-cost test. While 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 pre-K programs. However the benefits of such changes remain uncertain and they entail some downside risk.

Our essay 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 impacts. 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 impacts of Head Start on early cohorts of program participants, as well as about the short-term program impacts on more recent cohorts of children. Research on Head Start's long-term impacts suggests the program passed a benefit-cost test during the first few decades of operation [Currie and Thomas, 1995; Garces, Thomas and Currie, 2002; Ludwig and Miller, 2007]. These findings counter the view that only very intensive (and expensive) early childhood interventions can generate long-term benefits, and also 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 impacts of contemporaneous government programs – we can only estimate long-term impacts 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 impacts measured within one year of random assignment, which was sponsored by the federal government and carried out by Westat [Puma et al., 2005]. 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) impact. These impacts are presented by Westat separately for 3 and 4 year old program participants and are usually in the direction consistent with some beneficial impact of Head Start on children's short term outcomes, but are often not statistically significant.

The third objective of our paper is to provide some benchmarks for how large these short-term impacts would need to be in order to believe that any long-term benefits generated by today's Head Start program will be 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 non-cognitive skills of young children translate into long-term life outcomes. With this caveat in mind, the evidence that is available suggests that given Head Start's costs (around $7,000 per child on average), the program would pass a benefit-cost test if the short-term impacts on achievement test scores were equal to around .1 to .2 standard deviations, or maybe even much smaller still.

If our achievement test-score benchmarks are correct, then the expected benefits from Head Start are likely to exceed the program's costs. The effects of being assigned to the Head Start experimental treatment group (the intent to treat effects) are typically on the order of .1 to .2 standard deviations. However the estimated effects of Head Start participation itself (the "effects of treatment on the treated," or TOT) are estimated to be around 1.5 times as large as the intent-to-treat impacts. The difference between the effects of being assigned to the experimental treatment group and the effects of Head Start participation *per se* (that is, ITT versus TOT) stems from the fact that not all treatment-group children actually enrolled in Head Start, while some children assigned to the control group wound up enrolling in Head Start on their own.

Some observers have focused on the fact that many of the estimated impacts in the recent randomized experimental evaluation of Head Start are not statistically significant, and so follow the usual scientific convention of assuming that any estimates that cannot be statistically distinguished from zero are zero. Our own calculations suggest that if Westat had conducted analyses that pooled the 3 and 4 year old samples, instead of only presenting estimates separately for 3 versus 4 year olds, almost all of the estimated Head Start impacts on the main cognitive outcomes of interest would be statistically significant. But more importantly, as Cook and Ludwig [2006] argue, the expected value of a program's benefits and costs may be a more relevant framework for making policy decisions than statistical significance, and the expected net value of Head Start is positive. That is, odds are that Head Start passes a benefit-cost test for current cohorts of children.

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- 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. Given available evidence the expected net value of changing Head Start is ambiguous.

II. THE BENEFITS OF BENEFIT-COST ANALYSIS

The argument that we should judge the magnitude of Head Start's impacts by how the dollar value of these benefits compare 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 impacts is relative to the scale of the social problem that is being addressed. For example Besharov [2005] 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 of a program that serves almost 900,000 children at a cost of $9 billion a year."

But the right standard of success for a public program is not the elimination of a social problem. Consider, for example, that mortality rates from lung cancer in the U.S. in 2003 remain quite high – equal to 71.9 deaths per 100,000 people for males and 41.2 deaths per 100,000 for females [Thun and Jemal, 2006, p. 346]. 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 since diverting smokers from smoking appears to make them happier as well as healthier [Gruber and Koszegi, 2002; Gruber and Mullainathan, 2002].

Psychologists and education researchers often use the typology offered by Jacob Cohen [1977], who argues that an "effect size" (that is, program impact expressed as a share of a control group standard deviation) of .2 should be considered "small," while effect sizes of .5 should be considered "medium" and those of .8 or more are "large." Lipsey [1990] conducts a meta-analysis that draws on results from 6,700 studies in education and other related areas and finds that the empirical distribution of estimated effect sizes roughly corresponds to Cohen's categorization [see also Bloom, 2005]. This is the convention adopted by Westat in their report on the short-term results of the recent randomized Head Start experiment.[4]

Yet any assessment of what a program accomplishes should take into account not just the program's benefits but also its costs, which necessarily requires conversion of both into some common metric – that is, benefit-cost analysis. A program that improved test scores by .8 standard deviations – "large" in Jacob Cohen's [1977] scheme – but cost a total of $10 trillion per year would be difficult to support, since undertaking such an early childhood intervention would absorb the majority of the nation's gross domestic product with very little left to house, clothe, feed and protect the nation's child (and adult) population. At the other extreme a program that generated impacts on the order of

[4] The Westat report calls effect sizes below 0.2 standard deviations "small," while those of 0.2 to 0.5 standard deviations are "medium" and those over 0.5 are "large" [see Puma et al., 2005, p. ii, footnote 1].

.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.[5]

III. EVIDENCE ON HEAD START'S LONG-TERM IMPACTS

While researchers have been studying Head Start for over 40 years, only in recent years have social scientists made real headway in identifying the causal impacts of the program on participating children. There is now credible evidence on Head Start's long-term impacts that suggest the program passed a benefit-cost test over the first few decades of operation, but the program is changing over time and so these impacts might not be relevant to today's Head Start. Short-term impacts 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 obviously only be identified 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 the problem of 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, 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 are unable to

[5] For an excellent discussion of these points see Duncan and Magnuson [2006]. Harris [2007] presents a cost-effectiveness framework for judging program impacts and suggests that any intervention that generates increased test scores of .025 standard deviations per child per $1,000 spending should be considered

adequately measure all aspects of family disadvantage. The opposite bias may result if instead the more motivated and effective parents are the ones who are able to get their children into (or are selected by program administrators for) scarce Head Start slots.

Economists Eliana Garces, Duncan Thomas and Janet Currie [2002] evaluate Head Start by comparing the experiences of siblings who did and did not participate in the program. 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.

The research design employed by Garces and colleagues 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 impacts if parents enroll their more (or less) able children to participate in the program.

