NBER WORKING PAPER SERIES

DOES HEAD START DO ANY LASTING GOOD?

Chloe Gibbs
Jens Ludwig
Douglas L. Miller

Working Paper 17452
http://www.nber.org/papers/w17452

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2011

This is the working paper version of a chapter that will appear in The War on Poverty: A 50-Year Retrospective, edited by Martha Bailey and Sheldon Danziger. Writing of the paper was supported in part by a visiting scholar award from the Russell Sage Foundation to Jens Ludwig. Thanks to Laura Brinkman for research assistance and to David Deming, Jacob Vigdor and seminar participants at the American Economic Association meetings, the Brookings Institution, and MDRC for helpful discussions. All opinions and any errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Does Head Start Do Any Lasting Good?
Chloe Gibbs, Jens Ludwig, and Douglas L. Miller
NBER Working Paper No. 17452
September 2011
JEL No. I21,I38

ABSTRACT

Head Start is a federal early childhood intervention designed to reduce disparities in preschool outcomes. The first randomized experimental study of Head Start, the National Head Start Impact Study (NHSIS), found impacts on academic outcomes of .15 to .3 standard deviations measured at the end of the program year, although the estimated impacts were no longer significant when measured at the end of kindergarten or first grade. Assessments that Head Start is ineffective based on the NHSIS results are in our view premature, given our currently limited understanding of how and why early childhood education improves long-term life chances. Many of the specific changes to Head Start that have been proposed could potentially wind up doing more harm than good.

Chloe Gibbs
University of Chicago
1155 East 60th Street
Chicago, IL60637
chloeh@uchicago.edu

Jens Ludwig
University of Chicago
1155 East 60th Street
Chicago, IL 60637
and NBER
jludwig@uchicago.edu

Douglas L. Miller
University of California, Davis
Department of Economics
One Shields Avenue
Davis, CA 95616-8578
and NBER
dlmiller@ucdavis.edu

I. INTRODUCTION

Head Start is an early childhood education, health, and parenting intervention started in 1965 as part of the War on Poverty, which currently serves up to 900,000 mostly low-income children per year with total federal spending of over $7 billion (Haskins and Barnett, 2010). Recent results from the first randomized experimental study of Head Start, the National Head Start Impact Study (NHSIS), showed impacts on achievement test scores measured at the end of the program year on the order of .15 to .3 standard deviations, although impacts were no longer statistically significant in first grade. In response Joe Klein (2011) of *Time Magazine* argues the evidence is "indisputable" that "Head Start simply does not work," and that continued funding is "criminal, every bit as outrageous as tax breaks for oil companies." Ron Haskins (2010) argues "taxpayers get little for their annual investment of $8 billion in Head Start." W. Steven Barnett argues in a recent paper in *Science* (2011, p. 977) that Head Start yields "poor results" and might need to "focus more resources on the classroom to recruit and retain better teachers." The Obama Administration proposes to start measuring classroom quality of Head Start, and require low-quality programs to re-compete for – and possibly lose – funding.

In this paper we argue that these negative assessments of Head Start's lasting benefits for children are premature, given how little we currently understand about how and why early childhood interventions improve long-term life chances. The initial impacts of Head Start found in the NHSIS for recent cohorts of program participants are about the same size as those estimated for children who participated in the program in the 1960s through 1980s. For those cohorts of Head Start participants, the program seems to have produced long-term benefits large enough to outweigh program costs – *despite* fade-out of initial test-score impacts (Garces et al., 2002, Ludwig and Miller, 2007, Deming, 2009).

The main surprise from the NHSIS is *not* that Head Start's initial impacts on children are

"too small," *nor* that these test score impacts attenuate over time, but rather that the test score impacts attenuate so rapidly. For cohorts of children who were in Head Start in the 1960s through 1980s, the program's impacts on test scores seemed to persist at least through the early elementary school grades (Currie and Thomas, 1995, Deming, 2009). In contrast for the NHSIS study sample of children who were 3-4 years old in 2002, Head Start's initial impacts on test scores are no longer significant one year after children leave the program.

Is this faster rate of fade out worrisome? The pattern of short-term impacts and rate of fade out in the NHSIS is not so different from what has been found in a recent meta-analysis of the larger early-childhood intervention literature (Leak et al., 2010). Recent longitudinal studies show that even early childhood interventions whose initial test score impacts fade out within a single year can still – remarkably – generate lasting impacts on long-term outcomes such as adult earnings (for example, Chetty et al., forthcoming). By process of elimination, researchers typically assume the mediator driving long-term behavioral impacts must be non-academic (that is, socio-emotional and behavioral) skills. But almost none of the studies that follow people from early childhood into adulthood have good measures of these socio-emotional and behavioral skills. These skills currently play the role of what one might call "social policy dark matter."[1]

Given the widespread assumption that rapid fade-out of test score impacts is a problem for Head Start, policy analysts and policymakers have proposed numerous changes to the program. How much good can be accomplished by transforming Head Start depends in part on how inefficient the current program is. The key question is how much more human capital promotion is possible with annual expenditures of $9,000 per child (Ludwig and Phillips, 2007b). What will result from modifying Head Start may also depend on why the program's

[1] Previous studies do find some evidence that these interventions change outcomes like grade retention. This outcome reflects a mix of academic and non-academic skills, and so is far from a direct measure of social-cognitive skills like future orientation or impulse control.

impacts on test scores are attenuating so rapidly for recent cohorts of children as suggested by the NHSIS.

Barnett (2011) believes that there is substantial room for improvement in Head Start in part because he believes that much larger test-score impacts have been found with other early childhood programs that don't cost much more than Head Start, particularly newer state-sponsored universal pre-K programs. While ultimately Barnett could turn out to be right, the evidence that is currently available to date has not yet convinced us that universal pre-K programs generate substantially more output than does Head Start. First, the research design that has been used to study state pre-K programs is vulnerable to omitted variables bias. Second, the studies of state pre-K programs are not informative about whether impacts fade out – the key concern about Head Start.

Moreover, several plausible explanations for what seems to be increasingly rapid attenuation in Head Start effects suggest that many of the proposed changes to Head Start may be unproductive or even counterproductive. For example, one candidate explanation for why Head Start's impacts on test scores are fading out more rapidly now compared to previously is that the size of the program's initial impacts on socio-emotional and behavioral skills are not as stable over time compared to the apparent stability of the program's initial impacts on academic achievement test scores. A common assumption is that Head Start's impacts on academic outcomes should be proportional in magnitude to the program's impacts on social-emotional outcomes, but in principle this need not be the case. The recent NHSIS results show few statistically significant impacts on socio-emotional and behavioral outcomes, even for very short-term measures collected at the end of the Head Start program year. If Head Start impacts on social-emotional and behavioral skills help "prop up" lasting achievement test gains, and/or are the key mediators for Head Start's long-term impacts on schooling attainment and other adult life

outcomes, then any decline over time in Head Start's impacts on social-emotional outcomes could be quite costly from the perspective of improving poor children's life outcomes.

Unfortunately this explanation is not directly testable, since we do not have good measures of social-emotional and behavioral skills for previous cohorts of Head Start children. As discussed below, previous observational studies provide mixed evidence for the importance of early social-cognitive skills in predicting later academic test scores, but suggestive evidence for the importance of social-cognitive skills for long-term outcomes like schooling attainment.

The possibility that effects on social-emotional and behavioral skills are an important mediator for early childhood effects on other longer-term outcomes, together with the possibility that Head Start's effects on social-cognitive and behavioral skills could be declining over time, suggests that proposals to make Head Start "more academic" could actually exacerbate the problem of Head Start fade-out and reduce the program's overall effectiveness. And as noted above, the Obama Administration is proposing to use data-driven accountability reforms to "weed out" the least effective Head Start programs. The success of this type of proposal will depend in part on our ability to identify programs that have high value-added for children. The possibility that social-emotional and behavioral skills could be a key mediator of Head Start's impacts means that any performance-review and accountability system might benefit from focusing on outcomes beyond just academic skills. This would be complicated by our limited understanding of exactly which social-emotional and behavioral skills are most important.

A different candidate explanation for why Head Start's impacts on test scores fade out more rapidly now than in the past focuses on changes over time in the quality of the elementary schools that children attend after leaving Head Start. Many people assume that low-quality elementary schools squander the benefits of Head Start. However, in a previous paper by two of us (Ludwig and Miller, 2007), we find long-term benefits even for a group of children – African-

Americans living in the poorest parts of the Deep South in the 1960s – who were attending public schools of an average quality that would, by any fair assessment, be termed appalling.

