The Relative Performance of Head Start

Cory Koedel
University of Missouri

Teerachat Techapaisarnjaroenkit*
University of Missouri

June 2011

In early 2010, the U.S. Department of Health and Human Services released the findings from a large, experimental evaluation of the Head Start program. A common interpretation of the findings is that they show "small" effects, which has lead to, among other things, calls to improve the efficacy of Head Start. However, it is not clear that Head Start is performing worse than should be reasonably expected. To provide a frame of reference for evaluating the program, we compare the performance of Head Start childcare centers to the performance of non-Head Start childcare centers, the latter being the preferred childcare option of wealthier families. We find that, on average, Head Start centers perform similarly to non-Head Start centers. Our results suggest that expectations for the Head Start program may be too high.

* We thank Emek Basker, Daphna Bassok and Jeff Milyo for useful comments and suggestions, and the Economic & Policy Analysis Research Center (EPARC) at the University of Missouri for research support.

Head Start is a federal matching grant program that provides disadvantaged children with access to pre-kindergarten education. The program began as part of the War on Poverty in the 1960s, and today serves roughly 900,000 children each year at a cost of $6.9 billion.

There is a large and robust literature that evaluates the effects of Head Start on program participants. A recent and particularly influential study is the *Head Start Impact Study* (U.S. Department of Health and Human Services, 2010), which uses an experimental design and a nationally representative sample of children to evaluate the program. The *Impact Study* shows that Head Start has positive immediate-term effects, but these effects have been described as "disappointingly small" (Besharov, 2005).[1] It also shows that the effects of Head Start fade out quickly.[2] The *Impact Study* has rekindled an intense policy debate over the future of Head Start, and has lead to increased calls for change.

Prior research provides some insight as to why the initially positive effects of Head Start fade out over time – Head Start children go on to attend inferior K-12 schools (Currie and Thomas, 2000; Lee and Loeb, 1995; Lee et al., 1989).[3] However, this does not address the commonly voiced concern that even the immediate-term effects of the program are too small. It is often presumed, sometimes implicitly, that Head Start's small program impacts are attributable to the poor performance of the program itself. Taking this presumption as a point of departure, one logical course of action would be to intervene to improve the program. However, it is also possible that Head Start is performing well, but the program's margin for effect is smaller than is typically thought. These two explanations are substantively different, and they have different

[1] These findings are consistent with a large non-experimental literature showing that Head Start improves children's cognitive, social and health outcomes during and immediately after treatment (e.g., see Currie and Thomas, 1995; Currie, 2001; Frisvold and Lumeng, 2009).

[2] Prior non-experimental work has also documented the fade out of program impacts (e.g., see Currie and Thomas, 1995).

[3] The literature on K-12 education shows that it is not uncommon for educational interventions to have effects that fade out over time (see, for example, Bhatt and Koedel, 2010; Jacob et al., 2008).

policy implications. For example, if Head Start's small program impacts are an artifact of the nature of the intervention itself, it makes it less likely that adjustments to the program will meaningfully improve children's outcomes.

It is difficult to know what types of program impacts are reasonable to expect from Head Start because we have limited evidence on the effects of preschool interventions in general, particularly large-scale interventions.[4] In our study, we use the performance of non-Head Start childcare centers as a benchmark for evaluating the Head Start program. Although there is no other program exactly like Head Start that can be used for comparison, non-Head Start childcare centers offer a similar service, and are a viable, if imperfect, option.

A direct comparison between Head Start and non-Head Start childcare centers is difficult because the programs serve different populations of children. The key to our empirical strategy is that we have data from children who enrolled in Head Start and non-Head Start centers but were treated for only a short time. We use these children to control for selection into each childcare type, and compare their outcomes to outcomes from children who were treated for a much longer duration. This approach produces estimates of program impacts for Head Start and non-Head Start centers that correspond to differences in treatment duration within each program. We compare these estimates in a difference-in-difference framework.

Motivated largely by the *Impact Study* findings and subsequent policy discussions, we focus our evaluation on how Head Start and non-Head Start centers affect children's immediate-term test scores. We find that, on average, both childcare types perform similarly. Our findings are at odds with the common perception that Head Start is performing poorly, but this perception

[4] As noted by Barnett (1992), many of the earlier studies that evaluate preschool interventions, including Head Start, suffer from serious design limitations. In a more-recent and rigorous study, Currie and Thomas (1995) consider the effects of non-Head Start preschool interventions. They use a sibling-fixed-effects approach to estimate non-Head Start preschool effects on cognitive outcomes and estimate effects that are statistically indistinguishable from zero.

appears to be driven more by lofty expectations for the program than by the program's actual performance. Implicit in calls for improvements to the delivery of Head Start is the notion that the program is somehow "broken" and must be fixed, but this notion is not supported by the evidence presented here. We additionally note that there is a growing body of research showing that Head Start confers positive and economically meaningful long-term benefits to program participants (e.g., see Deming, 2009; Garces et al., 2002; Ludwig and Miller, 2007). Combined with this prior research, our study suggests that the negative reaction to the *Impact Study* findings may be overblown.[5]

I. Motivation and Research Design

The Head Start literature to date has focused primarily on estimating treatment impacts for treated children. The typical comparison in the literature is between Head Start and a composite of the childcare arrangements available to children who are similar to Head Start attendees, but who do not attend (including non-Head Start, non-relative care; relative care; and parent care). While this type of comparison gets at exactly the right question when we are interested in knowing whether Head Start participants are better off for having participated (a clear first-order issue), it is only partly informative about Head Start's performance in the broader sense. A key issue is that the non-Head-Start childcare arrangements available to socioeconomically disadvantaged families are likely to be of lower quality than the childcare arrangements available to other families. This means that from the current literature we only know how well Head Start performs relative to what are likely to be inferior childcare alternatives.

[5] Examples of negative responses to the *Impact Study* findings are abundant, and include Besharov (2005), Coulson (2010), Wetzstein (2010) and Whitehurst (2010). There are also other potential benefits of the provision of Head Start– for example, Lemke, Witt and Witte (2007) find that the availability of Head Start is associated with an increase in single mothers on welfare transitioning directly into work.

Even so, the *Impact Study*, like many other studies in the literature, finds that Head Start has small treatment impacts. Prompted by the *Impact Study* results, Whitehurst (2010) writes: "Head Start isn't doing the job the families it serves and the nation need. It must be improved. There are many proposals for doing so. Let's try them, test them, and do what works." Implicit in Whitehurst's argument, and the many others like it, is that Head Start *can* perform better if only we can figure out how to improve efficacy. But this is far from certain. In fact, we only know that Head Start seems to marginally improve upon the available alternatives for program participants. An important question remains: How much better could we reasonably expect Head Start to perform? There are very few studies in the literature that attempt to answer this question.

A notable exception is Gormley et al. (2010), who compare Head Start to an early-childhood program in Tulsa, Oklahoma. They find that the Tulsa program generally outperforms Head Start (with the exception that Head Start has larger effects on health outcomes). On the one hand, Gormley et al. (2010) establish that early-childhood interventions can have large effects on children's outcomes, which is important. However, on the other hand, comparability is an issue because the childcare inputs in the Tulsa program appear to be superior to those in Head Start. For example, Gormley and Gayer (2005) report that care providers in the Tulsa program are required to have four-year college degrees and are paid like public elementary-school teachers.[6] This raises concerns about the fiscal scalability of such a program.[7]

Instead of comparing Head Start to the composite of alternatives that are available to socioeconomically disadvantaged children, or to a program that will be difficult to bring to scale for fiscal or other reasons, the contribution of the present study is to compare Head Start to the

[6] Head Start has been slowly ramping up teacher degree requirements, but even by 2013 Head Start will only require half of its teachers to hold a bachelor's degree. Furthermore, teacher pay in Head Start is far below the level of the typical elementary-school teacher, particularly when benefits are included in the calculation.

[7] Also see Currie (2001), who documents evidence from several other studies of pre-kindergarten interventions that are funded at a considerably higher level than Head Start.

larger non-Head Start, center-based childcare sector. This comparison is appealing because over half of all children in the United States attend some form of non-Head Start, center-based care, making it the most common type of care in the country.[8] Furthermore, childcare costs in the non-Head Start, center-based childcare sector are more-closely aligned with those of Head Start than are the costs of previously studied, higher-quality childcare programs.[9]

One way to view the contribution of our study is as a complement to earlier work, like that of Gormley et al. (2010) and others, which shows that high-quality pre-kindergarten interventions can be quite successful. The prior literature provides insight into what we could expect to see if we were to dramatically scale up the quality of the Head Start program, and our fiscal commitment to its success. At the other end of the spectrum is the question of how well Head Start performs compared to more-typical childcare arrangements in the United States – prior to our study, we are not aware of any work that attempts to shed light on this question.

Our approach is conceptually straightforward: we estimate average program impacts on children's test scores for Head-Start and non-Head Start childcare centers, and compare these estimates using a difference-in-difference framework.[10] We estimate the effects of the programs using test score data collected at the time of treatment. Although much of the recent controversy surrounding Head Start is related to the lack of persistent program impacts, we cannot perform our comparison using later-year outcomes because these outcomes are also affected by the

[8] In the ECLS-B dataset, which includes a nationally representative sample of children (see Section II), over half of the sample attended some form of center-based care (non-Head Start) at the time of the age-4 survey (≈55 percent). The next-most common childcare arrangement reported in the data was non-parent, relative care (≈20 percent).