The Garces study might also understate Head Start's impacts if there are positive spillover effects of participating in the program on other members of the family, since in this case 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). In addition, their study relies on retrospective self reports of Head Start participation by people who have reached adulthood, which some people may

---

"large." The implication is that Head Start impacts of .175 standard deviations would be "large" under this framework, roughly consistent with our own benchmarks presented below.

misremember or misreport. If this measurement error is uncorrelated with respondents' characteristics and potential outcomes then misreporting will cause their estimated Head Start impacts to be attenuated to some degree (that is, biased towards 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 impact estimates are equal to around one-quarter and one-half of the "control mean." For African-Americans the estimated Head Start impact 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 very large effect.[6]

Ludwig and Miller [2007] 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 impacts for blacks as well as whites. Their design exploits a discontinuity in Head Start funding across counties generated by the way that the program was launched in 1965. Specifically, the Office of Economic Opportunity (OEO) 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

[6] The share of all children ever booked or charged with a crime in their data is 9.7% for the full sample and 10% for the sibling sample. These figures do not imply that Head Start achieves more than a 100% reduction in crime for program participants, since the right comparison for the estimated Head Start effect

that Head Start participation and funding rates are 50 to 100% higher in the counties with poverty rates that just barely put them into the group of the 300 poorest counties compared to 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 an increase in schooling attainment of about one-half year, and an increase in the likelihood of attending some college of about 15% of the control mean. 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 since the Head Start funding discontinuity across counties at the heart of this research design seems to have dissipated over time.

Taken together, these impact 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 might be at least as large as the 7-to-1 figure often cited for model early childhood programs such as Perry Preschool. Currie [2001] 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

---

on African-American participants is the average arrest rate for the siblings of these children, which does not seem to be reported in the study.

retention might together offset between 40 and 60 percent of the program's costs. Ludwig and Miller's [2007] 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 easily outweigh the extra program spending.[7] In addition Frisvold [2007] provides some evidence that Head Start might reduce childhood obesity.

These findings run counter to the common view that only very intensive and expensive early childhood interventions are capable of generating long-term benefits. The origin of this conventional wisdom is itself not entirely clear, since there is no logical reason that lower-cost programs will necessarily have lower benefit-cost ratios compared to those from higher-cost programs. The impacts 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. This was based in part on Head Start's origin as part of the Community Assistance Program of the War on Poverty with its emphasis on involvement of the poor in the design and implementation of new social programs [Vinovskis, 2005], including roles as classroom teachers and aides. 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

[7] Ludwig and Miller [2007] estimate the impact of an additional $400 per four year old in Head Start funding in a county. The dollar value of the decline in child mortality is equal to around $120 per four year old in the county. They also estimate an increase in schooling attainment of around one-half year per child. Card [1999] suggests an extra year of schooling increases earnings by 5 to 10 percent. We conservatively assume the extra $400 in Head Start funding raises lifetime earnings by 2 percent per child, which Krueger [2003] shows is worth at least $15,000 in present value using a 3 present discount rate (even assuming no productivity growth over time). The benefits would be even larger if we accounted for the fact that increased schooling also seems to reduce involvement with crime [Lochner and Morretti, 2004], and that the costs of crime to society are enormous – perhaps as much as $2 trillion per year [Ludwig, 2006].

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 impacts 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 since parent educational attainments and real incomes have increased since the 1960s and state-funded pre-school programs have been introduced. It is not clear which environment is improving more rapidly in this horse race.

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 sub-contractor [Puma et al., 2005]. Starting in 2002 nearly 4,700 three and four 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 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 seems to have been done well – randomization was implemented properly, and careful assessments were made of a wide variety of children's cognitive and non-cognitive outcomes, and parents were also studied. Response rates for both the child and parent assessments were usually around 10 percentage points lower for the

control than treatment group.[8] Note that 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 impacts of Head Start are more pronounced for more disadvantaged children, then the experimental impact estimates may under-state the effect of Head Start on the average program participant in the nation at large.

While the Head Start experiment is informative about a number of key policy questions about the program, in the absence of time travel we cannot directly measure the long-term impacts of Head Start for the children who only recently participated in the program. As a result we are forced to rely on indirect evidence about what the short-term impact estimates from the recent Head Start experiment might imply about the program's long-term effects. This point is taken up in the next section.

IV. SHORT-TERM BENCHMARKS FOR LONG-TERM SUCCESS

Head Start, as the program currently operates, spends about $7,000 per participating child.[9] How large would Head Start's short-term impacts 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, first by examining the short-term impacts 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 then by trying to assess directly the dollar value of a standard deviation

[8] Puma et al. [2005, p. 1-18] report that for the first data collection wave in Fall 2002, child response rates equaled 85% for the treatment group and 72% for the control group, and for parents equaled 89% and 81% for the treatment and control groups, respectively. For the Spring 2003 follow-up response rates for children equaled 88% and 77% for the treatment and control groups, and 86% and 79% for parents.

increase in early childhood test scores. Both approaches are subject to some uncertainty given important limitations with the available evidence. But with that qualification in mind, we believe there is a reasonable case to be made that positive impacts on achievement test scores on the order of .1 to .2 standard deviations (and perhaps even much smaller than that) would be large enough to generate long-term dollar-value benefits that would outweigh the program's costs.

A. Short-term Impacts of Yesteryear's Head Start and Perry Preschool

The findings from Garces et al. [2002] and Ludwig and Miller [2007] cited above suggest that Head Start as the program operated in the 1960s through 1980s seems to have generated long-term benefits that were larger than the program's costs. How large were the short-term impacts of Head Start on participating children, and in what outcome domains? If the short-term impacts of today's Head Start were about as large as the short-term impacts 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 in Garces et al. [2002], Currie and Thomas [1995] study children who would have been in Head Start in the 1980s or earlier and find that Head Start participation seems to increase scores on the PPVT vocabulary test by around .25 standard deviations in the short term for both white and African-American children. These impacts persist for whites, but fade out within three or four years for blacks.[10] Head Start's impacts on PIAT math scores might be around half as

[9] http://www.nhsa.org/download/advocacy/fact/HSBasics.pdf

[10] Currie and Thomas [1995, Table 6] use a sibling-difference research design and estimate a short-term effect of Head Start on PPVT test scores of nearly 7 percentile points in the national distribution for both blacks and whites. The standard deviation of percentile ranking scores (i.e. a uniform distribution with values between 1 and 100) will be around 29 points, implying short-term effect sizes in the Currie and Thomas study of around one-quarter of a standard deviation.