If changes in elementary schools are indeed contributing to accelerating fade-out of Head Start impacts, then a more plausible version of this explanation is that elementary schools are getting better at remediating skill deficits for children who are lagging behind. This hypothesis would be consistent with increases over time in reading and math scores for 9 year olds in the National Assessment of Educational Progress (NAEP) and narrowing of black-white gaps in NAEP scores. This hypothesis is also consistent with indications that fade-out from early childhood interventions seems to be less rapid in developing country contexts, where remediation efforts may be less common or less effective (Barnett, 2011). What looks like "fade out" in the NHSIS might actually be "catch up" by low performing non-Head Start children. Put differently, in this case part of Head Start's benefits accrue to children who do not participate in the program, by enabling teachers to focus extra classroom time remediating their skill deficits. In this model, the best way to address what looks like Head Start "fade out" might be to actually expand Head Start enrollment.

The next section reviews previous research on Head Start and other relevant early childhood interventions. The third section discusses the evidence about whether other early childhood programs that have similar costs to Head Start have been shown to generate larger or more lasting impacts. Section four considers candidate explanations for why Head Start's impacts on test scores may be fading out more rapidly over time, while section five discusses the implications for different reform proposals that have been offered for Head Start.

II. PREVIOUS RESEARCH ON HEAD START

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. The first true randomized experimental study of Head Start (the NHSIS) only started to track children who began participating in the program in 2002, so evidence about Head Start's longer-term benefits will for the foreseeable future necessarily come from studies that use other research designs. But a growing body of quasi-experimental research provides credible evidence (in our view) that Head Start generated lasting benefits for children in the first few decades of the program's existence, and that these benefits are likely large enough to justify program costs. In what follows we first summarize this quasi-experimental literature before turning to a discussion of results from the recent NHSIS experiment.

A. Evidence on Head Start's long-term impacts[2]

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. For example, Head Start recipients come from more disadvantaged families than other children. If researchers are unable to adequately measure and control for all aspects of family disadvantage then simple comparisons of Head Start recipients to other children may understate the program's effectiveness. 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) studied the long-

[2] This section draws heavily from Ludwig and Phillips (2007a).

term effects of Head Start by comparing the experiences of siblings who did and did not participate in the program. Their sample consists of children who would have participated in Head Start in 1980 or earlier, using data from the Panel Study of Income Dynamics (PSID). 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 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 misremember or misreport. This measurement error may also bias the impact estimates.

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 (high school) and one-half (college) 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 is a very large effect.[3]

Deming (2009) studies the long-term effects of Head Start by applying the same sibling-difference design to data from the children of the National Longitudinal Survey of Youth (CNLSY), which follows a national sample of children who would have participated in Head Start between 1984 and 1990 (the same sample for whom medium-term impacts were estimated by Currie and Thomas, 1995). CNLSY children were administered the Peabody Picture Vocabulary Test (PPVT) sometime between the ages of three and five, and then again sometime after age 10. Children were also administered the Peabody Individual Achievement Math (PIATMT) and Reading Recognition (PIATRR) subtracts every survey year for those ages 5-14.

Table 1 in our paper replicates Table 4 from Deming (2009), and shows that compared to siblings who did not participate in any other form of preschool, those enrolled in Head Start had higher average PIAT scores (averaging together the math and reading tests) of .145 standard deviations measured at ages 5-6, .133 standard deviations measured at ages 7-10, and .055 measured at ages 11-14. Initial impacts on the PIAT tests are larger for blacks than for whites or Hispanics, but show suggestive signs of fading out more rapidly for blacks than non-blacks.[4]

---

[3] 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 Head Start achieves more than a 100% reduction in crime, since the right comparison for the estimated Head Start effect 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.

[4] Currie and Thomas (1995) use the same research design and CNLSY dataset as Deming (2009), but for obvious reasons have a shorter follow-up window given the much earlier date of their study. They find that 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. More specifically, Currie and Thomas (1995, Table 6) 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. Head Start's impacts on PIAT math scores might be around half as large and are not statistically significant (p. 345, fn 10). 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

Despite the fade-out in test scores for the overall study sample, Deming estimates Head start impacts on an index of long-term outcomes (high school graduation, college-going, idleness, crime, teen parenthood and health, all measured after age 18) equal to .23 standard deviations.

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 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. One limitation of their study is that the datasets available to measure schooling attainment identify county of residence at the time the schooling measures are obtained, not county of residence at the time people would have been of Head Start age.

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

---

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.

control mean. Importantly, the estimated effects of extra Head Start funding on educational attainment are found for both blacks and whites. Ludwig and Miller (2007) find that this 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 as measured in the National Education Longitudinal Survey of 1988 (NELS), 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. Their estimates are calculated for children who would have participated in Head Start during the 1960s or 1970s, and cannot be calculated for more recent birth cohorts 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 seems to have generated benefits in excess of program costs, despite fade-out in initial achievement test impacts, 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 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.[5] In addition, Frisvold (2007) provides some

[5]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

evidence that Head Start might reduce childhood obesity.

Barnett (2011) suggests that this pattern of results from the previous Head Start studies – long-term impacts on behavioral outcomes despite fade-out of impacts on achievement test scores – may simply reflect "the high probability of false positives when many researchers conduct such studies" (p. 976). What Barnett has in mind is that the ratio of published studies with estimated impacts that are significant at the 5 percent threshold divided by the total number of studies carried out is not much more than 5 percent, and that this may not be apparent from carrying out a review of the literature because many studies with statistically insignificant findings may never get published – the "file-drawer" problem that causes the denominator in this ratio to look smaller than it actually is.

While acknowledging that two of us (Ludwig and Miller) are not entirely disinterested participants in any discussion of the plausibility of the quasi-experimental literature about Head Start's long-term effects, we do not find Barnett's false positive argument very persuasive. There are just not that many longitudinal datasets that include the two necessary ingredients to carry out a study of Head Start's long-term impacts – information on participation in Head Start during early childhood, plus long-term follow-up data on later life outcomes. We ourselves only know of three national datasets that meet these criteria: PSID, CNLSY, and NELS – that is, the three data sources that have been used in the long-term quasi-experimental studies mentioned above. Only two datasets include sibling pairs in the sampling frame (PSID and CNLSY) – the two datasets that have been analyzed this way.[6] There cannot be a "false positive" problem if there is not a large number of estimates hidden away in file drawers somewhere.

B. The National Head Start Impact Study (NHSIS)

---

trillion per year (Ludwig, 2006).

[6]The sample frame for Add Health, which interviewed a national sample of 7-12 graders in 1994-5, did include sibling pairs and twins. But Add Health did not collect data about whether people in the study sample participated in Head Start or not (from personal correspondence of Jens Ludwig with Kathie Harris on August 31, 2011).

The impacts of Head Start on children depend in part on the size of the disparities in outcomes across children that arise during the preschool years, which have been declining over time,[7] and in part on the difference in the developmental quality of the program versus the quality of the environments that low-income children would have experienced otherwise. Over time the Head Start program has improved in quality, but arguably so has the alternative to Head Start for poor children.[8] It is not clear which environment is improving more rapidly in this horse race, which also means we cannot necessarily forecast the effects of today's Head Start from studies of the program in the past.

Fortunately, the federal government sponsored a true randomized experimental study of Head Start, the National Head Start Impact Study (NHSIS) (see Puma et al., 2005, 2010).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.[9]

---

[7] Data from the mid-1960s showed that black-white gaps in achievement test scores were fully 1.5 standard deviations measured in first grade (Coleman et al., 1966). More recent studies suggest that early black-white gaps are on the order of .8 to 1.0 standard deviations measured during the preschool years (Jencks and Phillips, 1998). Data from the more recent Early Childhood Longitudinal Study of Kindergartners (ECLS-K) show smaller gaps still, but that may be due to the narrow focus of the ECLS-K tests (Murnane et al., 2006).

[8]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 program were even less developmentally productive than Head Start. Over time the quality of the Head Start program has improved, but arguably so have the alternatives to Head Start for poor children: parent educational attainments and real incomes have increased since the 1960s and alternative forms of center-based early education, such as state-funded pre-school programs, have been introduced.