[9] If anything, the ECLS-B data suggest that per-pupil expenditures in Head Start are higher, on average, than in the larger non-Head Start sector, although we note that accurate data on non-Head Start per-pupil expenditures are difficult to obtain. That said, the per-pupil costs in the non-Head Start sector appear to be much closer to Head Start's per-pupil costs than are the costs of other pre-kindergarten programs that have been evaluated in prior work (see Currie, 2001, for example). Also, Resnick and Zill (undated) show that in terms of measurable inputs, Head Start looks similar to preschools from the National Child Care Staffing Study.

[10] Head Start also has non-cognitive objectives, like improving children's social and health outcomes, but we focus on cognitive outcomes here.

different post-treatment experiences of Head Start and non-Head Start children (Currie and Thomas, 1995, 2000; Lee and Loeb, 1995; Lee et al., 1989). For our analysis, where interest is in the performance of Head Start itself, these post-treatment experiences are confounding factors.

The key empirical issue that we face is that children and their families non-randomly select into childcare treatments. To minimize the influence of selection bias in our study, we rely on a fairly unique feature of our dataset, the Early Childhood Longitudinal Survey, Birth Cohort (ECLS-B). In the survey, each caregiver for an age-4 child is asked to indicate how long the child has been receiving care. We use caregiver responses to this question to categorize children as either "treatments" or "controls", within care type, based on treatment duration. Specifically, some children had been enrolled in care for only a few months at the time of the survey, while others had been enrolled for much longer. The children who had been enrolled for only a short time received very little treatment, but they are useful for our analysis because they *selected into treatment*. We use these "control" children to capture selection effects, and compare their outcomes to outcomes from children who were in care for a longer duration, who we define as "treated". Although Head Start children differ markedly from non-Head Start children across virtually every observable measure, the within-program samples of treatments and controls are observationally very similar.[11]

Our research design greatly improves observational comparability in our study, which we show below, but we rely on variation across children in the timing-of-entry into childcare to identify program impacts.[12] For our within-program estimates of treatment effects to be

[11] To the best of our knowledge, this approach has been used twice in prior research, first by Hebbeler (1985) and more recently by Behrman et al. (2004). We note that our use of the term "control" is non-standard because the control children in our analysis did receive some treatment. Nonetheless, for ease of presentation we refer to the children who were in care for only a short duration as "controls" throughout our analysis.

[12] Variation in time-in-care, which determines treatment or control status in our analysis, can come from two sources in the data: (1) timing-of-entry into childcare, and (2) variation in assessment dates among children. Following Fitzpatrick et al. (2011), we attempted to isolate and exploit the latter source of variation in our study because it is

unbiased, timing-of-entry must be independent of potential outcomes conditional on observable information in each program. That is, for each program:

$$Y_0, Y_1 \perp T \mid X \qquad (1)$$

In (1), Y_0 and Y_1 are outcomes that correspond to the control and treatment states; T indicates treatment status, which is determined by duration in care; and X is a vector of observable information about children and their families.

The conditional independence assumption will be violated if timing-of-entry is correlated with unobserved factors that are not controlled for by the components of X, and if these factors also influence children's outcomes. Unfortunately we cannot directly test for the biasing effects of unobservables, but we do use the rich set of observable characteristics available in the ECLS-B data to show that within the same program, children who differ in terms of treatment duration over the range of durations that we consider are observationally very similar (see below). Additionally, we provide some validation of our empirical approach by comparing the treatment effects that we estimate for Head Start to analogous experimental estimates from the *Impact Study*. Our findings from this exercise are reported in Appendix B, which shows that our Head Start specific estimates are similar to the experimental results, particularly in math. Our interpretation of this result is that even if conditional independence is not mechanically satisfied in our analysis, any bias generated by its failure is likely to be small.[13]

---

more likely to be exogenous. We were unsuccessful because we could not disentangle the test-date variation from variation in age (unlike in Fitzpatrick et al., we do not have access to a pre-test). Nonetheless, we note that our estimates from the IV analysis are consistent with the findings we present throughout the text – Head Start and non-Head Start centers have similar effects – if we assume that the age bias is constant across Head Start and non-Head Start children. More details are available in Section V.

[13] The difference in treatment duration between treatment and control children in our primary analysis is roughly six months in each program, which is similar to the treatment duration for *Impact Study* children, which facilitates the comparison (see Appendix B).

We also use a difference-in-difference framework to further limit the impact of bias in our estimates. The difference-in-difference models estimate Head Start's relative effect as the difference between the within-program estimates for Head Start and non-Head-Start centers. Taking this difference will remove any bias generated by endogenous timing-of-entry decisions that is consistent across both programs. So, for example, if early entrants into both types of care are more disadvantaged, any resulting bias in the program-specific estimates will be reduced in our difference-in-difference models.[14]

II. Data

Our data are from the Early Childhood Longitudinal Survey, Birth Cohort (ECLS-B), provided by the National Center for Education Statistics (NCES). The ECLS-B began tracking children at birth and administered follow-up surveys when the children were aged 9-months, 2-years and 4-years (the survey is ongoing). During each wave of the survey information is collected from parents and children, and for the age-2 and age-4 follow-ups, childcare providers were also surveyed. We use the data file from the age-4 follow-up in our analysis.

The ECLS-B contains detailed information about children and their families. Two of the most important variables in the data are family income and parental-education levels, which are highly correlated with children's outcomes. Also, because the survey began at birth, it provides access to information that is not commonly available, like birth weight. Childcare arrangements, past and present, are reported for each child. The data on childcare arrangements are collected contemporaneously. Therefore, unlike much of the prior work in this area, we do not need to rely on retrospective questionnaires given to parents to assign childcare treatments.

[14] The difference-in-difference estimates could *increase* bias if unobserved, timing-of-entry selection effects move in opposite directions for Head Start and non-Head Start children. Although this possibility is difficult to evaluate directly, we note that the small observable differences between treatment and control children within programs are consistent with any such bias moving in the same direction (see Table 2).

Assessment data are collected from children during each wave of the survey. During the 9-month and 2-year follow-ups, children were given motor-skills and mental-skills tests, and during the age-4 follow-up they were given cognitive tests in math and literacy. We estimate childcare effects on the age-4 math and literacy scores, and condition on the history of children's scores in our models. We standardize all test scores using the universe of test takers in the data.

Our empirical strategy requires that we impose several restrictions on the dataset. First, to facilitate a clean comparison across programs, we restrict our analysis to include only children who were reported to be in care, exclusively, either at a Head Start facility or a non-Head Start, center-based care facility at the time of the age-4 survey. That is, we omit all children who received some other form of care, or split time between multiple care arrangements.[15] Among the children who meet this criterion, we further restrict the sample based on caregiver responses to the time-in-care question. In our primary models we identify children who were reported to be in care for three months or less as controls, and children who were reported to be in care for 6-12 months as treatments. In a robustness exercise in Section V, we show that our findings are not sensitive to reasonable adjustments to these treatment and control definitions.[16]

We impose three additional restrictions on the data. The first two restrictions are imposed because the caregiver question about time-in-care is specific to the current center, raising concerns about misclassification error. For example, we may misclassify some treated children as controls, and understate the amount of treatment that some treated children actually received, simply because they changed childcare centers. To minimize these occurrences in the data, we

[15] The majority of children in each program receive exclusive care. The exclusive enrollment rate is higher for the non-Head Start sample, with the difference driven primarily by the fact that more Head Start children also receive care from a non-parent relative (see Table 1).

[16] We cap treatment durations in the neighborhood of one year because children do not generally enroll in Head Start prior to age-3. Therefore, at the time of the age-4 survey, few Head Start children were reported to have been in care for over a year. As we extend the time-in-care window for treatments beyond 12 months, we observe longer care durations, on average, for non-Head Start children relative to Head Start children, reducing comparability.

omit all children whose parents reported changing residences between the age-2 and age-4 surveys because we expect that residence changes are likely to correspond to changes in childcare arrangements. We also exclude all children who were reported to attend any non-relative childcare during the age-2 survey. Neither treatments nor controls should have been in care at that time, and if they were, it suggests that a change in childcare arrangements occurred. We also omit children who were reported to be in care for more than 40 hours per week because Head Start centers do not care for children for over 40 hours.[17] Again, we consider the robustness of our findings to relaxing these restrictions in Section V.

Moving forward, when we reference our preferred data sample, we are referring to the sample where treatments were in care for 6-12 months, controls were in care for three months or less, and the above-described data restrictions are in place.