large and are not statistically significant [p. 345, fn 10]. [11] Ludwig and Miller [2007] find that a 50-100% increase in 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 impacts smaller than around .2 standard deviations. Nor do they have adequate sample sizes to examine impacts on test scores separately for blacks and whites. Unfortunately not much is currently known about Head Start's causal effects on short-term non-cognitive outcomes for earlier cohorts of program participants.[12]

While there remains some debate about the relative importance of different early childhood cognitive or non-cognitive skills in predicting subsequent outcomes,[13] the literature as a whole is consistent with the idea that there are multiple pathways to long-term success. For example while Duncan et al. [2005] find that early math skills are the strongest predictor of subsequent academic achievement, early reading and attention skills also predict later test scores but just not quite as strongly as do early math skills.[14]

[11] Currie and Thomas [1995], p. 345, footnote 10, note the PIAT math results are not statistically significant, but that version of the study does not report the math point estimates themselves. However an earlier version of the study, Currie and Thomas [1993], reports results for PIAT math, PIAT reading and PPVT scores but not results interacted with age, so we cannot recover short- versus long-term effects. However the overall impacts for whites for PIAT math scores are about half as large as the PPVT results, and PIAT reading scores are about 15% of the PPVT impacts.

[12] Currie and Thomas [1995, Table 4] do find some evidence that Head Start might reduce grade retention for white children who participated in the program in the 1980s or earlier.

[13] For example Duncan et al. [2005] do not find much evidence that non-cognitive outcomes measured during early childhood (aside from attention skills) predict later test scores, although other correlational studies have found that socio-emotional outcomes, notably aggressive behavior, do seem to contribute to children's achievement trajectories [Hinshaw, 1992; Jimerson, Egeland, and Teo, 1999; Miles and Stipek, 2006; Tremblay et al., 1992].

[14] These correlational data of course have important limitations in illuminating the causal relationships of early childhood outcomes with later outcomes. For example suppose that most parents read to their children, but what really distinguishes the most scholastically motivated parents from their peers is that the former try to impact math skills to their children even during the early childhood period. In this case the relatively strong correlation between early math and later scores could simply be a stand-in for the influence of parent motivation to help their children learn, and so an increase in early math skills induced by some intervention would yield longer-term impacts that are smaller than Duncan et al.'s correlations would suggest. Alternatively one can also imagine that children with early childhood socio-emotional

The fact that early childhood programs like Head Start achieve long-term behavioral impacts despite "fade out" of initial achievement test score gains suggests that lasting program impacts on non-cognitive skills might be the key drivers of long-term program impacts on outcomes such as school completion or employment [see for example Carniero and Heckman, 2003]. But it is possible that short-term boosts in academic skills are a key mechanism for improving non-cognitive skills such as motivation and persistence by for instance increasing children's confidence in school [Barnett, Young and Schweinhart, 1998]. 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 suggests that short-term impacts of around .25 standard deviations for vocabulary and perhaps .1 for math might be large enough to generate long-term benefits in excess of program costs.

We can also look at the short- versus long-term impacts of the widely-cited Perry Preschool program, which provided poor 3 and 4 year old children with two years of services at a total per-child cost of about twice that of Head Start.[15] At the end of the second year of services, Perry had increased PPVT vocabulary scores by around .91 standard deviations and scores on a test of nonverbal intellectual performance (the Leiter International Performance test) by around .77 standard deviations [Schweinhart et al., 2005, p. 61]. By age 9, the impact on vocabulary scores had faded out entirely, while around half of the original impact on nonverbal performance had dissipated. By age 14 impacts on reading and math scores are just over .3 standard deviations. Despite this

problems receive a variety of compensatory resources from their parents and schools to offset these early developmental challenges. In this case any intervention that improved children's early socio-emotional skills – holding all else constant – might have larger impacts than the Duncan et al. correlations would imply.

partial fade-out of test score impacts, Perry Preschool shows large long-term impacts on schooling, crime and other outcomes measured through age 40 [Schweinhart et al., 2005]. The dollar value of Perry Preschool's long-term benefits (in present dollars) range from nearly $100,000 calculated using a 7 percent discount rate to nearly $270,000 using a 3 percent discount rate [Belfield et al., 2006, p. 180-1].

Suppose that short-term test score impacts are proportional to the dollar value of long-term program benefits. In this case, even if we used a conservative 7 percent discount rate Head Start's short-term impacts would need to be at most around 7 percent as large ($7,000 / $100,000) as those of Perry Preschool (that is, around .08 and .05 standard deviations for vocabulary and nonverbal performance, respectively) to generate benefits that are large enough to outweigh Head Start's costs of around $7,000 per child. If we use a 3 percent discount rate instead, the necessary short-term impacts may be more on the order of .03 and .02 standard deviations, respectively.

Of course it might be possible that long-term gains are not strictly proportional to short-term impacts. For example, it could be the case that some minimum short-term impact is necessary in order to generate lasting cognitive or non-cognitive benefits. It could also be the case that the behavioral consequences of achievement impacts on the low-IQ sample of Michigan children in Perry Preschool are different from those arising from similar-sized impacts on a more representative Head Start population. But, at a minimum, the Perry Preschool data raise the possibility that "small" short-term impacts might be sufficient for a program with the costs of Head Start to pass a benefit-cost test.

B. The Value of Increasing Early Childhood Test Scores

[15] Currie [2001] cites Perry costs of $12,884 per child in 1999 dollars.

Another way to think about how large Head Start's short-term impacts would need to be in order 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 followed people from early childhood all the way through adulthood, this exercise is necessarily subject to some uncertainty. But the evidence that is available suggests that short-term effect sizes of .15 to .2 might be more than enough for Head Start to pass a benefit-cost test, consistent with the evidence from the previous section.