[9]The design of the NHSIS raises a subtle point about external validity: 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

The experiment seems to have been done well – randomization was implemented properly, careful assessments were made of a wide variety of children's cognitive and non-cognitive outcomes, and parents were also studied. As we would expect with random assignment, baseline characteristics for the children assigned to the treatment group look quite similar to those of the children assigned to the control group. Response rates for both the child and parent assessments were usually around 5 to 10 percentage points lower for the control than treatment group.[10]

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 summaries that accompany the three- or four-hundred page technical reports, are intention-to-treat (ITT) estimates that are not *intended* to reflect the effects of actual Head Start participation. These ITT estimates will not equal the effects of actually participating in Head Start (the effects of treatment on the treated, or TOT) in the NHSIS data because not all children assigned to the program group wind up in Head Start, while some of those assigned to the control group get into Head Start on their own. Specifically, 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 (Puma et al., 2005, p. 3-7).[11] If we are willing to assume that the average quality of the Head Start centers attended by those in the treatment group is the same as for the control group crossovers, and that randomization to the treatment group had no effect on the treatment group other than by affecting Head Start enrollment likelihoods, then the

---

participant in the nation at large.

[10] 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. Puma et al. (2010, p. 2-19) reports response rates for the 4-year-old cohort in spring 2005 (first grade) for the child assessment that are equal to 79% for the treatment group and 73% for the control group, and for the parent interview, equal to 82% and 75% for the treatment and control groups, respectively.

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

TOT effect will be around 1.5 times as large as the ITT estimates (Bloom, 1984, Angrist, Imbens and Rubin, 1996).[12]

In Table 2 (taken from Ludwig and Phillips, 2007a), 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),[13] as well as the TOT effects that come from re-scaling the ITT effects by the difference in the treatment and control groups in Head Start enrollment rates. If the Head Start programs treatment-group children attend are better than the Head Start programs the control group attends, this Bloom-style TOT estimate will somewhat overstate the effects of participating in Head Start. Point estimates and standard errors have been divided by the control group standard deviation for the relevant outcome measure so that they can be compared to other studies reporting results as "effect sizes."[14]

Table 2 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 somewhat larger (both absolutely and in relation to their standard errors) for 3 year olds than 4 year olds. For vocabulary, pre-reading and pre-writing skills Head Start's effects (i.e., TOT) range from .15 to

[12] The initial first-year report (Puma et al., 2005) 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 procedure in future reports.

[13] 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.

[14] 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.

.35 standard deviations. 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. It is not necessarily obvious whether student assessments are more or less reliable than those derived from parent reports.[15] Impacts on the Woodcock-Johnson applied math problems test are equal to .18 and .15 standard deviations for 3 and 4 year olds, respectively, but not statistically significant. Ludwig and Phillips (2007a,b) note that if one pooled the 3 and 4 year old cohorts in the NHSIS and analyzed them together, rather than separately, Head Start impacts would be statistically significant for every outcome shown in Table 2 except for oral comprehension.

Social-emotional outcomes are also potentially important targets for early childhood interventions as well, beyond academic outcomes. For example, research by Greg Duncan and colleagues has shown that early childhood measures of attention skills are important in predicting future test scores. The closest measure to this in the NHSIS is a variable for hyperactive behavior, where we see a Head Start impact of -.26 standard deviations for 3 year olds but a point estimate of essentially zero for 4 year olds. It is worth noting that the reliability of the social-emotional measures collected in the NHSIS (that is, the degree to which the assessments yield the same results when administered at different times to the same child) tend to be somewhat lower than for the NHSIS academic measures.[16]

---

[15]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 their children's literacy skills. On the other hand, tests provide just a snapshot of what children can do, and that snapshot suffers further from noise due to student nervousness, stereotype threat or any other stochastic shock that affects a student's performance on the test.

[16]A standard concern with assessing young children is that their performance on the test might not be a good indication of what they really know because of factors that vary minute-to-minute with young children: short attention span, variability in temperament and willingness to cooperate, and so on. 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). 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

Figure 1 compares Head Start impacts by age for the 4 year old cohort in the NHSIS[17] with the estimated Head Start impacts by age from Deming (2009), who as noted above studied a sample of children who would have been in Head Start no later than around 1990. Head Start impacts measured around age 5 are fairly similar for the recent cohort of Head Start children studied in the NHSIS and the earlier cohorts of children examined by Deming. But the figure shows that the impacts attenuate quite rapidly in the NHSIS, and by the end of first grade are very small in magnitude. In contrast in Deming's study we see estimates that are measured between the ages of 7-10 that are of about the same magnitude as those estimated at ages 5-6.

Figure 2 presents the results separately by race, pooling together Hispanics and whites in one panel, and separate results for blacks in the other. In Deming's study, results are larger initially for blacks than non-blacks, but fade out more rapidly for blacks than non-blacks (the p-value on Deming's test of equality of Head Start impacts by age is p=.003 for blacks and p=.24 for non-blacks; see also Currie and Thomas, 1995). In the NHSIS, Head Start's impacts seem to fade out *less* rapidly for blacks than non-blacks. This could be a fluke finding or it could be a useful diagnostic for understanding why attenuation over time in Head Start impacts is occurring more rapidly overall for more recent cohorts of Head Start participants.

Many people have assumed that the rate of test-score fade-out in the NHSIS implies that the program can have no long-term lasting benefits. The recent paper by Chetty et al. (forthcoming) of the long-term effects of kindergarten quality provides a striking example of how that conclusion might be premature. Figure 3 below reproduces Figure 6 from Chetty et al. Panel A shows the estimated effect of a one standard deviation increase in kindergarten quality

---

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

[17] We focus in the figure on the 4 year old cohort in the NHSIS because such a large share of control children in the 3 year old cohort received Head Start during the second year of the experiment, which complicates interpretation of the results for this group.

(as measured by teacher value-added) on the test scores of children measured in different grades. The figure shows that the initial impacts are about 6 percentile points, which would be just over about .2 standard deviations.[18] That impact declines by more than two-thirds by first grade.

Panel B of Figure 3 shows the results of regressing test scores measured at different grades against adult outcomes from IRS earnings data collected when study participants are in their late 20s. The first data point shows that if we used children's test scores measured in kindergarten, we would predict that being in a kindergarten of higher quality by 1sd translates into increased annual earnings of $588. If we looked at those children in first grade, when the effect of kindergarten quality was no longer really evident in their first grade scores, then if we knew nothing about their kindergarten test scores we would have concluded that children in different quality kindergarten classrooms should have similar earnings. However, the actual long-run impact on earnings (shown by the last dot, labeled "E") is $483. In this application, there is strong test score fade out, but still long run impacts – and these impacts are best predicted by the short-term test score impacts.

III. HOW INEFFICIENT IS HEAD START?

A key question is whether it is possible to substantially increase the amount of "output" that Head Start produces through budget-neutral changes to the program. Barnett (2011) believes the answer is yes, in part because of the impact estimates that have been found for state-sponsored universal pre-K programs. We are less convinced by the pre-K evidence, and are instead struck by the qualitatively similar findings in the NHSIS compared to a recent meta-analysis of a variety of early childhood interventions.

Almost all of the existing universal pre-K studies have used the same research design that

[18] The standard deviation for a variable like percentile rankings that has uniform distribution from 1 to 100 will about 28 percentile points.

exploits an age discontinuity that determines whether children are eligible for pre-K in a given academic year or not until the subsequent academic year. The results from these studies appear to be quite impressive. Gormley et al. (2005) evaluate the effects of Tulsa, Okahoma's pre-K program. Gormley and colleagues 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. Wong et al. (2008) report average effect sizes for these five states of .17 for the PPVT, .26 for the math scores, and fully .68 for a test of print awareness.

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.

Study Samples for State Pre-K RD Evaluations

	Year of eligibility for pre-K	
Volunteer for pre-K?	Year T	Year T+1
No	A	B
Yes	C	D

Most study samples in this literature rely on sample frames of children who attend pre-K in year T versus year T+1, and so compare basically outcomes for children in cells C and D above who have birthdays that are "close" to the cutoff. 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 we know that there is something else that changes discontinuously at the age threshold for eligibility for the program in year T – namely, the choice set that families face in deciding whether to send their children to pre-K or not.