Table 1 shows descriptive statistics for the full ECLS-B data file (age-4 survey), and within childcare type, for program participants, exclusive program participants and our preferred data samples. We highlight two aspects of the table. First, across programs, Head Start and non-Head Start children differ substantially along virtually every observable dimension. Second, within programs, our preferred data samples differ from the exclusive-care samples in some ways, and are much smaller.[18] However, given the sharp reduction in sample sizes when we move to our preferred samples, and the number of restrictions we impose on the data, the differences seem generally modest. That is, with a handful of exceptions, our preferred data

[17] Although we note that in the data there are a handful of exclusive-care Head Start children who are reported to be in care for over 40 hours per week. See Appendix Table A.1.
[18] Appendix Table A.1 shows how we arrive at our final data sample in each program.

samples do not seem to be meaningfully different from their respective exclusive-care samples in the full ECLS-B data file.[19]

Next, in Table 2, we split our preferred samples based on treatment status and compare treatments and controls within and across programs. We also further restrict the data to include only children for whom either a math or literacy score is reported during the age-4 survey (or both), which is required for inclusion in our analysis. The first row of the table shows the average difference in treatment duration between treatments and controls. The difference is roughly six months for both programs (although it is slightly larger in the non-Head Start sample – by about five days, or 0.16 months). The treatment effects that we report below correspond to these differences in treatment duration, which we consider to be equivalent across programs.

The second row highlights a large and statistically significant difference in age (in months) between treatments and controls, which is a mechanical artifact of our research design – treatment children, by definition, had been in care longer at the time of the age-4 ECLS-B survey, and this is correlated with age (note that there is no discrepancy in age across programs conditional on treatment status). To ensure that the age differences between treatments and controls do not bias our findings, we include age-in-months indicator variables in all of our models – that is, we estimate within-age treatment effects. We allow the age trend to differ by program type in our difference-in-difference models.[20]

[19] Because our samples are small and purposefully selected we do not use the NCES-provided sample weights; instead, we control for the oversampled populations directly in the covariate-adjusted models as in Rose and Betts (2004). The survey weights are designed to allow for generalizations about the survey population – because we have already drawn a non-random subsample of the data, it is not clear what weighting would imply for our estimates. In any case, for completeness, in an omitted analysis we also estimate models that use the survey weights. Our primary findings are not qualitatively sensitive to using the weights, but the weighted estimates are less precise.

[20] Most of the sample is between 48 and 59 months of age during the age-4 survey. However, some children are younger, and some older. Children are grouped by exact age, in months, for ages 48-59 months. The remaining children are categorized as either age < 48 or age > 59. Our results are robust to excluding children outside of the primary age range of 48 to 59 months. If anything, excluding these children modestly favors Head Start in our comparison.

The remainder of Table 2 reports average values for the demographic, socioeconomic, and lagged test-score variables. Within both programs there are some differences between treatments and controls. The differences are more-often statistically significant in the non-Head Start sample; however, as we show in Table 3, the nominal differences are slightly larger in the Head Start sample (the differences in the t-tests are driven by differences in sample size). We also note that the within-program differences between treatments and controls are similar across childcare types. For example, treatments from both the Head Start and non-Head Start samples are more likely to be non-white, and generally come from lower-income and less-educated families than controls. This suggests that any bias in our childcare-specific estimates owing to these differences will be reduced when we compare the programs.

Are the differences that we document in Table 2 large? We answer this question by calculating the standardized differences in observables between three groups: (1) Head Start treatments and non-Head Start treatments, (2) Head Start treatments and Head Start controls, and (3) non-Head Start treatments and non-Head Start controls. We calculate the standardized difference in observable characteristic X_k between groups *A* and *B* as:

$$STDIF(X_k) = \left| \frac{\bar{X}_k^A - \bar{X}_k^B}{\sqrt{\frac{Var(X_k^A) + Var(X_k^B)}{2}}} \right| * 100 \qquad (2)$$

Equation (2) is motivated by Rosenbaum and Rubin (1985), who suggest using a similar metric to evaluate whether matching methods are effective in producing observationally comparable treatment and control units. Analogously to the matching literature, we expect our treatment and control observations to be similar, in which case the within-program standardized differences should be small.

Table 3 reports standardized differences that are averaged across all of the demographic, socioeconomic and lagged test-score variables in Table 2, by comparison group.[21] Not surprisingly, the average standardized differences are much smaller in the within-program comparisons than in the across-program comparison. Because the treatment and control definitions are chosen independently of the observable information in the data, the improvement in observed comparability generated by our research design suggests that unobserved comparability is also improved. We cannot directly control for unobserved differences between children in our models, meaning that the reduction in these differences implied by Table 3 is particularly important.

It is easy to see that our within-program samples are more comparable than our between-program samples, but whether the within-program standardized differences are small or large is not obvious. To provide some insight, for each comparison we construct the empirical distribution for the average standardized difference under random assignment. Because the standardized difference is an absolute measure, even with random assignment and large (but finite) samples, the center of the distribution will be above zero.

We illustrate our approach for constructing the empirical distributions using the comparison between Head Start treatments and controls. We begin with the 852 children in the full ECLS-B data file who were reported to exclusively attend Head Start (see Table 1). From this sample we randomly draw 96 observations and calculate the mean and variance of each variable, X_k, and then randomly draw 108 observations and calculate the mean and variance of each variable. The first group corresponds to our primary control sample from Table 2, where N = 96, and the second group to our primary treatment sample, where N = 108. We calculate the

[21] We exclude differences in age in our calculations because these differences are generated by our research design. We address the age discrepancies between treatments and controls in our analysis by estimating within-age treatment effects.

average standardized difference between the two random samples across the *X*'s, and then repeat this procedure 500 times to construct the empirical distribution of the average standardized difference under random sampling. For each comparison, Table 3 reports the average and 95-percent confidence interval of this distribution.[22,23]

Table 3 shows that for both programs, the observed standardized difference is comfortably within the 95-percent confidence interval of the random-sampling distribution (although, consistent with Table 2, the standardized difference is larger relative to the empirical distribution in the non-Head Start sample). Row (4) of the table shows that our "difference-in-difference" standardized difference is also well within the 95-percent confidence interval of the random-sampling distribution. Although we lack a true mechanism for random assignment in our study; at least observationally, our treatment and control samples look very similar to what we would expect to see if such a mechanism were actually in place.

III. Methodology

We estimate program-specific regression models where we only include observations that are identified as either treatments or controls. These models take the following form:

$$Y_i = \beta_0 + A_i\beta_1 + X_i\beta_2 + T_i\pi + \varepsilon_i \quad (3)$$

In (3), Y_i indicates an outcome for child i, A_i is a vector of age-in-months dummies, X_i is a vector of observable information about child i and her family (the child-specific variables in X_i include the demographic, SES and lagged test-score variables shown in Tables 1 and 2), and T_i indicates

[22] To generate the random-sampling distribution for our comparison across programs, we randomly draw from the combined exclusive-care sample in Table 1 (N=4,403). For the non-Head Start comparison, we draw from the sample of 3,551 non-Head Start, exclusive-care children. Finally, for the difference-in-difference, we simply subtract the non-Head Start difference from the Head Start difference at each draw.

[23] The re-sampling is done without replacement. We also estimate the empirical distributions by re-sampling with replacement, which yields very similar results. However, we prefer re-sampling without replacement because when we re-sample with replacement, the same observations can be designated as treatments and controls. Of course, in practice this can never occur.

treatment status, where treated children belong to the group that spent more time in care. $\hat{\pi}$ estimates the treatment effect.

The identifying variation in the models comes from differences in treatment status among children who were the same age at the time of their assessments (because we include age fixed effects). This means that we identify the program impacts by comparing children who differ by their timing-of-entry into childcare, as discussed above.[24] More specifically, we compare relatively late-entering control children to relatively early-entering treatment children. Noting that the ECLS-B assessments were administered mostly in the fall – in September, October and November – the control children in our analysis primarily entered care in the late summer and early fall immediately preceding the tests. Treatment children entered care at some point during the prior year – some enrolled at the beginning of the year and others enrolled midyear.[25]

We compare the program-specific estimates from equation (3) using a difference-in-difference framework:

$$Y_i = \beta_0 + A_i\gamma_1 + X_i\gamma_2 + T_i\gamma_3 + HS_i\,\delta_1 + (A_i * HS_i)\delta_2 + (X_i * HS_i)\delta_3 + HS_i^T\theta + u_i \quad (4)$$

In (4), Y_i, A_i and X_i are defined as above. T_i also maintains its definition from equation (3), but now applies to treated individuals in either program. HS_i is an indicator for Head Start attendance, both treatment and control, and HS_i^T further indicates Head Start treatment. The covariates and age indicators are interacted with the Head Start indicator to allow them to differentially affect Head Start and non-Head Start children. F-tests reject the null hypothesis that

[24] Children who were assessed at the same age cannot have different treatment durations if they also entered childcare at the same age (treatment duration = {assessment age – entry age}). Therefore, our estimates are identified from variation in timing-of-entry into childcare across children.

[25] Midyear enrollment policies at non-Head Start centers are likely to be heterogeneous. The policy for Head Start centers is that midyear enrollments are permitted subject to space constraints. The ECLS-B data show that midyear enrollments occur regularly in both programs, although the most common enrollment period is in the late summer.