The British National Child Development Study (NCDS) is one of the few datasets available for this purpose, and includes achievement test scores measured at age 7 and earnings measured at age 33 for a sample of people born in the U.K. in 1958. Krueger [2003] argues that analyses of these data suggest that an increase in early childhood test scores in either reading or math of 1 standard deviation might plausibly be associated with higher lifetime earnings of about 8 percent.[16] If Krueger's argument is correct, then the short-term impacts on reading or math that would be needed to generate $7,000 in benefits from increased future earnings would be on the order of around .07 (using a 3 percent discount rate and assuming no productivity growth).[17] If we assume productivity growth of 2 percent, then the short-term impact on reading or math scores necessary to generate $7,000 in benefits could be as little as .04 standard deviations.

[16] Krueger [2003] notes that Currie and Thomas' [1999] analyses of these data imply that a 1 standard deviation increase in test scores increases lifetime earnings by around 8 percent. This impact is smaller than what has been estimated for a 1 standard deviation increase in test scores measured during adolescence for more recent US samples, which typically suggest earnings gains of around 20 percent. The difference is presumably due as Krueger notes to some combination of differences in the time period studied, the US vs UK labor markets, the fact that Currie and Thomas control for both reading and math scores simultaneously while most US studies examine one type of test score at a time in their effects on earnings.

[17] Krueger [2003] reports increased lifetime earnings from a .2 standard deviation increase in test scores using a 3 percent discount rate and assuming no productivity growth of $15,174 in 1998 dollars, equal to around $18,800 in current dollars. So the effect size required to generate $7,000 in benefits is equal to ($7,000 / $18,800)*.2 = .37*.2 = .07.

One possible objection is that we are trying to use non-experimental correlations between early test scores and adult earnings to extrapolate to earnings gains from short-term experimental impacts, which fade over time. But as noted above there is fade out in non-experimental achievement test advantage as well – that is, test scores measured in early childhood and adolescence are correlated, but imperfectly.[18] Alternatively the correlation between early childhood cognitive test scores and subsequent earnings might be different from the earnings impacts associated with a program-induced change in cognitive test scores. One reason is that the correlation between achievement test scores and earnings observed in population data may reflect, in part, the association of achievement test scores and IQ scores, for which early childhood programs have had a harder time generating lasting impacts [see for example Schweinhart et al., 2005].

On the other hand, the calculations presented above assume that the only benefit from increased early test scores is higher adult earnings. But anything that increases early childhood test scores and subsequently future earnings could affect other outcomes as well. Crime is one of the most important of these other outcomes, given that the social costs of crime might be on the order of $2 trillion per year [Ludwig, 2006]. In the Perry Preschool experiment, around two-thirds of the total dollar-value of the program's benefits came from crime reductions [Belfield et al., 2006].

There is to date no entirely satisfactory way of determining how early test score impacts relate to longer life outcomes. But the two different approaches used here both suggest that short-term impacts that would be considered very small by the usual standards of education research – on the order of .05 standard deviations or less – could

[18] Jencks and Phillips [1998, p. 28] think a plausible estimate is that the correlation between 1st and 12th grade test scores is around .52. The implication is that a child starting at the 16th percentile of the test score

potentially generate long-term benefits that would at least equal Head Start's cost per participant (around $7,000). Given the uncertainties with these calculations, a more conservative approach would be to require that Head Start improve short-term test scores by .1 to .2 standard deviations in order to believe that the program generates long-term benefits that are large enough to justify the costs.

V. HOW LARGE ARE HEAD START'S CURRENT SHORT-TERM IMPACTS?

The best available evidence on current Head Start's impacts on children comes from the Head State 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" [Besharov, 2005, p. 1], although much of the public discussion of these findings seems to confuse the intent to treat effects emphasized in Westat's report on the experimental results with the effects of Head Start participation per se (that is, the effects of treatment on the treated. The short-term impacts of Head Start participation are usually equal to or greater than the .1 or .2 standard deviation benchmark that is necessary for Head Start to pass a benefit-cost test. While many of the point estimates that Westat calculates separately for 3 and 4 year old program participants are not statistically significant, our calculations suggest that pooling data for 3 and 4 year olds leads impact estimates for almost all of the main cognitive outcome measures emphasized in the Westat report's Executive Summary to be statistically significant. But more importantly the expected value of the program is positive.

A. Intent-to-Treat Effects vs. the Effects of Head Start Participation

One common source of confusion about the recent randomized Head Start

---

distribution in first grade will on average be at the 27th percentile of the distribution in 12th grade.

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 intent to treat impact. 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 effect of treatment on the treated) can be different – sometimes quite different – from the effects of treatment-group assignment.

The problems of drawing inferences about Head Start participation from the effects of treatment-group assignment can be easily seen by imagining an example in which everyone assigned to the treatment group participates in Head Start ... but because of their own efforts, so does everyone in the control group. If the average quality of the Head Start programs experienced by children in the treatment and control groups were the same, the effects of treatment group assignment (the intent-to-treat estimate) would be equal to exactly zero. It would obviously be incorrect to infer from these estimates that Head Start does nothing to improve the life chances of participating children. The

central point is that if Head Start participation rates are less than 100% among children assigned to the treatment group or greater than 0% among those in the control group, or both, then the effects of actual Head Start enrollment (the effect of treatment on the treated) will be larger than the estimated effect of being assigned to the treatment group (the intent-to-treat effect).

In the Head Start experimental data we see that around 86% of 4 year olds assigned to the experimental treatment group enrolled in Head Start, while 18% of 4 year olds assigned to the control group wound up in Head Start on their own [p. 3-7, Puma et al., 2005].[19] The body of the report does mention that the intent-to-treat estimates will understate the effects of actually participating in Head Start. But the report's description of how it tries to convert the intent-to-treat estimates into something like an estimate for the effect of Head Start participation is confusing and the actual approach they employ might be misleading. In any case these results are relegated to one of the appendices and perhaps as a result seem to have been largely ignored in public discussions compared to the intent-to-treat estimates included in the Executive Summary.[20]

More than 20 years ago, Howard Bloom [1984] proposed a method for translating intent-to-treat effects into estimates for the effects of treatment on the treated. He noted that under some conditions we can learn about the effects of treatment participation – in this case, Head Start enrollment – by scaling differences in the treatment and control

[19] The figures for 3 year olds assigned to the treatment and control groups equal 89% and 21%, respectively.