For instance, suppose that among the children whose birthdays just barely excluded them from enrolling in pre-K during year T, the parents of the most academically ready children respond to being prevented from sending their children to pre-K that year by giving up on the public system and sending their children to private pre-K in year T and then private kindergarten in year T+1. These children would not be in the study sample of children who are assessed in the fall of year T+1 for any of these age-discontinuity pre-K studies. This type of selection would reduce the share of high ability children among the control group in the pre-K studies and lead them to overstate the benefits of pre-K participation. One fix to this selection concern would be to expand the sample frame to be the population of children in a given age cohort in a given

jurisdiction, rather than limit the study sample to children who actually participate in pre-K,[19] but to date no study we know of has followed that data collection strategy.

Another relevant point about these age-discontinuity pre-K studies is that by their nature they are not capable of producing evidence about the degree of fade-out of pre-K impacts. Note that the one-year follow-up analysis would occur in year T+2 in our example above, comparing children who participated in pre-K in year T and then kindergarten in year T+1 with children who participated in pre-K in year T+1 – thereby becoming a test of the effects of kindergarten, rather than about whether pre-K impacts persist over time. Put differently, the pre-K studies are unavoidably silent about fade out. And fade out is the key issue that has generated concern about Head Start.

An alternative comparison for Head Start comes from Leak et al's (2010) meta-analysis of a variety of early childhood programs. It should be said that there is some uncertainty about whether their meta-analysis accurately captures what we should expect for initial impacts and rate of decay of initial test score impacts with the mix of early childhood programs included in their study sample, since a large share of the studies use a quasi-experimental design that could still be subject to some bias.[20] But we still think there is some useful information in their results.

Table 3 in our paper reproduces some key results table from their study (Leak et al.'s Table 4). The table presents the results of a meta-analytic regression that is run on a sample consisting of regression estimates from individual early childhood studies, conducted at different points in calendar time. Around forty percent of these studies are from randomized experiments, the rest use different quasi-experimental designs. Each row in the table presents a different

[19] That is, we have in mind comparing children in the pooled cells A and C above with the children in pooled cells B and D above, who have birthdays "close" to the cutoff, which would be analogous to an intention-to-treat analysis in a randomized experiment.

[20] The next-to-last column in the table shows that studies that do not use random assignment have initial program impacts that are .07 standard deviations lower than what is observed for experimental studies (.21 versus .28), although the Leak et al. analysis does not make it possible to examine whether the research design affects the rate at which impacts decay with time since the time of program participation.

characteristic of a study, such as characteristics of the program being studied or the study's research design.

The first column of Table 3 shows that the average early childhood program had an effect size of .28 standard deviations measured at the end of the treatment itself, not much different from the average achievement impact of .21 we observe in the NHSIS for the 4-year-old cohort of children (averaging across all achievement measures). Interestingly, Table 3 also shows that whether the study was carried out before or after 1980 does not seem to have much effect on the estimated initial program effect, despite the fact that the counterfactual early childhood environments of low-income children have presumably changed over time.

The final column of Table 3 controls for study fixed effects. This enables us to measure the within-study rate of decay of treatment effects. This analysis suggests that almost the entire effect of these early childhood programs dissipates the year after the program – a picture that is not so different from what we see in the NHSIS experimental study of Head Start.

IV. EXPLANATIONS FOR ACCELERATING HEAD START "FADE-OUT"

The effects of Head Start reform proposals will depend on what the reason is for accelerating Head Start "fade-out" over time. In what follows we consider two types of explanations: changes over time in Head Start's impacts on non-academic outcomes – that is, socio-emotional and behavioral skills; and changes over time in elementary school quality.

A. Changing impacts on socio-emotional and behavioral skills

Forecasting what the short-term results from the NHSIS imply for the long-term outcomes of low-income children is complicated by the fact that we currently know relatively little about how or why Head Start or other early childhood interventions such as improved kindergarten quality generate lasting gains in schooling attainment or other key life outcomes.

Our literature review above shows that for older cohorts initial test score gains seem to decay in whole or part, yet long-term outcomes seem to improve for program participants. This mirrors the pattern found with other early childhood interventions like Perry Preschool and Abecedarian.

Moreover, little is known about the relationship or potential interplay between program impacts on academic skills and social-cognitive skills. Previous correlational studies of the relationship between early skills and later skills suggest that there are multiple pathways to success – both academic and social-cognitive skills measured early in life seem to predict academic outcomes measured later in life.[21] Research discussions about early childhood programs typically seem to assume that program impacts on academic skills are proportional to impacts on social-cognitive skills – that is, programs that generate relatively larger gains in academic skills generate relatively larger gains in social-cognitive skills as well. Whether this is actually true in practice is hard to determine in a world in which far too few studies actually try to directly measure socio-emotional and behavioral skills.

If we believed that initial program effects on social-emotional and behavioral skills are key to sustaining program effects on academic test scores, then one candidate explanation for why test score impacts seem to fade out more rapidly in the NHSIS could be changes over time in Head Start's impacts on social-emotional and behavioral skills. For this explanation to work,

[21]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. Duncan et al. (2005) do not find much evidence that other social-cognitive skills 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). 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 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 early socio-emotional skills – holding all else constant – might have larger impacts than the Duncan et al. correlations would imply.

we would need to rule out the possibility that short-term boosts in academic skills are a key mechanism for improving socio-emotional and behavioral skills such as motivation and persistence by, for instance, increasing children's confidence in school (e.g., Barnett, Young and Schweinhart, 1998).

What we do know from the NHSIS is that there are relatively few statistically significant effects on what that report refers to as "socio-emotional" outcomes such as aggressive, hyperactive, withdrawn or other problem behavior, social competencies, social skills, "closeness," conflict and positive relationships, particularly for the 4 year old cohort in the NHSIS study sample (Puma et al., 2010). In principle these variables might miss those social-emotional and behavioral skills that are most important for children's life success, either because of measurement error or because the variables try to measure the wrong things.

The possibility that Head Start's impacts on social-emotional and behavioral skills has declined over time means that policy proposals to make Head Start "more academic" could in principle exacerbate the underlying problems that contribute to test-score fade out. This concern is also relevant to proposals to make greater use of data and performance measures to identify and eliminate low-value-added Head Start programs, since any such assessment regime would ideally capture those short-term skills that are most important for long-term life outcomes – quite a challenge given we know so little about what those key short-term skills actually are!

B. Changes Over Time in Elementary School Environments

A common hypothesis is that Head Start gains fade-out for children because they attend low-quality elementary schools after exiting the Head Start program. For example, Currie and Thomas (2000) analyze data from the NELS:88 and find that black children who participated in Head Start and attended an elementary school with higher average test scores have higher achievement test scores than black children who had been in Head Start but attended elementary

schools with lower average test scores. Currie and Thomas note that it is difficult with their data to distinguish between the role played by school quality versus peer "quality" in explaining this difference. Their analysis is suggestive but not definitive, since there remains the possibility that there is some self-selection of higher-scoring Head Start children into higher-scoring schools. (As Currie and Thomas also note, the use of the NELS:88 data also makes it challenging to isolate the main effect of Head Start on children's outcomes as well).

One challenge to the explanation that low elementary school quality squanders Head Start gains comes from the findings by Ludwig and Miller (2007) of what appear to be long-term gains in schooling attainment even among African-Americans living in the poorest counties of the Deep South in the 1960s. It is hard to imagine that very many poor children in modern-day America attend schools that are actually worse than those that states like Mississippi provided to blacks living in the Delta 40 or 50 years ago.

More generally, low elementary school quality could only explain accelerating Head Start "fade-out" if elementary schools in the U.S. overall were declining in their average quality. But by at least some indications the reverse may be occurring. For example, Figure 4 presents average reading scores on the National Assessment of Educational Progress (NAEP) for 9 year old children over the period from 1971 to 2008. Overall reading scores are increasing over time. Moreover, the black-white gap in average reading scores seems to be narrowing. Figure 5 shows that we see a similar pattern for math scores.

This pattern of increased elementary school quality, particularly for relatively more disadvantaged students (in Figures 4 and 5, minorities), suggests an alternative hypothesis for how changing elementary school conditions might be related to accelerated attenuation of Head Start impacts: Perhaps elementary schools are becoming increasingly effective at remediating academic deficits among children who enter kindergarten or first grade behind their peers. That

is, perhaps what looks like "fade out" among Head Start children is actually "catch up" by low-skilled controls (non-participants). It might be the case that elementary school teachers focus more and more on helping students who did not participate in Head Start to catch up. If individualized teacher attention is developmentally productive, then Head Start generates social returns (in the form of reducing the number of students for whom the teacher needs to do remediation) that may exceed the private returns. Part of the benefits of Head Start come in the form of improved outcomes for non-Head Start children, and so would be invisible in a study that compares the average outcomes of Head Start and non-Head Start children.