$\delta_2 = \delta_3 = 0$.[26] If our empirical strategy effectively mitigates selection bias, θ can be interpreted as the average causal effect of Head Start treatment relative to treatment at a non-Head Start childcare center. δ_1 is also informative – it indicates how selection into Head Start compares to selection into non-Head Start, center-based care. Noting that $HS_i^T = HS_i * T_i$, equation (4) is a basic difference-in-difference model.[27]

IV. Results

We estimate program impacts on children's age-4 math and literacy scores, which we standardize using the entire sample of age-4 test takers in the ECLS-B data. Table 4 reports estimates from the childcare-specific models. The first column in each panel shows raw differences in outcomes between treatments and controls, and the second column is covariate adjusted. For non-Head Start centers, treatment is positively associated with outcomes. For Head Start centers, the estimates are nominally positive but statistically indistinguishable from zero.

Table 5 reports several different sets of estimates. The first two columns show estimates from OLS regressions that directly compare Head Start and non-Head Start treatments. Column (1) reports unconditional differences in outcomes, and column (2) reports covariate-adjusted differences. Columns (3) and (4) report estimates from difference-in-difference models where we impose the constraint that $\delta_2 = \delta_3 = 0$. Columns (5) and (6) report estimates from difference-in-difference models where we relax this constraint, which is our preferred specification (these estimates are equivalent to subtracting the non-Head Start effects from the Head Start effects in

[26] When the interaction terms involving the covariates are included, the difference-in-difference models return estimates that are equivalent to the differences between the program-specific estimates from model (3). Imposing the constraint $\delta_2 = \delta_3 = 0$ changes the estimates slightly, and in a way that tends to favor Head Start, although we empirically reject this model (see Section IV).

[27] Bertrand et al. (2004) show that difference-in-difference estimators can overstate statistical significance when outcomes are evaluated for several years before and after an intervention, and there is no correction for serial correlation. However, this concern is not relevant to our analysis.

Table 4). The covariates greatly increase predictive power in all of the models, as would be expected. However, in the difference-in-difference models they negligibly affect our estimates, suggesting that the control observations largely capture the effects of selection.

The unconditional differences in column (1) show that Head Start children perform remarkably worse on the cognitive tests. The regression adjusted estimates in column (2) are near zero, indicating that the differences in column (1) reflect selection. The difference-in-difference estimates are generally small, and none can be statistically distinguished from zero. Overall, Table 5 shows that Head Start centers, on average, perform similarly to non-Head Start centers in terms of raising children's cognitive scores.[28]

It is notable that the estimates from the model in column (2), which we refer to as the regression-adjusted model, are similar to the difference-in-difference estimates. If there is selection into childcare type based on unobservables, the estimates from the regression-adjusted model will be biased by this selection, while in the difference-in-difference models bias from unobserved selection is likely to be reduced. The most reasonable explanation for the similarity in the findings is that selection into childcare is largely captured by the observable information in the data. Although this is somewhat surprising, it is a testament to the exceptional quality of the ECLS-B survey. *Ex post*, our findings would have been qualitatively similar had we simply assumed selection on observables from the onset. However, we would be less confident in the regression-adjusted estimates without the confirmation provided by the difference-in-difference models.

[28] Appendix Table A.2 reports estimates for all coefficients from the models in columns (2), (4) and (6) in Table 5.

V. Robustness and Other Issues

Robustness of findings

We first evaluate the robustness of our findings to adjustments to the treatment and control definitions. We consider defining treatments as being in care for 7-12 months, and 9-15 months, and compare these treatments to controls who were reported to be in care for 0-2 and 0-3 months (we also compare the 6-12 month treatment group to the new control group). We then return to our preferred treatment and control samples and relax some of the other data restrictions. Specifically, we include children whose parents reported moving no more than once, and no more than twice, between the age-2 and age-4 surveys; and children who were reported to be in care for more than 40 hours per week. For the models that include movers we note that treatment misclassification errors will be more common. However, in practice, any bias from the misclassification errors appears to roughly cancel out in the difference-in-difference models (of note, approximately half of the movers are classified as treatments, and half as controls, in each program). Including the movers increases our sample size by over 50 percent.

Our results are reported in Table 6. For brevity, we only report estimates from the covariate-adjusted, unrestricted difference-in-difference models (the first row of the table shows the baseline estimates from Table 5 for comparison). In math, Table 6 provides no evidence to overturn our primary finding that Head Start and non-Head Start centers perform similarly. In literacy, while none of the estimates can be statistically distinguished from zero, it is noteworthy that the point estimates are consistently negative. This raises the possibility that Head Start underperforms in literacy, but the evidence is not strong enough to scientifically support this claim.

Next, we consider an alternative specification analogous to the model in (4) where we replace the treatment indicator with a quadratic function of months-in-care for each child.

$$Y_i = \beta_0 + A_i\gamma_1 + X_i\gamma_2 + MIC_i\gamma_{31} + MIC_i^2\gamma_{32} + HS_i\,\delta_1 + (A_i * HS_i)\delta_2 + (X_i * HS_i)\delta_3 + (HS_i * MIC_i)\theta_1 + (HS_i * MIC_i^2)\theta_2 + u_i \quad (5)$$

In (5), MIC_i and MIC_i^2 replace T_i from equation (4), where *MIC* indicates months-in-care. The coefficients of interest in (5) are θ_1 and θ_2, which characterize the effects of months-in-care in Head Start relative to months-in-care in non-Head Start centers. A benefit of defining treatment duration as in (5) is that we can include children who were reported to be in care for four or five months at the time of the age-4 survey – these children were not identified as either treatments or controls in the previous models.[29]

Table 7 reports estimates that are analogous to those in Table 5, but based on the new estimation sample and the model in (5). The first column of the table shows the unconditional differences in outcomes between Head Start and non-Head Start children, and the second column shows the regression-adjusted differences. The unconditional differences again show that Head Start children perform considerably worse on the cognitive tests. The regression-adjusted differences in math and literacy are both nominally negative, but statistically insignificant. Moving to the difference-in-difference estimates, the results in Table 7 are qualitatively consistent with those in Table 5 – they do not indicate any differences in performance between Head Start and non-Head Start childcare centers.

Finally, using a much larger sample, we also considered an IV strategy where we instrumented for months-in-care using assessment dates from the ECLS-B survey as a source of

[29] Adding these children, net of the other data restrictions, increases our sample size by approximately 12 percent. Also, similarly to the analysis above, our findings from this model are not qualitatively sensitive to relaxing the other data restrictions.

exogenous variation (following Fitzpatrick et al., 2011).[30] However, the IV approach was unsuccessful, and therefore, we omit the results for brevity.[31] The IV estimates were clearly too large, which was obvious when we compared the Head Start IV estimates to the *Impact Study* estimates. The primary problem with the instrument is that we do not have access to a pre-test, and we suspect that we could not adequately separate the test-date variation from variation in age (we tried to deal with this issue in several ways without success). Of interest, however, is that even in the IV models the estimated Head Start and non-Head Start program impacts are very similar, which would occur under two conditions: (1) Head-Start and non-Head Start centers have similar impacts, and (2) the biases in the IV estimates are similar for Head Start and non-Head Start children.

Does our Analysis Favor Head Start?

Our analysis may favor Head Start because of the relative disadvantage of Head Start children. Although we would typically be concerned about the reverse, we cannot preclude this possibility. One of the most likely ways that our results would be biased in favor of Head Start would be if the testing instruments in the ECLS-B had strong ceiling effects. If this were the case, the test scores for the higher-achieving non-Head Start children would be mechanically restricted relative to the Head Start children, which would give Head Start an advantage in our analysis. Following Koedel and Betts (2010), we test for ceiling effects in the ECLS-B testing instruments and find no evidence of a test-score ceiling in either test. In fact, inferring from Koedel and Betts (2010), there are mild *floor* effects in the literacy test, which would work against Head Start in our comparison (scores for the lowest performers are mechanically

[30] This strategy does not require great care in identifying treatment and control groups based on treatment duration so we were able to use most of the observations from the ECLS-B for children who were in one of the two care-types.

[31] Results are available from the authors upon request.

inflated). The test-score floor in the literacy test is consistent with our estimates, which show that Head Start has a nominally smaller relative effect in literacy.[32]

It is also possible that disadvantaged children gain more from childcare relative to non-attendance. For example, it could be that advantaged children gain little from attending childcare relative to staying at home, perhaps because they have more-positive home environments. Alternatively, disadvantaged children could benefit more from going to childcare, if for no other reason than because they are not staying at home. If the marginal benefit to childcare attendance over non-attendance is higher for disadvantaged children, this would favor Head Start in our analysis. We use the heterogeneity in family income within the non-Head Start sample to evaluate whether this explanation is likely to be driving our findings (see Table 1).[33] Specifically, using the income categories reported in Tables 1 and 2, we divide the non-Head Start children into bins and estimate income-specific childcare effects. Separating the income controls from the other *X*'s for illustration, and defining the vector of income indicators for child *i* by INC_i, we estimate:

$$Y_i = \phi_0 + A_i\phi_1 + X_i\phi_2 + T_i\phi_3 + INC_i\lambda_1 + (T_i * INC_i)\lambda_2 + \varepsilon_i \quad (6)$$

The estimates of λ_2 are of interest in (6), and we report our findings in Table 8. The omitted group is the highest income category. Although the estimates are noisy, they are not consistent with disadvantaged children gaining more from childcare treatment, at least within the non-Head Start sample.[34]

[32] Koedel and Betts (2010) use skewness to measure ceiling effects. The skewness in the distribution of math scores is approximately -0.04, and in literacy it is 0.83.