[20] The report describes the Bloom [1984] procedure for handling "no shows" in the treatment group, but does not use this procedure to handle the problem of control group members who wind up in Head Start on their own [p. 4-29, 4-35]. Instead the report seems to drop control group families who wind up in Head Start on their own and then re-weight the remaining control group members; see pp. 4-35,6. The report mentions the Bloom [1984] approach we use to calculate TOT impacts accounting for compliance rates in both the treatment and control groups on p. 4-36 but notes only that Westat will explore how findings from this procedure compare to their default procedure in future reports.

groups in average outcomes by the difference in the treatment and control groups in treatment participation rates. This procedure assumes random assignment is in fact random, and that treatment group assignment has no effect on children who do not participate in Head Start.[21] In addition, the Bloom procedure assumes that everyone who would participate in Head Start if assigned to the control group would also participate if they had been assigned to the treatment group instead. It further assumes that the average quality of the Head Start programs attended by children assigned to the treatment versus control groups is comparable. This latter assumption may be more problematic, but even fairly large differences in Head Start program quality between Head Start enrollees in the treatment and control group would impart relatively modest bias to the estimates derived from Bloom's procedure.

Why focus on the effects of actually participating in Head Start rather than the intent-to-treat estimates? One answer is that the effect sizes for the Head Start experiment's intent-to-treat estimates are often compared to estimates from Perry Preschool, Carolina Abedarian 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 apples (TOT) -to- oranges (ITT) comparison will obviously understate the relative effectiveness of Head Start.

A more important reason for focusing on estimates for the effects of actually participating in Head Start (treatment on the treated) is to avoid confusion in conducting a benefit-cost analysis of Head Start. In public discussions about Head Start's costs, the

[21] Stated differently, the latent propensity to participate in Head Start *if* assigned to the treatment group is assumed to be equivalent for children who were, in fact, assigned to the treatment and control groups. This should be true if random assignment was in fact random, since the propensity to participate in Head Start –

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.

It is possible that Westat's report on the Head Start experiment focuses more on intent-to-treat impacts than on the effects of treatment on the treated because either Westat or the U.S. Department of Health and Human Services might be uncomfortable with presenting treatment-on-the-treated estimates that do require imposing some additional assumptions on the data beyond those necessary to calculate the intent-to-treat effects. Our own view is that even if some of the assumptions for calculating the treatment-on-the-treated impacts are not strictly true (for example if there is some difference in program quality for children enrolled in Head Start in the experiment's treatment versus control groups), the treatment-on-the-treated estimates still provide more useful approximations for the effects of actually enrolling in Head Start, and help avoid confusion along the lines described above.

Readers who are uncomfortable with the assumptions required for the treatment-on-the-treated calculations can conduct their own benefit-cost analysis using the intent-to-treat impact estimates, but must then be careful to adjust the cost side of the equation appropriately. If the difference in Head Start enrollment rates between the Head Start experiment's treatment and control groups equals (86% - 18%) = 68%, then the right "cost" for comparison to the intent-to-treat "benefit" would equal the average Head Start cost per child assigned to the HSNIS treatment group minus the average Head Start cost

---

like all other baseline characteristics – will be equally distributed between treatment and control groups (subject to sampling error).

per child assigned to the control group. This cost figure equals 68% * $7,000 = $4,760.[22]

B. Head Start's Short-Term Impacts

In Table 1 we show the ITT impacts 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 [Puma et al., 2005]. While the published Westat report did not show standard errors for impact estimates, Ronna Cook at Westat has very 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).[23]

Table 1 also presents our own estimates for the effects of actually participating in Head Start (the effects of treatment on the treated) derived using Bloom's approach together with information about Head Start enrollment rates in the experiment's treatment and control groups. 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 intent-to-treat effects that are commonly misinterpreted to represent the effects of Head Start participation. These

[22] It is easy to see that since both the costs and benefits of the ITT calculation are proportional to the TOT calculations by the treatment-control difference in Head Start enrollment rates, evidence for benefits in excess of costs for the ITT approach implies the same must be true with the TOT approach and vice versa. The important thing is to avoid comparing the dollar value of the ITT impact estimates with the costs per Head Start enrollee.

[23] In the body of the report Westat presents a series of different impact estimates for each outcome domain, including those that do not adjust for baseline characteristics, those that adjust for baseline socio-demographic characteristics only, and those that also adjust for fall outcome measures in looking at spring test scores. Because the fall outcome measures are collected mostly by mid-November (collected over the period October to December), in principle controlling for these measures could understate Head Start's impacts due to program effects that arise during the early parts of the academic year. Table 1 presents Westat's own preferred regression-adjusted point estimates and standard errors, based on Westat's examination of whether there is any evidence of program gains between the beginning of the school year and when the fall outcome measures are collected.

results are best interpreted as providing a range within which the "true" effects of Head Start likely fall. If the average Head Start program quality is somewhat higher for the treatment than control groups then our Bloom-style estimates for the effects of treatment on the treated might be biased upward somewhat.

Note also that our estimates for the effects of Head Start participation also assume that the 10 percentage point difference in response rates between the Head Start experiment treatment and control groups [Puma et al. 2005, p. 1-18] do not impart any bias to the basic intent-to-treat estimates. Of course if there is selective sample attrition that biases the basic intent-to-treat estimates, this would represent a more fundamental problem with the Head Start experiment that cannot be solved by focusing on the intent-to-treat effects rather than the effects of treatment on the treated.

Table 1 shows that at least for cognitive skills all of the Head Start impact estimates point in the direction consistent with beneficial program impacts, 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 effects of treatment on the treated estimates because we believe they are likely to be much closer approximations of the true effect of Head Start participation per se than are the intent-to-treat estimates. Nevertheless, it should be understood that the true impact is probably somewhere in between the ITT and TOT estimates.

For vocabulary, pre-reading and pre-writing skills Head Start's effects (the effects of treatment on the treated) range from .15 to .35 standard deviations, while for 4 year olds the impacts are one-third to one-half as large as for 3 year olds on the PPVT and

smaller for pre-reading and pre-writing. Parent-reported literacy skills show much more pronounced Head Start impacts, equal to .5 and .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 domain may be more reliable than those derived from parent reports.[24]

Given the findings by Greg Duncan and his colleagues that early math scores are the strongest predictor of subsequent achievement test scores, one particular concern with the Head Start experiment results has been that the impact estimates on early math scores (measured by the Woodcock-Johnson applied problems test) are not statistically significant. Head Start's impact on this test equals .18 and .15 standard deviations for 3 and 4 year olds, respectively. Duncan's study also finds that attention skills are important in predicting future test scores. The closest measure to this in the HSNIS is a variable for hyperactive behavior, where we see a Head Start impact of -.26 standard deviations for 3 year olds but a zero point estimate for 4 year olds.