Testing among the "fade-out" and "catch-up" hypotheses is challenging in part because children's achievement test scores change a great deal naturally as children age during their earliest years of life. Setting aside for now the considerable technical problems associated with comparing children's achievement levels at different ages, we see that children in both the Head Start and control groups in the NHSIS are experiencing considerable increases in overall achievement levels over time. By definition the rate of increase is slower for the treatment than control groups (since the initial Head Start impact attenuates over time), but from this fact alone we cannot conclude whether poor school quality is suppressing the trajectories of Head Start children or whether remedial efforts by teachers are accelerating the trajectories of non-Head Start children.

A different way to test for catch up versus fade out is to try to directly examine the degree to which teachers try to "smooth" outcomes between Head Start and other children. Unfortunately, most existing datasets are not very good at capturing very detailed classroom process measures like the amount of time teachers spend with particular individual students.

The rising NAEP scores among 9 year olds and narrowing of black-white gaps provide evidence that is consistent with the "catch-up" hypothesis. Another data point consistent with the

"catch-up" hypothesis is that as Barnett (2011) notes, meta-analyses of early childhood interventions done in developing country contexts – where we assume elementary school quality is not as strong as in developed countries – do not show the same sort of fade out as in the U.S.

Whether convergence in treatment-control outcomes in the NHSIS is due to "fade out" versus "catch up" is important for policy design. Fade-out is an argument for focusing on efforts to improve elementary schools. Catch-up provides an argument for measures such as increasing enrollment rates in Head Start or other early childhood programs. In the catch-up scenario, tracking in the early elementary grades might also help preserve Head Start benefits.

V. CONCLUSIONS

Many people believe that early childhood interventions represent a promising way to invest in the human capital of disadvantaged children and improve their long-term life chances. However, the follow-up results of the NHSIS experiment have led several analysts to conclude that Head Start is in need of changes, or might even best be folded together with other early childhood programs such as state pre-K. The arguments outlined in this paper suggest that the pessimism surrounding Head Start is premature, given what we currently know about how and why early childhood programs generate long term gains to low-income children.

The most striking feature of the NHSIS results is not the size of Head Start's initial test-score impacts, but what happens after children exit from the Head Start program. Our lack of understanding about why there is more rapid test score convergence makes it difficult to identify the benefits and costs that would be associated with different policy responses.

For example, one plausible explanation for NHSIS test score convergence is that the program may be having less pronounced effects on social-emotional and behavioral skills. If this were true it would mean that making Head Start "more academic" could exacerbate the key

problem with Head Start.

Similarly, if social-emotional and behavioral skills could be the key mediator for long-term benefits from Head Start, this has implications for another common policy proposal: to rely more on data and performance measurement to identify and de-fund Head Start programs that have low value-added (see for example Haskins and Barnett, 2010, Ramey and Ramey, 2010, and the Obama Administration's new proposals). Implicit in these proposals is the ability to accurately measure the relevant skills. Ensuring that practitioners are measuring and appropriately weighting the right social-emotional and behavioral skills is challenging given that the reliabilities of the social-emotional measures in the NHSIS seem to be lower than for academic outcome measures, and given that the research community has limited understanding at the present time about what the key skills and mediators are that matter most for driving long-term program benefits.

Another proposal from Haskins and Barnett (2010) is to make it easier for states and localities to coordinate or co-mingle funding from different early childhood programs, like Head Start, the child care block grant, Earl Head Start, Title I, etc. This is a step in the direction of other proposals that have been made to rely more and more on state-sponsored universal pre-K programs, which a growing number of studies suggest have impressive short-term impacts on achievement test scores. Eventually social science research might demonstrate that universal pre-K programs produce larger short-term impacts than Head Start, but the existing studies of pre-K do not yet convince us of this point. Moreover the research design used in the current batch of pre-K studies by their nature cannot tell us anything about whether pre-K impacts fade out or not. Most researchers who think pre-K dominates Head Start must assume that the size of a program's impact on initial test scores is directly proportional to how much skill impacts persist over time (bigger initial impacts mean bigger impacts later on in elementary school). But the fact

that Head Start's initial test score impacts seem to be about the same for recent cohorts of children as for those who were in Head Start several decades ago, yet the rate at which these impacts attenuate has changed, would seem to invalidate the assumption that initial impacts are always predictive of what sorts of impacts we can expect down the road.

Every educational program of which we are aware (including those at our home universities) always has some room for improvement. There is no reason to think that Head Start is an exception in this regard; after all, the program began nearly 50 years ago in a social environment that is quite different from today's. The main point of this essay is to emphasize the uncertainty that remains in our understanding of what today's version of Head Start accomplishes – and what might result from different proposed changes to the program. Specifically the current state of research leaves great uncertainty about key factors: how much more "output" we could expect for current Head Start expenditures; the exact mechanisms through which early childhood programs generate lasting benefits to people; and what is causing Head Start's impacts on test scores to attenuate more rapidly over time (and whether that is even a worrisome trend). Many of the proposals that have been made to modify Head Start have the very real prospect of having no value or even negative value.

Table 1. The Effect of Head Start Overall and by Subgroup (replication of Table 4 from Deming, 2009)

	Test scores				Nontest score	Long term
	5-6	7-10	11-14	5-14	7-14	19+
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Overall						
Head Start	0.145*	0.133**	0.055	0.101	0.265***	0.228***
	(0.085)	(0.060)	(0.062)	(0.057)	(0.082)	(0.072)
Other preschools	–0.079	0.048	–0.022	–0.012	0.172*	0.069
	(0.085)	(0.065)	(0.069)	(0.062)	(0.088)	(0.072)
p (HS = preschool)	0.021	0.254	0.315	0.118	0.372	0.080
Panel B: By race						
Head Start (black)	0.287***	0.127*	0.031	0.107	0.351***	0.237**
	(0.095)	(0.075)	(0.076)	(0.072)	(0.120)	(0.103)
Head Start (white/Hispanic)	–0.057	0.111	0.156	0.110	0.177	0.224**
	(0.120)	(0.092)	(0.095)	(0.090)	(0.111)	(0.102)
p (black = nonblack)						
Panel C: By gender						
Head Start (male)	0.154	0.181**	0.141**	0.159**	0.390***	0.182*
	(0.107)	(0.079)	(0.081)	(0.076)	(0.123)	(0.103)
Head Start (female)	0.128	0.059	0.033	0.055	0.146	0.272**
	(0.106)	(0.083)	(0.085)	(0.081)	(0.108)	(0.106)
p (male = female)	0.862	0.287	0.357	0.346	0.135	0.553
Panel D: By maternal AFQT score						
Head Start (AFQT≤ -1) (n=361)	0.171	0.016 –	0.023	0.015	0.529***	0.279**
	(0.129)	(0.095)	(0.102)	(0.094)	(0.156)	(0.114)
Head Start (AFQT > -1) (n=890)	0.133	0.172**	0.144*	0.154**	0.124	0.202**
	(0.094)	(0.073)	(0.074)	(0.071)	(0.091)	(0.091)
p (low = high AFQT)	0.809	0.198	0.192	0.245	0.024	0.595
Panel E: P-values for equality of tests scores by age group						
	Black	Nonblack	Male	Female	Low AFQT	High AFQT
p (all effects equal)	0.003	0.240	0.262	0.254	0.198	0.205

Notes: All results are reported using the specification in column 5 of Table 3, which includes a family fixed effect, all pre-treatment covariates, and controls for gender, age, and firstborn status. Race and gender subgroup estimates are obtained by interacting the Head Start treatment effect with a full set of dummy variables for each subgroup. Standard errors are in parentheses and are clustered at the family level. The test score indices include the PPVT and PIAT Math and Reading Recognition tests. The nontest score index includes indicator variables for grade retention and learning disability diagnosis. The long-term outcome index includes high school graduation, college attendance, idleness, crime, teen parenthood, and self-reported health status.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table 2. Intention-to-treat (ITT) Effect Sizes from the National Head Start Impact Study and Estimated Effects of Treatment on the Treated (TOT) (*source*: Table 1 from Ludwig and Phillips, 2007a)

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)

Notes: 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.