[33] There is not enough heterogeneity in income, and there are too few observations, to reasonably estimate the parameters in (6) using the Head Start sample.

[34] This test is imperfect, but it was the best we could do with the available data. The primary problem with the test, of course, is that there is likely to be selection into childcare centers within the non-Head Start sector. If lower-income children attend lower-quality centers, our estimates from equation (6) will understate their benefits from childcare inputs.

VI. Conclusion

For Head Start and non-Head Start childcare centers, we estimate average program impacts by comparing outcomes from children who received treatment to outcomes from children who selected into treatment but received very little care. We use a difference-in-difference framework to evaluate the relative performance of Head Start. We find that Head Start centers, on average, perform similarly to non-Head Start centers.

Our analysis suggests that the common perception that Head Start is performing poorly is driven more by lofty expectations for the program than by the program's actual performance. Implicit in calls for improvements to the delivery of Head Start is the notion that the program is somehow "broken" and must be fixed, but our analysis does not support this contention. Policymakers and other interested parties may or may not be satisfied with the outcomes generated by Head Start, but speaking comparatively, the program does not appear to be underperforming.

Two qualifications to our study merit attention. First, our findings do not preclude the possibility that Head Start can perform better. For example, Gormley and Gayer (2005) and Gormley et al. (2010) evaluate a high-quality pre-kindergarten program in Tulsa, Oklahoma, and they find that it greatly outperforms Head Start in terms of affecting cognitive achievement. Additionally, Currie (2001) discusses several other studies that document large effects for programs that were funded at higher levels than Head Start. However, these previously studied programs are unlikely to be scalable to the level of Head Start without considerable increases in funding, and even then there would be challenges. In this modern era of fiscal constraints, this is an important consideration. A unique contribution of our study is that we compare Head Start to the larger non-Head Start, center-based childcare sector, which provides care to the majority of

children in the United States and where childcare costs are more-closely aligned with the costs of Head Start.[35]

Second, our study does not address the growing body of evidence showing that Head Start has large effects on children's longer-term outcomes (e.g., Garces et al., 2002; Deming, 2009; Ludwig and Miller, 2007). The longer-term benefits of Head Start participation may be sufficient to justify current expenditures even with only small immediate-term effects on test scores and other outcomes (Ludwig and Philips, 2007). Our findings, in conjunction with those from the literature on Head Start's longer-term impacts, suggest that we should be cautious about putting too much emphasis on the immediate-term impacts of Head Start on test scores. If Head Start is performing within expectations in this regard (which our study suggests should be low), but participants meaningfully benefit in the long term, overreacting to the test-score results from the *Impact Study* could do more harm than good.

[35] Data from the ECLS-B suggest that childcare costs at non-Head Start centers are lower than in Head Start, but similar. However, we note that measuring non-Head Start childcare costs is difficult because they are paid by various sources (parents, non-profit groups, government, etc.). Details are available from the authors upon request.

References

Barnett, Steven W. 1992. "Benefits of Compensatory Preschool Education," *Journal of Human Resources* 27 (2), 279-312.

Behrman, Jere R., Yingmei Cheng and Petra E. Todd. 2004. "Evaluating Preschool Programs when the Length of Exposure to the Program Varies: A Non-Parametric Approach," *The Review of Economics and Statistics* 86 (1), 108-132.

Bertrand, Marriane, Esther Duflo and Sendhil Mulllainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1), 249 – 275.

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

Bhatt, Rachana and Cory Koedel. 2010. "A Non-Experimental Evaluation of Curricular Effectiveness in Math," University of Missouri Working Paper 09-13.

Coulson, Andrew J. 2010. "Head Start: A Tragic Waste of Money," *New York Post* (01.28.2010).

Currie, Janet. 2001. "Early Childhood Intervention Programs: What Do We Know?" *Journal of Economic Perspectives* 15 (2), 213-238.

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

-- 2000. "School Quality and the Longer Term Effects of Head Start," *Journal of Human Resources* 35 (4), 755-774.

Deming, David. 2009. Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics* 1(3), 111-134.

Fitzpatrick, Maria D., David Grissmer and Sarah Hastedt. 2011. "What a Difference a Day Makes: Estimating Daily Learning Gains During Kindergarten and First Grade Using a Natural Experiment," *Economics of Education Review* 30 (2), 269-279.

Frisvold, David E. and Julie C. Lumeng. 2011. "Expanding Exposure: Can Increasing the Daily Duration of Head Start Reduce Childhood Obesity?" *Journal of Human Resources* 46(2), p. 373-402.

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

Gormley, William T., Deborah Phillips, Shirley Adelstein, and Catherine Shaw. 2010. "Head Start's Comparative Advantage: Myth or Reality," *Policy Studies Journal* 38(3), 397-418.

Gormley, William T. and Ted Gayer. 2005. "An Evaluation of Tulsa's Pre-K Program," *Journal of Human Resources* 40(3), p. 533-558.

Hebbeler, Kathleen. 1995. "An Old and a New Question on the Effects of Early Education for Children from Low Income Families," *Education Evaluation and Policy Analysis* 7 (3), 207-216.

Jacob, Brian and Lars Lefgren and David Sims. 2008. "The Persistence of Teacher-Induced Learning Gains," NBER Working Paper No. 14065.

Koedel, Cory and Julian R. Betts. 2010. "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation," *Education Finance and Policy* 5 (1), 54-81.

Lee, Valerie E. and Susanna Loeb. "Where do Head Start Attendees End Up? One Reason Why Preschool Effects Fade Out," *Educational Evaluation and Policy Analysis* 17 (1), 62-82.

Lee, Valerie E., Jeanne Brooks-Gunn, Elizabeth Schnur and Fong-Ruey Liaw. 1990. "Are Head Start Effects Sustained? A Longitudinal Follow-up Comparison of Disadvantaged Children Attending Head Start, No Preschool, and Other Preschool Programs," *Child Development* 61 (2), 495 – 507.

Lemke, Robert J., Robert J. Witt and Ann Dryden Witte. 2007. "The Transition from Welfare to Work," *Eastern Economic Journal* 33, 359-373.

Ludwig, Jens and Douglas 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. Philips. 2007. "The Benefits and Costs of Head Start," NBER Working Paper 12973.

Resnick, Gary and Nicholas Zill. Undated. "Is Head Start Providing High-Quality Educational Services? Unpacking Classroom Processes," Westat, Inc.

Rose, Heather and Julian R. Betts. 2004. "The Effect of High School Courses on Earnings," *Review of Economics and Statistics* 86(2), 497-512.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "The Bias due to Incomplete Matching," *Biometrika* 41 (1), 103-116.

U.S. Department of Health and Human Services, Administration for Children and Families. 2010. *Head Start Impact Study*. Final Report. Washington, DC.

Wetzstein, Cheryl. 2010. "Is Head Start a 'Sacred Cow'?" *The Washington Times* (03.30.2010).

Whitehurst, Grover J. 2010. Is Head Start Working For American Students? Brookings Institution Up Front Blog Entry (01.21.2010).

Table 1. Average Characteristics of Various Subsamples of the Data Based on the Age-4 Survey.

			Head Start			Non-Head Start, Center Based Care	
	All Observations	Any Head Start	Exclusive Head Start	Our Preferred Sample	Any Center Care	Exclusive Center Care	Our Preferred Sample
Any Head Start	0.166	1.00	1.00	1.00	0.038	0	0
Head Start Only	0.095	0.575	1.00	1.00	0	0	0
Any non-Head Start, Center Care	0.562	0.130	0	0	1.00	1.00	1.00
Non-Head Start, Center Care Only	0.397	0	0	0	0.706	1.00	1.00
Age (in months)	52.951	53.244[b]	53.216[c]	52.529[b,c]	53.155[e]	53.187[f]	52.426[e,f]
Female	0.492	0.496	0.474	0.497	0.488	0.485	0.471
Asian†	0.101	0.041	0.047	0.037	0.129	0.139	0.153
Black	0.151	0.297[b]	0.265[c]	0.178[b,c]	0.125[e]	0.121[f]	0.079[e,f]
Hispanic	0.198	0.262[a,b]	0.302[a]	0.329[b]	0.141[e]	0.138	0.112[e]
White	0.435	0.238[b]	0.249[c]	0.320[b,c]	0.506[e]	0.507[f]	0.559[e,f]
Other	0.115	0.162	0.137	0.129	0.100	0.095	0.097
Birthweight = Low†	0.155	0.167	0.173	0.156	0.151	0.148	0.137
Birthweight = Very Low†	0.106	0.117	0.113	0.120	0.109	0.102	0.094
Family Income < $20,000	0.214	0.445	0.469	0.440	0.131[e]	0.117[f]	0.081[e,f]
$20,000 ≤ Family Income < $35,000	0.201	0.303	0.313	0.347	0.142	0.137	0.143
$35,000 ≤ Family Income < $50,000	0.147	0.137	0.124	0.102	0.131[e]	0.132[f]	0.162[e,f]
$50,000 ≤ Family Income < $75,000	0.158	0.061	0.050	0.058	0.182	0.184	0.207
Family Income ≥ $75,000	0.281	0.053	0.042	0.040	0.413	0.430	0.408
Highest Parental Educ < High School	0.095	0.169	0.195	0.187	0.045	0.044	0.048
Highest Parental Educ = High School	0.238	0.393	0.404	0.404	0.163[e]	0.149	0.125[e]
Highest Parental Educ = Some College	0.302	0.339	0.310	0.320	0.287	0.275	0.256
Highest Parental Educ = 4-Year Degree	0.204	0.070	0.060	0.059	0.268[e]	0.280[f]	0.342[e,f]
Highest Parental Educ = Graduate School	0.161	0.030	0.031	0.021	0.237	0.252	0.229
Number of family members less than 18	2.524	2.60[a,b]	2.72[a]	2.77[b]	2.37[e]	2.38[f]	2.54[e,f]
Age 9-Months Motor-Skills Test Score	0	0.073	0.049	0.025	-0.052	-0.064	-0.176
Age-9 Months Motor Score is Unknown/Missing	0.041	0.042	0.041	0.044	0.042	0.042	0.036
Age 9-Months Mental-Skills Test Score	0	0.005	0.008	0.021	-0.010	-0.017	-0.057
Age-9 Months Mental Score is Unknown/Missing	0.038	0.035	0.036	0.040	0.039	0.039	0.032
Age-2 Motor-Skills Test Score	0	0.011	-0.012	0.026	0.005	0.006	-0.048
Age-2 Motor Score is Unknown/Missing	0.096	0.096	0.098	0.097	0.094[e]	0.089[f]	0.066[e,f]
Age-2 Mental-Skills Test Score	0	-0.210	-0.244	-0.284	0.106	0.109	0.105
Age-2 Mental Score is Unknown/Missing	0.089	0.087	0.092	0.089	0.086	0.080	0.066
N	8,941	1,482	852	225	5,028	3,551	743