A different concern that has been raised about these impact estimates comes from the ability of the available assessments to detect reliable impacts of this size in young children. One criterion we have for cognitive or non-cognitive assessments is that they are reliable – that is, they generate similar results when applied on different occasions. A standard concern is that assessments of very young children may not be very reliable, for reasons that will be obvious to anyone who has ever been the parent of a young child (short attention span, variability in temperament and willingness to cooperate, and so on).

[24] Rock and Stenner [2005, p. 21] note that for the Early Childhood Longitudinal Study of the Kindergarten Class of 1998-99 (ECLS-K) parent reports of children's social competence and skills have not proven reliable, with "the main concern [being] that parents often have little basis for determining whether behavior is age appropriate." Analogous concerns could in principle apply to parent reports about

Reliability scores for achievement tests administered to adolescents are usually on the order of .8 to .9 [see for example Murnane et al., 1995]. 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 non-cognitive skills compared to cognitive outcomes [see also Rock and Stenner, 2005].[25]

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 and the result would simply be less precise estimation of Head Start impacts (i.e., larger standard errors). This concern would provide a candidate explanation for why so many of the Head Start experimental impact estimates are not statistically significant, but does not pose a threat for interpretation of those impact estimates that are statistically significant.

C. Statistically Insignificant Impact Estimates

For policy purposes what we want to know is whether Head Start passes a benefit-cost test. Most of the estimates for the effects of Head Start participation presented in Table 1 are above the .1 to .2 standard deviation threshold we think necessary for Head Start to pass a benefit-cost test. But many of these point estimates are not statistically significant. So what should policy makers conclude about the program?

The Head Start experiment enrolled nearly 4,700 children, which is large by the

---

their children's literacy skills.

[25] The reliabilities of the different cognitive and non-cognitive tests used in the Head Start experiment are as follows (3 year old figure shown first in parentheses, followed by figure for 4 year olds; only reliabilities for pooled 3 and 4 year old samples are available for the non-cognitive outcomes): WJ Word (.87, .9); Letter naming (.96, 97); McCarthy Drawing Score (.65, .73); WJ Spelling (.74, .78); PPVT (.66, .8); Color naming (.94, .94); WJ Oral Comprehension (.8, .88); WJ Applied Problems (.9, .91); Social skills and

standards of many social program evaluations but tiny compared to many randomized clinical trials in medicine, and in any case means that the standard errors around the resulting point estimates are subject to some non-trivial 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. While this splitting of the analytic sample makes sense for developmental reasons (program impacts may differ by age; see for example Knudsen et al. [2006]), it further reduces statistical power.

Table 1 illustrates the basic issue confronting policymakers. Recall that the estimates presented by Greg Duncan and his colleagues [2005] suggest that early math scores might be the strongest predictors of children's later cognitive outcomes. Section III above suggests that short-term impacts of .1 to .2 might be large enough for Head Start to pass a benefit-cost test. The effects of Head Start participation implied by the Head Start experimental data for early math scores for 3 and 4 year olds equal .18 and .15 standard deviations, respectively – neither of which is statistically significant!

One alternative analytic approach would have been to pool the 3 and 4 year old samples in the Head Start experiment. While the Westat report does not present these analyses, with data on the separate impact estimates, sample sizes and standard deviations for the 3 and 4 year old samples we can approximate what the impact estimates and standard errors would be if Westat had pooled the two age groups together 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

---

approaches to learning (.62); Social competencies (.58); Total problem behavior (.74); Hyperactive behavior (.58); Aggressive behavior (.6); and Withdrawn behavior (.45).

standard errors are if anything conservative. Our calculations suggest that for a pooled sample of 3 and 4 year olds the Head Start impact estimate would be statistically significant for every cognitive outcome domain shown in Table 1 except oral comprehension.

Perhaps more importantly, while scientific convention is to ignore estimates that are not statistically significant at the usual 95 percent cutoff (that is, assume they are zero), we believe that a more productive way to proceed for policy purposes is to focus on the expected value of the program benefits and costs, as suggested by Cook and Ludwig [2006]. The reason is that following the course of action associated with the null hypothesis of no statistically significant impact is itself a policy decision that winds up being overly privileged if we only follow through on point estimates that meet the usual standard for statistical significance.

To see the difference, we revisit the hypothetical program we introduced in Section II above, which we assume increases children's test scores by .2 standard deviations at a cost of just a nickel per child. Suppose that a randomized experimental evaluation of this intervention yielded a point estimate for a treatment effect of .2 standard deviations, but that the standard error was somewhat large and so the p-value for this estimate was equal to .8. While no referee worth her salt would endorse a scientific manuscript that claimed that this intervention "works," at the same time she would surely wish that her own child's school district jumped at the chance to adopt this program.

This sort of expected value framework suggests that Head Start as it currently operates is likely to pass a benefit-cost test. There are good reasons to believe that short-term impacts on reading and math scores on the order of .1 to .2 standard deviations, and

perhaps much smaller than that, would be large enough for Head Start to generate benefits in excess of costs. Table 1 shows that most of the point estimates for Head Start's effects on cognitive skills for both 3 and 4 year olds are of about this magnitude, even when these estimates are not statistically significant for the two samples.

VI. HEAD START ALTERNATIVES

The fact that the current incarnation of Head Start seems 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 focused on providing a broad range of academic, health, nutrition, and social services to disadvantaged children. 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 impact 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 four-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 at the present time.

The recent Head Start experimental evaluation provides rigorous information about the short-term impacts of Head Start as it operated since the program's inception,

as a comprehensive program focused on nutrition, physical and mental health, parenting and social services as well as education. The studies of Currie and Thomas [1995], Garces, Thomas and Currie [2002], and Ludwig and Miller [2007] also provide what is to us persuasive evidence for the long-term impacts of Head Start as the program was originally designed. To date there is no evaluation evidence available about what would be achieved for current recipients by a new version of Head Start that was more academically oriented.