Table 3. Regression Results from Leak et al. (2010) Meta-analysis (*source*: Table 4 of Leak et al., 2010)

	Mean Effect Size	All-Weighted	All-Unweighted	Starting Age 3+ years-Weighted	Study Fixed Effects-Weighted
Starting age of treatment (in yrs)					
less than 3 years old	.39	.09	.06		--
		(.15)	(.16)		--
between 3 and 4 years old	.20	.01	-.13*	.01	--
		(.04)	(.06)	(.04)	--
older than 4 years old	.28	ref	ref	ref	--
Length of treatment (in yrs)					
shorter than half a year	.30	-.11	.05	-.09	--
		(.07)	(.06)	(.07)	--
between half a year and 1 year	.21	ref	ref	ref	--
between 1 and 2 years	.28	.02	.11	.06	--
		(.10)	(.07)	(.11)	--
longer than 2 years	.42	-.16	.22	-.15	--
		(.12)	(.19)	(.13)	--
Time of measures (in yrs)					
during treatment	.10	-.10	-.06	-.12†	.30***
		(.07)	(.06)	(.07)	(.04)
end of treatment	.28	ref	ref	ref	ref
0 to 1 year beyond treatment	.13	.00	-.06	-.01	-.31***
		(.07)	(.06)	(.08)	(.04)
1 to 2 years beyond treatment	.04	-.18	-.27*	-.18	-.45***
		(.13)	(.12)	(.13)	(.07)
2 to 4 years beyond treatment	-.02	-.11	-.21*	-.10	-.36***
		(.07)	(.08)	(.07)	(.11)
more than 4 years beyond treatment	.01	-.20†	-.18**	-.20†	-.54***
		(.11)	(.06)	(.11)	(.08)
Passive control group	.28	.11*	.14*	.11	--
		(.05)	(.07)	(.06)	--
Study did not use random assignment	.28	-.06	-.02	-.07†	--
		(.04)	(.06)	(.04)	--
Any significant differences at baseline	.37	-.17*	-.19**	-.17*	--
		(.07)	(.06)	(.07)	--
Bias was observed in study	.03	.03	.12*	.02	--
		(.07)	(.06)	(.07)	--

Measurement method					
Observational rating	.27	-.07	-.17	-.06	--
		(.07)	(.07)	(.07)	--
Performance test	.25	ref	ref	ref	--
					--
Other measurement method	.27	-.02	-.08	-.07	--
		(.10)	(.07)	(.10)	--
Data collector not blinded	.27	.25**	.23*	.26**	--
		(.08)	(.10)	(.08)	--
Study not from a peer refereed journal	.27	-.06	.00	-.05	--
		(.06)	(.06)	(.06)	--
Treatment on the treated	.27	.02	-.04	.03	--
		(.06)	(.08)	(.07)	--
Baseline covariates not included	.27	.22***	-.07	.22***	--
		(.05)	(.06)	(.05)	--
Attrition					
High attrition (>.25)	.18	.17**	-.17**	-.18**	--
		(.06)	(.06)	(.07)	--
Medium attrition (.16-.25)	.25	-.17**	-.04	-.18**	--
		(.06)	(.06)	(.06)	--
Low attrition (<.16)	.32	ref	ref	ref	--
Attrition information missing	.10	-.07	-.10	-.07	--
		(.09)	(.09)	(.09)	--
Low reliability (<.93)	.27	-.07	-.01	-.06	--
		(.06)	(.07)	(.07)	--
Study conducted before 1980	.29	-.06	.04	-.06	--
		(.07)	(.07)	(.08)	--
Achievement measure	.25	.09†	.01	.09†	-.05
		(.05)	(.04)	(.05)	(.03)
Constant	--	.14	.24	.11	.45***
		(.11)	(.15)	(.11)	(.05)
Number of effect sizes	--	1978	1978	1705	1978
R-square	--	.210	.141	.218	.233

Huber-White standard errors in parenthesis. †p<.10, * p<.05, ** p<.01, *** p<.001. "ref" denotes the reference category for each set of dummy variables

Figure 1.

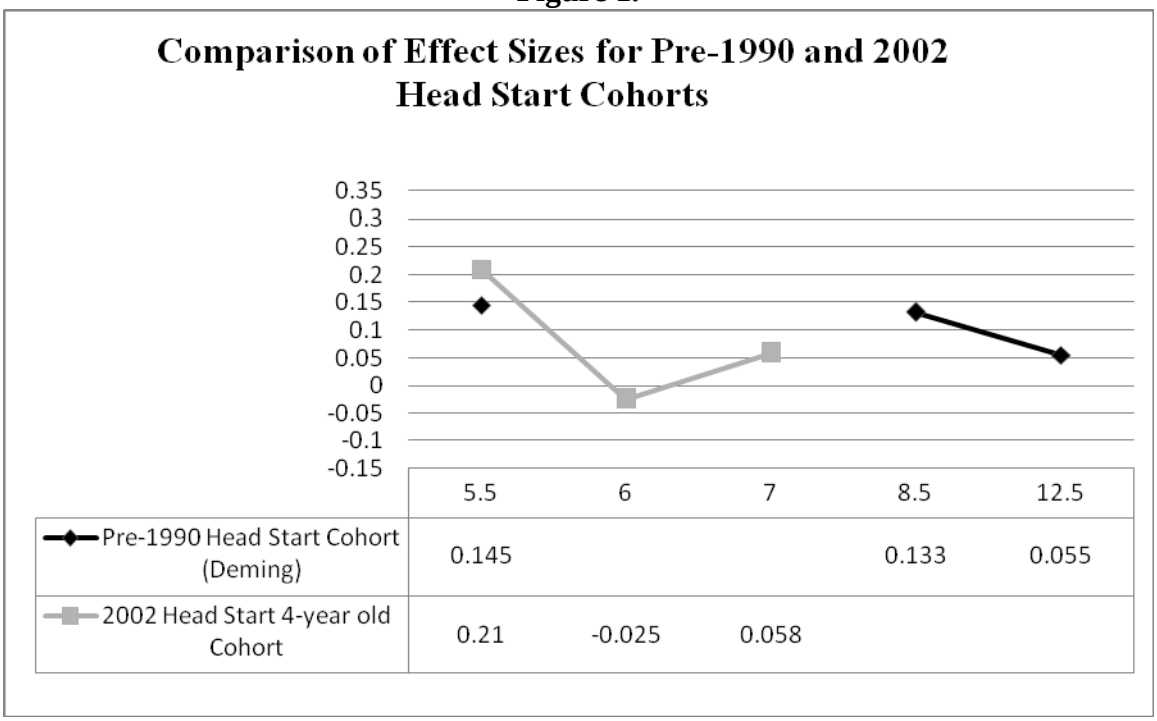
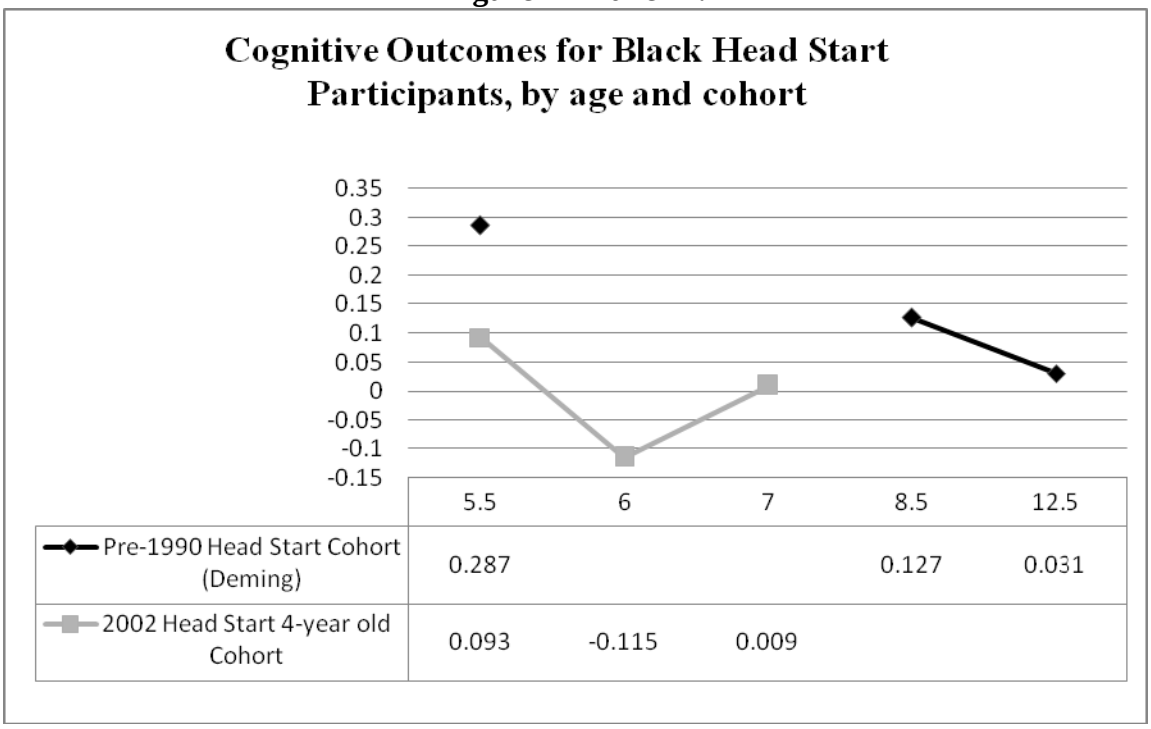
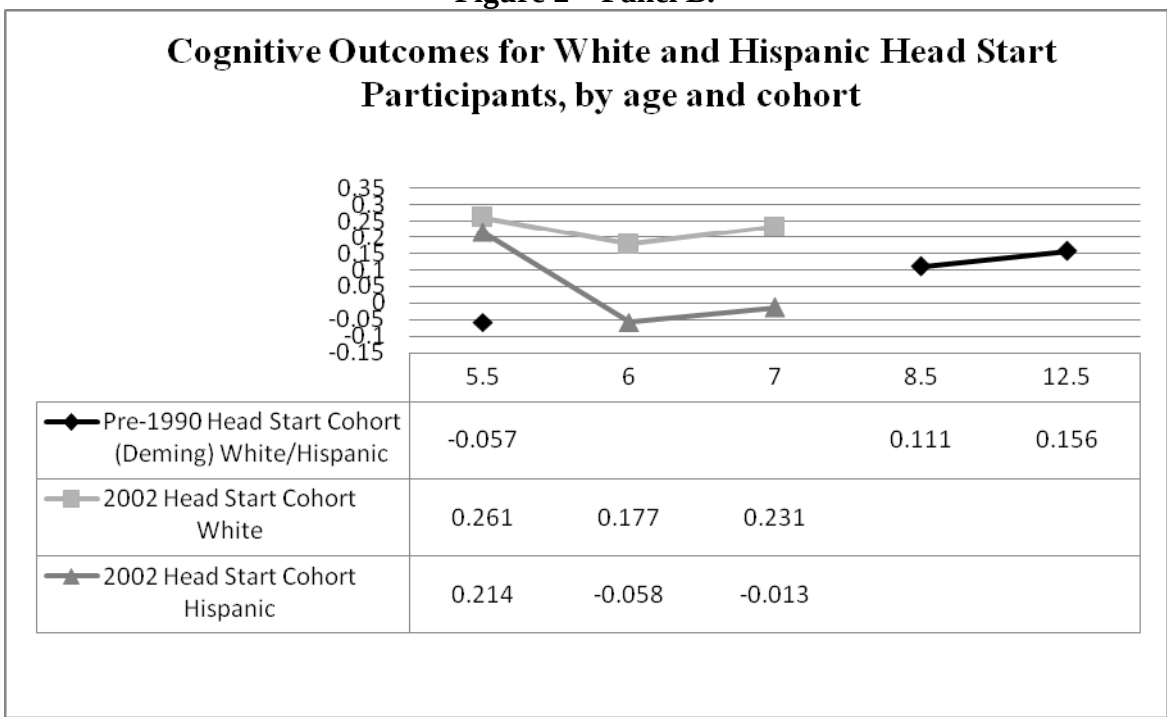