†Indicates oversampling group in ECLS-B data.

Notes: The average characteristics in the "our preferred sample" columns are from our estimation samples, including treatments and controls, where treatment status is defined by 6-12 months in care, and control status is defined by 0-3 months in care. "a" indicates that the differences between columns (2) and (3) are statistically significant, "b" indicates the differences between (2) and (4) are statistically significant, "c" indicates the differences between (3) and (4) are statistically significant, "d" indicates that the differences between columns (5) and (6) are statistically significant, "e" indicates the differences between (5) and (7) are statistically significant, "f" indicates the differences between (6) and (7) are statistically significant. Significance tests are at the 5-percent level.

Table 2. Average Characteristics for Treatment and Control Samples Within Childcare Type, Based on the Age-4 Survey.

	Head Start			Non-Head Start, Center-Based Care		
	Controls	Treatments	P-Value	Controls	Treatments	P-Value
Time in Care (in months)	2.09	8.17	0.00	2.18	8.42	0.00
Age (in months)	51.49	53.50	0.00	51.80	53.20	0.00
Female	0.45	0.53	0.26	0.48	0.47	0.66
Asian†	0.031	0.065	0.27	0.140	0.166	0.33
Black	0.208	0.157	0.35	0.061	0.089	0.15
Hispanic	0.229	0.333	0.10	0.072	0.135	0.01
White	0.375	0.315	0.37	0.631	0.503	0.00
Other	0.156	0.130	0.59	0.096	0.106	0.66
Birthweight = Low†	0.135	0.176	0.43	0.157	0.118	0.13
Birthweight = Very Low†	0.135	0.093	0.34	0.088	0.080	0.71
Family Income < $20,000	0.365	0.481	0.09	0.074	0.069	0.78
$20,000 ≤ Family Income < $35,000	0.375	0.352	0.73	0.105	0.172	0.01
$35,000 ≤ Family Income < $50,000	0.125	0.074	0.22	0.190	0.141	0.08
$50,000 ≤ Family Income < $75,000	0.094	0.056	0.30	0.226	0.193	0.28
Family Income ≥ $75,000	0.042	0.037	0.87	0.405	0.425	0.58
Highest Parental Educ < High School	0.146	0.213	0.22	0.041	0.037	0.79
Highest Parental Educ = High School	0.396	0.398	0.97	0.091	0.152	0.01
Highest Parental Educ = Some College	0.406	0.269	0.04	0.240	0.287	0.15
Highest Parental Educ = 4-Year Degree	0.042	0.074	0.33	0.386	0.307	0.03
Highest Parental Educ = Graduate School	0.010	0.046	0.13	0.242	0.216	0.39
Number of family members less than 18	2.78	2.82	0.80	2.53	2.50	0.75
Age 9-Months Motor-Skills Test Score	0.062	0.067	0.97	-0.151	-0.163	0.86
Age 9-Months Motor Score is Unknown/Missing	0.052	0.046	0.85	0.033	0.032	0.91
Age 9-Months Mental-Skills Test Score	-0.038	0.107	0.30	-0.068	-0.026	0.55
Age 9-Months Mental Score is Unknown/Missing	0.042	0.046	0.87	0.030	0.029	0.90
Age-2 Motor-Skills Test Score	0.144	-0.042	0.16	-0.024	-0.018	0.93
Age-2 Motor Score is Unknown/Missing	0.073	0.120	0.26	0.055	0.069	0.44
Age-2 Mental-Skills Test Score	-0.202	-0.315	0.39	0.142	0.135	0.93
Age-2 Mental Score is Unknown/Missing	0.063	0.111	0.22	0.052	0.066	0.44
N	96	108		363	348	

†Indicates oversampling group in ECLS-B data.

Notes: Control status is defined by 0-3 months in care and treatment status is defined by 6-12 months in care. P-values indicate probabilities that the differences between treatment and control observations occur by chance within programs.

Table 3. Standardized Differences in Observables Across Subgroups, and Distributions of Standardized Differences under Random Assignment to Treatment.

		Empirical Distributions	
Comparison Group:	Average Standardized Difference	Average	95-Percent Confidence Interval
1) Head Start Treatments – Non-Head Start Treatments	32.9	8.8	5.7-12.2
2) Head Start Treatments – Head Start Controls	12.8	11.3	8.1-15.2
3) Non-Head Start Treatments – Non-Head Start Controls	7.7	6.0	4.2-8.5
4) (Head Start Treatments – Head Start Controls) – (Non-Head Start Treatments – Non-Head Start Controls)	14.3	12.8	8.9-17.3

Table 4. Program-Specific Estimates. Treatment is Defined as 6-12 Months in Care.

	Head Start		Non-Head Start, Center-Based Care	
Treatment Effect on Math Scores	0.144	0.084	0.071	0.103
	(0.127)	(0.121)	(0.070)	(0.059)*
Treatment Effect on Literacy Scores	0.093	0.064	0.102	0.143
	(0.116)	(0.110)	(0.077)	(0.066)**
R^2 (Math)	0.19	0.42	0.09	0.43
R^2 (Literacy)	0.17	0.46	0.05	0.35
Age-in-Months Indicators	X	X	X	X
Other Covariates		X		X
N (Treated)	108	108	348	348
N (Control)	96	96	363	363

** Denotes statistical significance at the 5 percent level.

* Denotes statistical significance at the 10 percent level.

Notes: Robust standard errors in parenthesis.

Table 5. Relative Effects of Head Start. Treatment is Defined as 6-12 Months in Care.

	Basic OLS Model (Treated Samples Only)		Difference-in-Difference Estimates (set $\delta_2 = \delta_3 = 0$)		Difference-in-Difference Estimates (unrestricted)	
Math Scores						
Head Start Selection Effect	--	--	-0.712	-0.133	-0.726	-0.385
			(0.100)**	(0.099)	(0.174)**	(0.312)
Head Start Treatment Effect	-0.578	0.002	0.121	0.134	0.073	-0.019
	(0.101)**	(0.109)	(0.138)	(0.124)	(0.143)	(0.129)
Literacy Scores						
Head Start Selection Effect	--	--	-0.619	-0.046	-0.535	-0.106
			(0.094)**	(0.095)	(0.171)**	(0.292)
Head Start Treatment Effect	-0.588	-0.001	0.016	0.016	-0.009	-0.080
	(0.098)**	(0.106)	(0.134)	(0.121)	(0.138)	(0.123)
R^2 (Math)	0.06	0.47	0.16	0.44	0.18	0.47
R^2 (Literacy)	0.06	0.41	0.11	0.37	0.12	0.41
Age-in-Months Indicators		X	X	X	X	X
Other Covariates		X		X		X
N (Head Start Treated)	108	108	108	108	108	108
N (Head Start Control)	--	--	96	96	96	96
N (Center Care Treated)	348	348	348	348	348	348
N (Center Care Control)	--	--	363	363	363	363

** Denotes statistical significance at the 5 percent level.

* Denotes statistical significance at the 10 percent level.

Notes: Robust standard errors in parenthesis. Appendix Table A.2 reports coefficient estimates for the covariates from columns (2), (4) and (6), excluding the age indicators.

Table 6. Robustness Exercise 1. Difference-in-Difference Estimates from Covariate-Adjusted, Unrestricted Models Using Alternative Definitions of Treatments and Controls, and Relaxing Key Data Restrictions from the Primary Analysis.