Several recent studies of universal state pre-K programs suggest impressively large impact estimates. Gormley et al. [2005] evaluate the effects of Tulsa, Okahoma's pre-K program and report TOT estimates equal to .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 .65 for the WJ-R spelling test (almost three times as large as those reported for four year olds in the Head Start experiment) and of .38 for the WJ-R applied problems math test (more than twice as large as for four year olds in the Head Start experiment), all of which are statistically significant. Barnett et al. [2005] examine pre-K programs in five separate states and report effect sizes of .26 for the PPVT vocabulary test and .28 for the WJ-R applied problems test, both of which are statistically significant.

What explains the difference in impact 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 four-year college degrees, while Head Start does not impose that requirement. Teachers in these state pre-K programs will presumably also have higher salaries than 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,[26] perhaps because they employ higher class sizes (for example 10:1 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.

An alternative explanation comes from the fact that the state pre-K programs that have been recently evaluated are universal while 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 impacts between pre-K and Head Start.

A third candidate explanation for the difference in impact estimates for state universal pre-K programs and Head Start is the possibility of bias within the recent evaluations of state pre-K programs. While 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 birthdates close to the cutoff for having enrolled last year with fall tests of children who are just starting pre-K by virtue of having birthdates 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 since there

[26] For example Gormley and Gayer [2005] report per-pupil costs for 2005 for Tulsa pre-K of $3,500 to $6,000 for the full-day version of the program, which is less than the $7,000 figure for Head Start that

is a discrete change at the birthday threshold in terms of the choice set that families face in making this decision.

For instance, 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 wound up being 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. Presently, this is an unanswerable question. The recent Head Start experimental evaluation, as well as the on-going evaluation of Early Head Start, have pointed in the direction of beneficial impacts on both cognitive and non-cognitive outcome domains (e.g., social, emotional, and health outcomes), even if not all of the impact estimates are statistically significant. Previous studies have also found beneficial Head Start impacts on health outcomes and on crime reduction [Garces et al., 2002; Ludwig and Miller, 2007; Frisvold, 2007]. 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

---

represents an average of half- and full-day students.

systems could potentially generate improved academic outcomes, but the possible impacts on these other important domains of development remain unknown. While evaluations of high-quality, intensive early childhood interventions have found positive short- and long-term impacts 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 [Magnuson, Ruhm and Waldfogel, 2004; Zaslow, 2006]. Studies of state-funded universal pre-K programs have not yet reported findings for social-emotional outcomes. As a result policy actions that would shift or withdraw resources from Head Start are risky.

It is important to recognize that a different kind of risk from changing Head Start comes from the fact that the resources required to implement some of the proposed changes have some opportunity cost, since the funding in question could in principle have been devoted to other uses, including other social programs. For instance, an increasingly common proposal is to require Head Start teachers to hold 4 year college degrees. This change would require higher salaries to recruit and retain more highly-educated teachers, which would require either more spending for the Head Start program as a whole or else reductions in other parts of the Head Start budget. Even knowing that requiring BA-level teachers leads to improved student outcomes would not be sufficient to endorse this policy from an economist's perspective. We would want to know how these gains compare to what could be achieved from devoting those extra resources to other uses such as further reducing class sizes in Head Start,[27] expanding the program's coverage to more eligible low-income children, or improving pre-natal health and

outreach services to low-income women. Given these downside risks, it is possible to determine whether alternative uses of Head Start funding that have been proposed have positive or negative expected value.

VII. CONCLUSIONS

There is credible evidence that Head Start generates long-term benefits and passes a benefit-cost test, at least 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 impacts from a recent experimental evaluation but no direct data on long-term effects since experimental subjects have just recently finished participating in the program. However there are reasons to believe that with a cost of $7,000 per child Head Start does not need to yield very large short-term test score impacts in order to pass a benefit-cost test. Effect sizes of .1 or .2 might be enough, and impacts even smaller than this, perhaps much smaller, might be sufficient. The estimated effects of Head Start enrollment on children – the effects of treatment on the treated – 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 the expected value of the program is still positive.

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 [Zaslow, 2006], or requiring teachers to hold a four-year college degree, could make the program more effective or even more

[27] Currie and Neidell [forthcoming] suggest that redirecting resources to increase teacher qualifications and salaries within the existing Head Start budget at the expense of small class sizes would on net lead to worse

cost-effective. But there is some uncertainty about the benefits that would be achieved by such changes, and there is some downside risk associated with each of 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 various ways that have figured prominently in recent policy discussions may not make the program any better, and could make things worse.

---

student outcomes.

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 ITT	3 year olds TOT	4 year olds ITT	4 year olds TOT
Woodock-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	.120* (.052)	.17* (.077)	.051 (.052)	.075 (.076)
Color naming	.098* (.043)	.144* (.064)	.108 (.071)	.159 (.107)
Parent-reported literacy skills	.340* (.066)	.499* (.097)	.293* (.075)	.435* (.112)
Oral comprehension	.025 (.062)	.036 (.091)	-.058 (.052)	-.086 (.077)
Woodcock-Johnson applied problems	.124 (.083)	.182 (.122)	.100 (.070)	.147 (.103)

First and third columns reproduce ITT impact 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, i.e. program impacts divided by the control group standard deviation (Puma et al., 2005). Standard errors are shown in parentheses 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 treatment on the treated (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 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 percent cutoff.

References

Barnett, W. Steven and Kenneth B. Robin. (Undated) "How much does quality preschool cost?" Rutgers University, National Institute for Early Education Research.

Barnett, W. Steven, Cynthia Lamy and Kwanghee 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.

Barnett, W. Steven, J.W. Young and L.J. Schweinhart (1998) "How preschool education influences long-term cognitive development and school success." In W.S. Barnett and S.S. Boocock, Eds. *Early care and education for children in poverty: Promises, programs, and long-term results* (pp. 167-184). Albany: State University of New York Press.

Belfield, Clive R., Milagros Nores, Steve W. Barnett and Lawrence J. Schweinhart (2006) "The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Followup." *Journal of Human Resources*. XLI(1): 162-190.

Besharov, Douglas J. (2005) "Head Start's Broken Promise." American Enterprise Institute, On the Issues.