Comparison of Effect Sizes for Pre-1990 and 2002 Head Start Cohorts
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
5.5
6
7
8.5
12.5
Pre-1990 Head Start Cohort (Deming)
0.145
0.133
0.055
2002 Head Start 4-year old Cohort
0.21
-0.025
0.058

Figure 2 – Panel A.



	5.5	6	7	8.5	12.5
Pre-1990 Head Start Cohort (Deming)	0.287			0.127	0.031
2002 Head Start 4-year old Cohort	0.093	-0.115	0.009		

Figure 2 – Panel B.



	5.5	6	7	8.5	12.5
Pre-1990 Head Start Cohort (Deming) White/Hispanic	-0.057			0.111	0.156
2002 Head Start Cohort White	0.261	0.177	0.231		
2002 Head Start Cohort Hispanic	0.214	-0.058	-0.013		

Figure 3: Results from Chetty et al. (forthcoming) on effects of kindergarten classroom quality on short-term test scores and long-term earnings

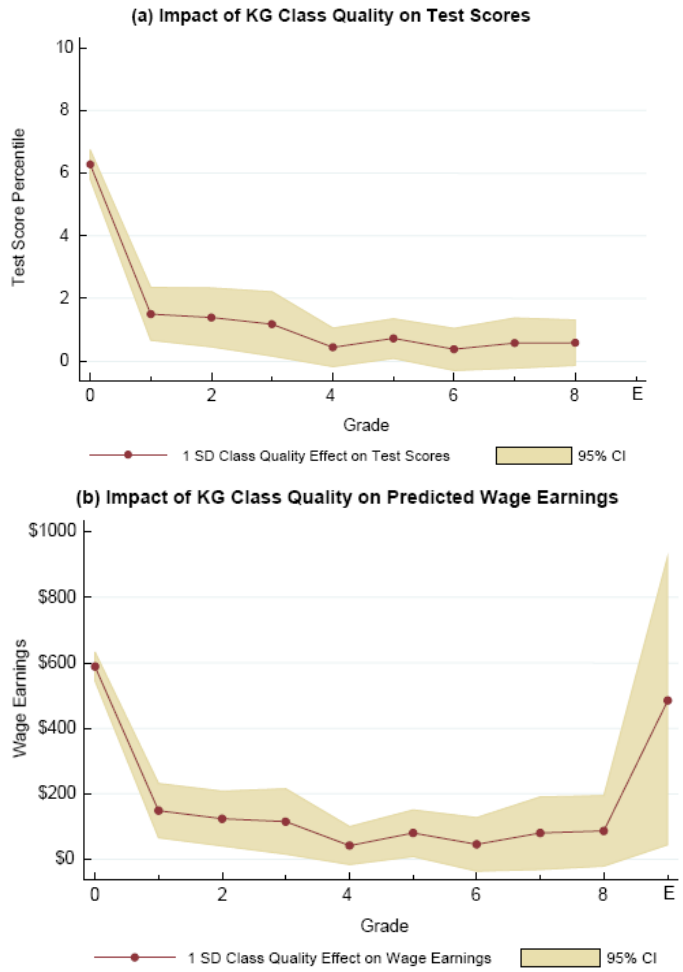


Figure 4: Trends in reading scores for 9 year olds, 1971-2008, from National Assessment of Educational Progress (NAEP)