	Treatment Effects		
	Math	Literacy	Total Observations
Treatments: 6-12 months	-0.019	-0.080	915
Controls: 0-3 months	(0.129)	(0.123)	
Robustness Checks			
Treatments: 6-12 months	-0.073	-0.131	741
Controls: 0-2 months	(0.137)	(0.130)	
Treatments: 7-12 months	0.020	-0.088	807
Controls: 0-3 months	(0.137)	(0.125)	
Treatments: 7-12 months	-0.014	-0.126	633
Controls: 0-2 months	(0.142)	(0.129)	
Treatments: 9-15 months	0.055	-0.121	696
Controls: 0-3 months	(0.155)	(0.145)	
Treatments: 9-15 months	0.102	-0.060	522
Controls: 0-2 months	(0.149)	(0.140)	
Include Movers (one move only)	0.022	-0.093	1,398
	(0.102)	(0.097)	
Include Movers (one or two moves)	-0.058	-0.106	1,561
	(0.096)	(0.091)	
Remove Hours-in-Care Restriction	0.004	0.049	966
	(0.127)	(0.121)	
Age-in-Months Indicators	X	X	
Other Covariates	X	X	

Notes: Robust standard errors in parenthesis.

Table 7. Robustness Exercise 2. Relative Effects of Months-In-Care in Head Start.

	Basic OLS Model (entire sample)		Difference-in-Difference Estimates (set $\delta_2 = \delta_3 = 0$)		Difference-in-Difference Estimates (unrestricted)	
Math Scores						
Head Start Indicator	-0.652	-0.052	-0.811	-0.106	-0.853	-0.506
	(0.068)**	(0.071)	(0.185)**	(0.169)	(0.226)**	(0.350)
Head Start Months-in-Care	--	--	0.060	0.002	0.075	0.063
Linear Effect			(0.074)	(0.071)	(0.074)	(0.072)
Head Start Months-in-Care	--	--	-0.004	0.001	-0.006	-0.006
Quadratic Effect			(0.006)	(0.006)	(0.006)	(0.006)
Literacy Scores						
Head Start Indicator	-0.624	-0.017	-0.577	0.150	-0.473	-0.208
	(0.065)**	(0.070)	(0.173)**	(0.164)	(0.217)**	(0.337)
Head Start Months-in-Care	--	--	-0.014	-0.072	0.003	0.023
Linear Effect			(0.073)	(0.068)	(0.071)	(0.068)
Head Start Months-in-Care	--	--	0.000	0.006	-0.002	-0.004
Quadratic Effect			(0.006)	(0.006)	(0.006)	(0.006)
R^2 (Math)	0.076	0.444	0.169	0.444	0.181	0.469
R^2 (Literacy)	0.063	0.392	0.121	0.392	0.138	0.424
Age-in-Months Indicators		X	X	X	X	X
Other Covariates		X		X		X
N (Head Start)	222	222	222	222	222	222
N (Center Care)	805	805	805	805	805	805

** Denotes statistical significance at the 5 percent level.

* Denotes statistical significance at the 10 percent level.

Notes: Robust standard errors in parenthesis. In the difference-in-difference models, the coefficient on the Head Start indicator reflects the "selection effect" as in Table 5.

Table 8. Income-Specific Childcare Effects for Non-Head Start, Center Care Facilities.

	Income-Specific Effects	
	Math	Literacy
Treatment*(Family Income < $20,000)	-0.332	-0.151
	(0.238)	(0.238)
Treatment*($20,000 ≤ Family Income < $35,000)	-0.020	-0.044
	(0.195)	(0.198)
Treatment*($35,000 ≤ Family Income < $50,000)	-0.127	0.025
	(0.177)	(0.199)
Treatment*($50,000 ≤ Family Income < $75,000)	-0.069	-0.138
	(0.152)	(0.180)

Notes: Robust standard errors in parenthesis. Omitted group is family income > 75,000.

Appendix A
Supplementary Tables

Table A.1. Main Data Sample Details.

	Head Start	Non-Head Start
Full Exclusive-Care Sample	852	3,551
Months-in-Care Variable Missing/Unknown	-158	-647
Months in Care > 12	-68	-413
Months in Care = 4 or 5	-62	-349
Hours in Care Exceeds 40	-15	-247
In Childcare During the Age-2 Survey	-106	-621
Changed Residences Between the Age-2 and Age-4 Surveys	-218	-531
Missing Math and Literacy Scores	-21	-32
Final Sample	204	711

Table A.2. Output from Models in Table 5: Columns 2, 4 and 6 (age-indicator coefficients not reported).

	Simple Regression Models (Column 2)		Restricted Diff-in-Diff Models (Column 4)		Unrestricted Diff-in-Diff Models (Column 6)	
	Math	Literacy	Math	Literacy	Math	Literacy
Head Start Selection Effect	--	--	-0.133	-0.046	-0.385	-0.106
			(0.099)	(0.095)	(0.312)	(0.292)
Head Start Treatment Effect	0.002	-0.001	0.134	0.016	-0.019	-0.080
	(0.109)	(0.106)	(0.124)	(0.121)	(0.129)	(0.123)
Female	-0.081	-0.004	-0.036	0.013	-0.061	-0.001
	(0.071)	(0.081)	(0.051)	(0.110)	(0.059)	(0.068)
Asian†	0.395	0.519	0.414	0.449	0.384	0.454
	(0.122)**	(0.135)**	(0.083)**	(0.091)**	(0.088)**	(0.101)**
Black	-0.165	0.019	0.018	0.149	0.033	0.183
	(0.145)	(0.143)	(0.100)	(0.108)	(0.129)	(0.145)
Hispanic	-0.124	-0.131	-0.074	-0.125	-0.121	-0.175
	(0.111)	(0.123)	(0.084)	(0.088)	(0.107)	(0.119)
Other	-0.031	-0.023	0.065	0.032	0.130	0.118
	(0.125)	(0.139)	(0.087)	(0.099)	(0.105)	(0.127)
Birthweight = Low†	0.023	0.165	0.020	0.014	-0.041	0.017
	(0.122)	(0.125)	(0.081)	(0.086)	(0.096)	(0.107)
Birthweight = Very Low†	-0.245	-0.031	-0.189	-0.032	-0.221	0.026
	(0.159)	(0.159)	(0.106)**	(0.099)	(0.132)*	(0.124)
Twin	-0.181	-0.059	-0.123	0.001	-0.171	-0.070
	(0.115)	(0.109)	(0.082)	(0.082)	(0.097)*	(0.101)
$20,000 ≤ Family Income < $35,000	0.073	0.245	0.104	0.188	0.153	0.171
	(0.140)	(0.119)**	(0.099)	(0.089)**	(0.148)	(0.138)
$35,000 ≤ Family Income < $50,000	0.265	0.369	0.238	0.277	0.241	0.241
	(0.159)*	(0.157)**	(0.110)**	(0.108)**	(0.150)	(0.145)*
$50,000 ≤ Family Income < $75,000	0.176	0.145	0.269	0.263	0.253	0.187
	(0.154)	(0.156)	(0.108)**	(0.111)**	(0.146)*	(0.145)
Family Income ≥ $75,000	0.221	0.167	0.273	0.252	0.294	0.226
	(0.167)	(0.174)	(0.111)**	(0.114)**	(0.147)**	(0.149)
Parental Education = High School	0.030	-0.134	0.156	0.143	0.378	0.269
	(0.161)	(0.149)	(0.119)	(0.096)	(0.169)**	(0.141)*
Parental Education = Some College	0.126	0.019	0.286	0.176	0.509	0.374
	(0.160)	(0.157)	(0.117)**	(0.101)*	(0.166)**	(0.146)**
Parental Education = 4-Year Degree	0.438	0.376	0.591	0.526	0.796	0.688
	(0.181)**	(0.184)**	(0.129)**	(0.123)**	(0.171)**	(0.159)**
Parental Education = Grad School	0.605	0.891	0.705	0.866	0.874	0.976
	(0.194)**	(0.210)**	(0.137)**	(0.134)**	(0.178)**	(0.170)**
Number of family members < 18	-0.074	-0.156	-0.089	-0.148	-0.134	-0.190
	(0.041)*	(0.041)**	(0.026)**	(0.027)**	(0.033)**	(0.035)**
Age 9-Month Motor-Skills Test	-0.007	0.036	0.029	0.048	0.039	0.045
	(0.053	(0.064)	(0.031)	(0.041)	(0.044)	(0.049)
Age 9-Month Mental-Skills Test	0.009)	-0.036	-0.034	-0.036	-0.042	-0.033
	(0.054)	(0.059)	(0.035)	(0.041)	(0.042)	(0.050)
Age-2 Motor Skills Test	-0.105	-0.092	-0.029	-0.045	-0.022	-0.051
	(0.045)**	(0.050)*	(0.031)	(0.033)	(0.036)	(0.041)
Age-2 Mental Skills Test	0.370	0.361	0.326	0.282	0.320	0.293
	(0.043)**	(0.054)**	(0.031)**	(0.035)**	(0.036)	(0.041)**
Age 9-Month Motor-Skills Test Missing/ Unknown	-0.007	-0.172	-0.159	0.028	-0.179	-0.271
	(0.193)	(0.233)	(0.113)	(0.219)	(0.152)	(0.210)
Age 9-Month Mental-Skills Test Missing/ Unknown	-0.014	0.073	0.413	0.109	0.412	0.338
	(0.347)	(0.336)	0.260)*	(0.305)	(0.290)	(0.325)
Age-2 Motor Skills Test Missing/ Unknown	0.795	0.872	0.406	0.503	0.405	0.784
	(0.265)**	(0.696)	(0.228)	(0.410)	(0.279)	(0.391)**
Age-2 Mental Skills Test Missing/ Unknown	-0.917	-0.922	-0.536	-0.479	-0.605	-0.811
	(0.313)**	(0.729)	(0.271)**	(0.422)	(0.318)*	(0.420)*