Bloom, Howard S. (1984) "Accounting for no-shows in experimental evaluation designs." *Evaluation Review*. 8(2): 225-46.

Bloom, Howard S. (2005) "Randomizing groups to evaluate place-based programs." In *Learning More from Social Experiments: Evolving Analytic Approaches*, Edited by Howard S. Boom. NY: Russell Sage Foundation.

Card, David (2001) "Estimating the returns to schooling: Progress on some persistent econometric problems." *Econometrica*. 69(5): 1127-60.

Carniero, Pedro and James J. Heckman (2003) "Human Capital Policy," In *Inequality in America: What Role for Human Capital Policies?* James J. Heckman and Alan B. Krueger. Cambridge, MA: MIT Press.

Cohen, Jacob (1977) *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cook, Philip J. and Jens Ludwig (2006) "Aiming for evidence-based gun policy." *Journal of Policy Analysis and Management*. 25(3): 691-736.

Currie, Janet (2001) "Early Childhood Education Programs." *Journal of Economic Perspectives*. 15(2): 213-238.

Currie, Janet and Duncan Thomas (1993) "Does Head Start Make a Difference?" Cambridge, MA: NBER Working Paper 4406.

Currie, Janet and Duncan Thomas (1995) "Does Head Start Make a Difference?" *American Economic Review*. 85(3): 341-364.

Currie, Janet and Duncan Thomas (1999) "Early test scores, socioeconomic status, and future outcomes." NBER Working Paper 6943.

Currie, Janet and Matthew Neidell (forthcoming) "Getting Inside the 'Black Box' of Head Start Quality: What Matters and What Doesn't?" *Economics of Education Review*.

Duncan, Greg J., Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huyston, Pamela Klebanov, Linda Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Crista Japel (2006) "School Readiness and Later Achievement." Working Paper, Northwestern University.

Duncan, Greg J. and Katherine Magnuson (forthcoming) "Penny size and effect size foolish." *CDP*.

Frisvold, David (2007) "Head Start Participation and Childhood Obesity." Paper presented at the Allied Social Science Association Meetings, January 2007, Chicago.

Garces, Eliana, Duncan Thomas, and Janet Currie (2002) "Longer Term Effects of Head Start." *American Economic Review*. 92(4): 999-1012.

Gormley, William T. and Ted Gayer (2005) "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program." *Journal of Human Resources*. XL(3): 533-558.

Gormley, William T., Ted Gayer, Deborah Phillips and Brittany Dawson (2005) "The effects of universal pre-K on cognitive development." Working Paper, Georgetown University, Center for Research on Children in the United States.

Gruber, Jonathan and Botond Koszegi (2002) "A Theory of Government Regulation of Addictive Bads: Optimal Levels and Tax Incidence for Cigarette Excise Taxation." NBER Working Paper 8777.

Gruber, Jonathan and Sendhil Mullainathan (2002) "Do Cigarette Taxes Make Smokers Happier?" NBER Working Paper 8872.

Harris, Douglas N. (2007) "New benchmarks for interpreting effect sizes: Combining effects with costs." Working Paper, University of Wisconsin at Madison.

Haskins, Ron (2004) "Competing Visions." *Education Next*.

Hinshaw, SP (1992) "Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms." *Psychological Bulletin*. 111: 127-154.
Holzer, Harry, Diane Whitmore Schanzenbach, Greg J. Duncan, and Jens Ludwig (2007) *The Economic Costs of Poverty*. Washington, DC: Center for American Progress.

Jencks, Christopher and Meredith Phillips (1998) "The black-white test score gap: An introduction." *The Black-White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips. Washington, DC: Brookings Institution Press. pp. 1-54.

Jimerson, S., Egeland, B., & Teo, A. (1999). A longitudinal study of achievement trajectories: Factors associated with change. *Journal of Educational Psychology, 91*(1) 116-126.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff (2006) "Economic, neurobiological, and behavioral perspectives on building America's future workforce." *Proceedings of the National Academy of Sciences*. 103: 10155-10162.

Krueger, Alan B. (2003) "Economic considerations and class size." *Economic Journal*.

Lochner, Lance and Enrico Moretti (2004) "The effect of education on crime: Evidence from prison inmates, arrests, and self-reports." *American Economic Review*. 94(1): 155-189.

Ludwig, Jens (2006) "The Costs of Crime." Testimony to the U.S. Senate Judiciary Committee, September, 2006.

Ludwig, Jens and Douglas L. Miller (2007) "Does Head Start Improve Children's Life Chances? Evidence from a Regression-Discontinuity Design." *Quarterly Journal of Economics*. 122(1): 159-208.

Magnuson, Katherine A., Christopher J. Ruhm, and Jane Waldfogel (2004) "Does prekindergarten improve school preparation and performance?" NBER Working Paper 10452.

Mayer, Susan E. (1997) *What Money Can't Buy*. Cambridge, MA: Harvard Press.

Miles, Sarah B. and Deborah Stipek (2006) "Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children." *Child Development*. 77(1): 103-117.

Murnane, Richard J., John B. Willett, and Frank Levy (1995) "The growing importance of cognitive skills in wage determination." *Review of Economics and Statistics*. 77(2): 251-266.

Phillips, D., McCartney, K., & Sussman, A. (2006). Child care and early development.

In McCartney, K., & Phillips, D. (Eds.), *The Handbook of Early Child Development*. Blackwell Publishers.

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, Michael Lopez, et al. (2005) *Head Start Impact Study: First Year Findings*. Westat. Report Prepared for the U.S. Department of Health and Human Services.

Rock, Donald A. and A. Jackson Stenner (2005) "Assessment Issues in the Testing of Children at School Entry." *The Future of Children*. 15(1): 15-34.

Schanzenbach, Diane Whitmore (forthcoming) "What have researchers learned from Project STAR?" *Brookings Papers on Education Policy*.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield and Milagros Nores, *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*. (Ypsilanti, Michigan: High/Scope Press, 2005).

Thun, Michael J. and Ahmedin 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.

Vinovskis, Maris A. (2005) *The Birth of Head Start: Preschool Education Policies in the Kennedy and Johnson Administrations*. Chicago: University of Chicago Press.

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.

Zaslow, Martha (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.