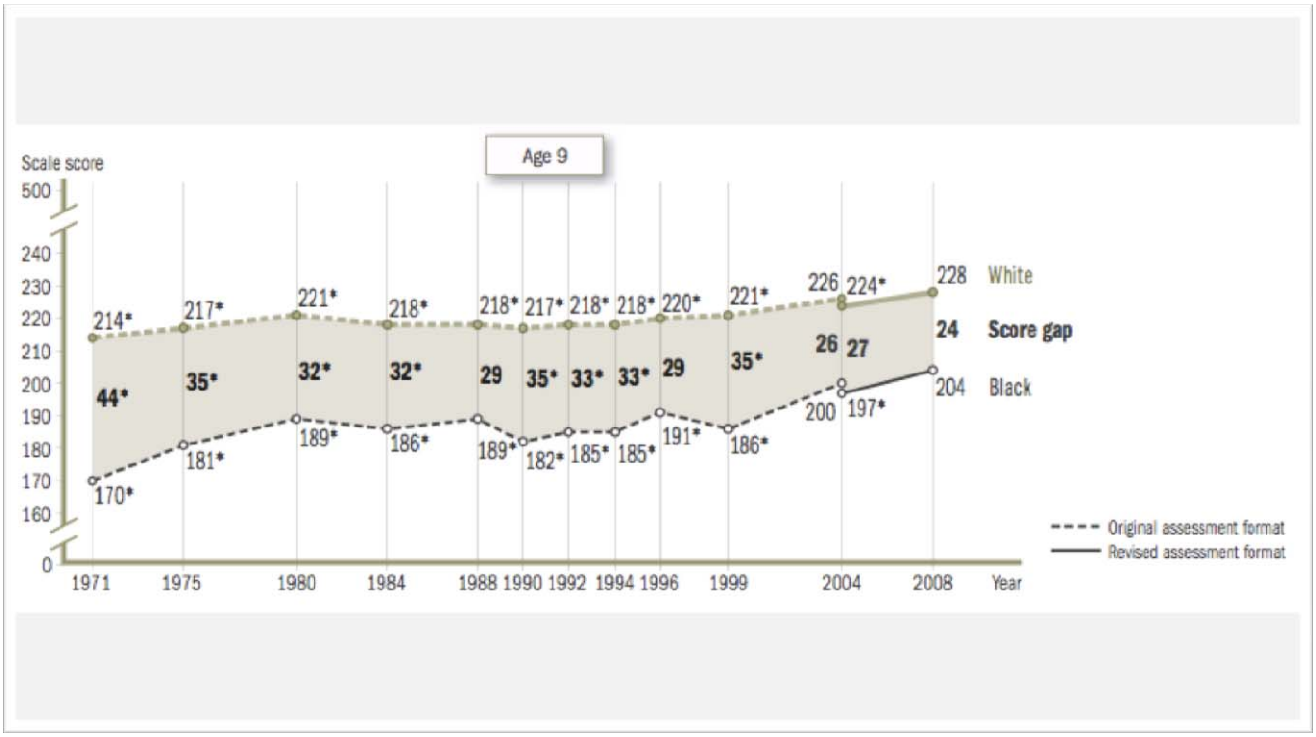
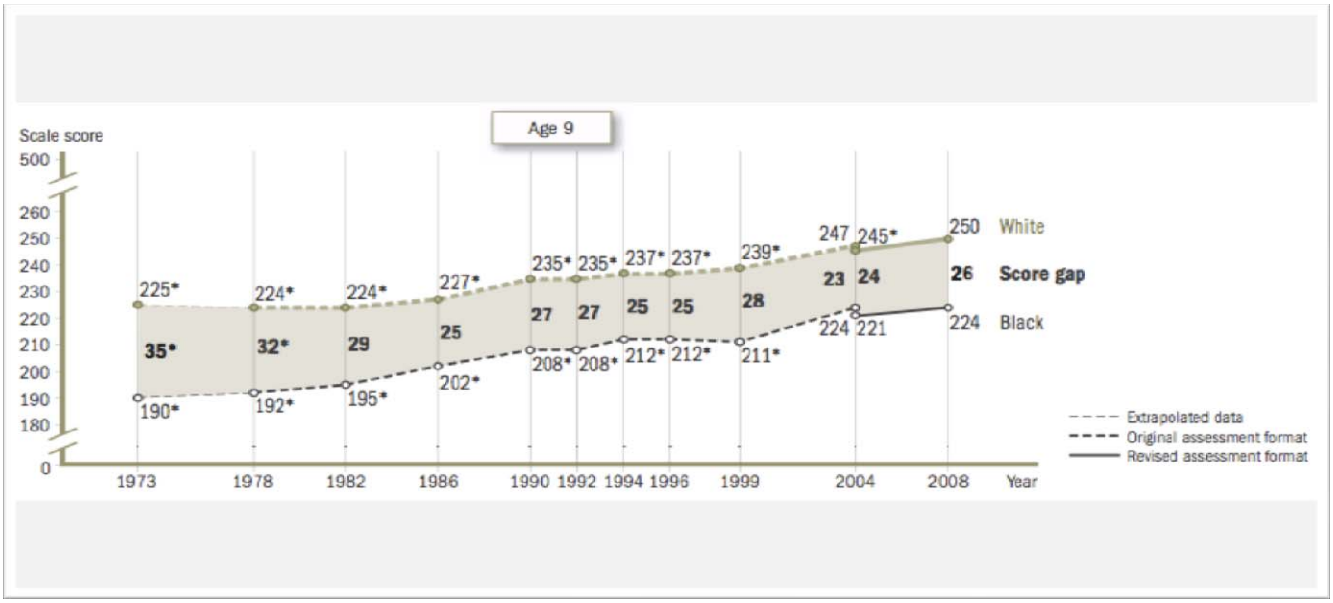


Figure 5: Trends in math scores for 9 year olds, 1971-2008, from National Assessment of Educational Progress (NAEP)



References

Barnett, W. Steven (2011) "Effectiveness of early educational intervention."*Science*. 333(August 19): 975-978.

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: StateUniversity 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.http://www.aei.org/issue/23373

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.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan (forthcoming) "How does your kindergarten classroom affect your earnings? Evidence from Project STAR."*Quarterly Journal of Economics*.

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.

Cunha, Flavio and James Heckman (2007) "The technology of skill formation." Cambridge, MA: NBER Working Paper 12840.

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 Duncan Thomas (2000) "School quality and the longer-term effects of Head Start." *Journal of Human Resources*. 35(4): 755-774.

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*.

Deming, David (2009) "Early childhood intervention and life-cycle skill development: Evidence from Head Start." *American Economic Journal: Applied Economics*. 1(3): 111-134.

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 BotondKoszegi (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 SendhilMullainathan (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*.

Haskins, Ron (2010) "Finally, the Obama Administration is putting Head Start to the test." *Washington Post*.

Haskins, Ron and W. Steven Barnett (2010) "New directions for America's early childhood policies."In*Investing in Young Children: New Directions in Federal Preschool and Early Childhood Policy*, Edited by Ron Haskins and W. Steven Barnett. Washington, DC: Brookings Institution. pp. 1-28.

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.

Klein, Joe (2011) "Time to ax public programs that don't yield results."*Time Magazine*. July 7, 2011.

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*.

Leak, Jimmy, Greg J. Duncan, Weilin Li, Katherine Magnuson, Holly Schindler, and Hirokazu Yoshikawa (2010) "Is timing everything? How early childhood education program impacts vary by starting age, program duration, and time since the end of the program." UC-Irvine working paper, presented at the fall 2010 meetings of the Association for Public Policy Analysis and Management, Boston, MA.

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.

Ludwig, Jens and Deborah A. Phillips (2007a) "The benefits and costs of Head Start." Cambridge, MA: NBER Working Paper 12973.

Ludwig, Jens and Deborah A. Phillips (2007b) "The benefits and costs of Head Start."*Social Policy Report*, Volume XXI, Number 3. Society for Research on Child Development.

Ludwig, Jens and Deborah A. Phillips (2010) "Leave no (young) child behind: Prioritizing access in early education." In *Investing in Young Children: New Directions in Federal Preschool and Early Childhood Policy*, Edited by Ron Haskins and W. Steven Barnett. Washington, DC: Brookings Institution. pp. 49-58.

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.

Murnane, Richard J., John B. Willett, K.L. Bub, and K. McCartney (2006) "Understanding trends in the black-white mathematics achievement gap during the first years of school."*Brookings-Wharton Papers on Urban Affairs*.97-135.

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.
http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, et al. (2010) *Head Start Impact Study: Final Report*. Westat. Report Prepared for the U.S. Department of Health and Human Services.http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/hs_impact_study_final.pdf

Ramey, Craig T. and Sharon Landesman Ramey (2010) "Head Start: Strategies to improve outcomes for children living in poverty." In *Investing in Young Children: New Directions in Federal Preschool and Early Childhood Policy*, Edited by Ron Haskins and W. Steven Barnett. Washington, DC: Brookings Institution. pp. 59-68.

Reynolds, Arthur J., Judy A. Temple, Suh-Ruu Oh, Irma A. Arteaga, and Barry A.B. White (2011) "School-based early childhood education and age-28 well-being: Effects by timing, dosage, and subgroups." *Science*. 333: 360-364.

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 (2006) "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 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.

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.

Wong, Vivian C., Thomas D. Cook, W. Steven Barnett, and Kwanghee Jung (2008) "An effectiveness-based evaluation of five state pre-kindergarten programs."*Journal of Policy Analysis and Management*. 27(1): 122-154.

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.