	Simple Regression Models (Column 2)		Restricted Diff-in-Diff Models (Column 4)		Unrestricted Diff-in-Diff Models (Column 6)	
	Math	Literacy	Math	Literacy	Math	Literacy
HS*Female	--	--	--	--	0.032	0.24
					(0.124)	(0.117)
HS*Asian†	--	--	--	--	-0.162	-0.493
					(0.298)	(0.290)*
HS*Black	--	--	--	--	-0.117	-0.246
					(0.214)	(0.212)
HS*Hispanic	--	--	--	--	0.049	-0.029
					(0.181)	(0.169)
HS*Other	--	--	--	--	-0.349	-0.434
					(0.198)*	(0.192)**
HS*Birthweight = Low†	--	--	--	--	0.158	-0.127
					(0.182)	(0.159)
HS*Birthweight = Very Low†	--	--	--	--	0.307	-0.094
					(0.225)	(0.202)
HS*Twin	--	--	--	--	0.254	0.335
					(0.198)	(0.178)*
HS*$20,000 ≤ Family Income < $35,000	--	--	--	--	-0.259	-0.118
					(0.204)	(0.174)
HS*$35,000 ≤ Family Income < $50,000	--	--	--	--	-0.262	-0.128
					(0.248)	(0.159)
HS*$50,000 ≤ Family Income < $75,000	--	--	--	--	0.016	0.198
					(0.284)	(0.251)
HS*Family Income ≥ $75,000	--	--	--	--	-0.259	0.008
					(0.204)	(0.328)
HS*Parental Education = High School	--	--	--	--	-0.369	-0.207
					(0.244)	(0.198)
HS*Parental Education = Some College	--	--	--	--	-0.335	-0.386
					(0.241)	(0.206)*
HS*Parental Education = 4-Year Degree	--	--	--	--	-0.496	-0.387
					(0.322)	(0.339)
HS*Parental Education = Grad School	--	--	--	--	0.489	0.642
					(0.416)	(0.369)*
HS*Number of family members < 18	--	--	--	--	0.201	0.178
					(0.058)**	(0.057)**
HS*Age 9-Month Motor-Skills Test	--	--	--	--	-0.030	0.021
					(0.094)	(0.088)
HS*Age 9-Month Mental-Skills Test	--	--	--	--	0.106	0.055
					(0.101)	(0.088)
HS*Age-2 Motor Skills Test	--	--	--	--	0.083	0.040
					(0.089)	(0.075)
HS*Age-2 Mental Skills Test	--	--	--	--	-0.005	-0.139
					(0.081)	(0.084)*
HS*Age 9-Month Motor-Skills Test Missing/ Unknown	--	--	--	--	-0.620	0.649
					(0.464)	(0.450)
HS*Age 9-Month Mental-Skills Test Missing/ Unknown	--	--	--	--	0.591	-0.557
					(0.676)	(0.670)
HS*Age-2 Motor Skills Test Missing/ Unknown	--	--	--	--	0.089	-1.576
					(0.353)	(0.550)**
HS*Age-2 Mental Skills Test Missing/ Unknown	--	--	--	--	0.417	1.968
					(0.499)	(0.623)**

** Denotes statistical significance at the 5 percent level.
* Denotes statistical significance at the 10 percent level.

Appendix B
Comparison of our Findings to the *Impact Study* Findings

We briefly compare our Head Start specific estimates from Table 4 to analogous experimental estimates from the *Impact Study*. Unfortunately, there are several important differences between the studies that limit inference from this comparative exercise. We highlight four issues prior to performing the comparison:

1) The duration of treatment in the *Impact Study* may have been somewhat longer, on average, than in our preferred estimation sample, although any differences are likely to be small. The *Impact Study* compares children's fall to spring test scores. The report indicates that the fall data collection began in October and was "mostly" complete by mid-November, but we were unable to find documentation of the spring data-collection timeframe. Assuming that the spring data collection occurred around May of the following year, the average treatment duration would have been roughly six months for children who were fall-tested in mid-November in the *Impact Study*. Therefore, we expect the treatment durations across studies to be quite similar.
2) The childcare arrangements for control children in the *Impact Study* differ from those of the control children in our Head-Start specific analysis. Our controls are selected based on not receiving any care, while many of the controls from the *Impact Study* attended some alternative care. Also, the controls from the *Impact Study* that did not seek out any form of alternative care (roughly 38 percent) differ from the controls in our study in that they initially sought to be placed in Head Start, while our controls may not have. The implications of these differences for the comparison of the estimates are not immediately clear.
3) Testing Instrument Issues: the testing instruments are different across studies, and the *Impact Study* reports findings from numerous instruments. There are surely differences in the content of the exams. Also, while both studies report estimates in standard deviations of the tests, facilitating some comparison, we standardize scores across the entire age-4 sample in the ECLS-B while the *Impact Study* uses only Head-Start eligible children (of course). The variance of test scores among Head-Start eligible children is smaller than the variance of scores for all children in our data. For the purposes of this comparative exercise, we re-standardize children's outcomes in our analysis using only children from the ECLS-B data who attended some Head Start (N=1,482, see Table 1). This scales up our math and reading estimates by roughly 10 and 20 percent, respectively.
4) The *Impact Study* estimates intention-to-treat effects (ITT), while we estimate treatment effects. Under fairly modest assumptions the ITT effects can be scaled to be comparable to our estimates, although the scaling likely adds some noise to the comparison.

To make the *Impact Study* results comparable to our own, we being by taking the within-subject averages of the ITT effects from the *Impact Study* in math and language arts. These averages serve as rough summary measures of the *Impact Study* findings by subject, which can be compared to our results. Across assessments, the average ITT effect in math for the age-3 cohort is 0.105. For language arts, the across-assessment average is 0.168 (the age-3 cohort is the correct comparison cohort, although the comparison is very similar if we use the age-4 cohort from the *Impact Study* instead).[36] The appropriate scaling factor to convert the ITT's to treatment effects is roughly 1.5 (as noted by Ludwig and Philips (2007), and verified here). Therefore, the *Impact Study* estimates that are most comparable to our estimates are 0.159 and 0.252 for math and language arts (which we label as literacy), respectively.[37]

Table B.1 compares our estimates to these rough summary measures of the *Impact Study* findings. We do not attempt to estimate the noise in the summary measures from the *Impact Study*, but these measures are based on estimates that are themselves quite noisy (see the *Impact Study*'s Main Tables supplement). The table shows that while the *Impact Study* estimates are nominally larger than our program-specific estimates in both subjects, particularly in reading, one set of estimates does not rule out the other.

[36] The *Impact Study* does not report estimates that are statistically insignificant but, of course, these estimates must be included in the comparison. They are available in supplementary tables from the study, obtainable from the Department of Health and Human Services. Also, the language-arts ITT includes one estimate based on parent-reported literacy that is much larger than the other ITT's – this may be related to parents' perceptions of the benefits of their children winning the lottery, rather than actual program benefits. Excluding the parent-reported outcome, the average language-arts ITT is 0.148.

[37] Taking our finding that Head Start and non-Head Start centers have similar effects at face value, an argument could be made that the *Impact Study* scaling factors should be adjusted up to roughly 2.1 (some of the treatments and controls attend non-Head Start centers, which can be viewed as equivalent to attending Head Start). We have some reservations about this adjustment, however, because as noted in the text, our findings are relevant for the entire center-care sector, and the control children in the *Impact Study* likely attended inferior centers. But, in the spirit of these rough comparisons, it is worth noting that a scaling factor of 2.1 could be viewed as an upper bound.

Despite inference from this comparative exercise being limited by several key differences between the studies, we do not uncover a conflict between the experimental *Impact Study* findings and our Head-Start specific estimates. This is consistent with our analysis in the text, which uncovers little evidence to suggest that our findings are driven by selection bias. Finally, we also note that this across-study comparison does not take into account the difference-in-difference aspect of our empirical strategy – that is, even if there is some bias in our Head-Start specific estimates in Table 4, it may be partly mitigated when we compare the Head Start estimates to the non-Head Start estimates.

Table B.1. Comparison between the *Impact Study* estimates and our Head-Start-Specific Estimates.

	Impact Study		Our Study
	Average Intent-to-Treat Effect	Approximate Treatment Effect	Primary Sample Treatment Effect* [95 percent confidence interval]
Math	0.105	0.159	0.092 [-0.169,0.353]
Language Arts	0.168	0.252	0.077 [-0.182,0.336]

* The estimates in column (3) are taken from Table 4, except that for comparative purposes they are re-scaled using the standard deviations of test scores from only the ECLS-B children who were enrolled in Head Start.