Experimental Evidence on Distributional Effects of Head Start

Marianne P. Bitler
University of California, Irvine and NBER

Hilary W. Hoynes
University of California, Berkeley and NBER

Thurston Domina
University of California, Irvine*

This version: June 2014

*Correspondence to Bitler at UC Irvine Department of Economics, 3151 Social Science Plaza, Irvine, CA 92697, phone (949) 824-5606, fax (949) 824-2182, or mbitler@uci.edu; Hoynes at hoynes@berkeley.edu; or Domina at tdomina@uci.edu. This research also supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number PO1HD065704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are grateful to Child Care & Early Education Research Connections for providing access to the experimental data used here. We would also like to thank Josh Angrist, Peg Burchinal, Janet Currie, Greg Duncan, George Farkas, Avi Feller, Jesse Rothstein, Adam Isen, Doug Miller, Luke Miratrix, Deborah Vandell, Chris Walters and members of the UC Irvine Network on Interventions in Development and our INID Advisory Board members Jeff Smith, Howard Bloom, Susanna Loeb, Sean Reardon, Robert Crosnoe, Dale Farran, and Jacquelynne Eccles as well as seminar participants at Claremont McKenna, Stanford, Texas, Tulane, UCLA, UC Davis, UC Irvine, UC Santa Cruz, University of Houston, APPAM, the Institute for Research on Poverty, the NBER Summer Institute, and SREE for helpful comments. Kevin Williams provided expert research assistance.
Abstract

The federal Head Start program funds public preschools for eligible low-income families. In so doing it aims to raise educational attainment levels and narrow educational inequalities. Existing research demonstrates that Head Start has short run positive impacts on cognitive outcomes that fade out for many groups by elementary school (e.g., Currie & Thomas (1995)). However, there appear to be some lasting positive impacts of Head Start on long-run educational and other outcomes (Deming (2009); Garces, Currie & Thomas (2002); Ludwig & Miller (2007)). In this study, we provide the first comprehensive analysis of the distributional effects of Head Start. We do so using the first national randomized experiment testing the effects of the Head Start program (the Head Start Impact Study). We examine effects on cognitive and non-cognitive outcomes and explore the heterogeneous effects of the program through 1st grade by estimating quantile treatment effects under endogeneity (IV-QTE) as well as various types of subgroup mean treatment effects and two-stage least squares treatment effects. We find that (the experimentally manipulated) Head Start attendance leads to large and statistically significant gains in cognitive achievement during the preschool period and that gains are largest at the bottom of the distribution for several measures. However, once the children enter elementary school, the cognitive gains fade out. Importantly, we find that one important group—Spanish speakers—has cognitive gains that persist through 1st grade. We find little effect of Head Start on non-cognitive (social-emotional) outcomes. Our results provide strong evidence in favor of a compensatory model of educational process in this context.
1 Introduction

Created in 1965, the federal Head Start program is among the more prominent educational initiatives in the US. By giving matching grants to programs providing comprehensive early education, health care, and nutritional services to poor children and providing parenting training to their parents; Head Start aims to raise educational attainment levels and narrow educational inequalities. Head Start enrolls more than 900,000 children and has an annual operating budget of nearly $7 billion (US DHHS, Office of Head Start 2008). Despite the recent growth of state public pre-K programs, Head Start remains a central piece of early childhood education for low-income families (Cascio & Schanzenbach (2013)). Additionally, educational inequities—as measured by preschool enrollment—still exist across the income distribution (Magnuson & Duncan (2013)).

Head Start and other efforts to expand preschool education are predicated on the notion that the positive effects that educational interventions have on young children multiply across their life course (Cunha, Heckman, Lochner & Masterov (2006); Heckman (2006), Heckman (2007)). For example, Knudsen, Heckman, Cameron & Shonkoff (2006) summarize research across various fields (economics, neurobiology, and developmental psychology), concluding that early experiences are key to later development, and that early in life is the most promising period for investments in disadvantaged children, where such investments will have high rates of return. Experimental data from the Perry Preschool program, the Abecedarian Project, and the Chicago Child-Parent Centers lend credence to this argument, demonstrating that early education programs can have positive effects on participants’ academic achievement and attainment (Barnett (1996); Campbell & Ramey (1995); Currie (2001); Schweinhart, Barnes & Weikart (1993)).

Decades of research have documented that Head Start generates important and statistically significant impacts on cognitive outcomes (e.g., see Currie & Thomas (1995) and the review in Currie (2001)). Furthermore, recent evidence using a wide range of quasi-experimental methods demonstrate that exposure to Head Start in the 1960s, 1970s, and 1980s had positive long-term effects on youth educational attainment, health, and labor market outcomes (Garces et al. (2002), Ludwig & Miller (2007), Deming (2009), and Carneiro
However, concerns about the ‘fadeout’ of Head Start’s impact on cognitive test scores have surfaced repeatedly in the program’s history. While it is clear that the program boosts attendees’ academic skills in the short-term, several studies demonstrate that these effects fade out as students move from Head Start into Kindergarten and elementary school (e.g., Currie & Thomas (1995), Ludwig & Phillips (2008)). On its face, this pattern of effects seems at odds with the idea that the effects of early skill formation multiplying across later periods of youth development.

In this study, we provide the first comprehensive analysis of the distributional effects of Head Start. We examine effects on cognitive and non-cognitive (social-emotional) outcomes and explore the heterogeneous effects of the program through 1st grade by estimating quantile treatment effects under endogeneity (IV-QTE) as well as various types of subgroup mean treatment effects and two-stage least squares treatment effects. Our analysis is based on data from the Head Start Impact Study, the first national randomized experiment testing the effects of Head Start.

By moving beyond the analysis of mean impacts, we test two competing hypotheses concerning how Head Start impacts vary across the skill distribution. Observational studies indicate that low-achieving children stand to gain the most by enrolling in early education (Manguson, Meyers, Ruhm & Waldfogel (2004); NICHD Early Child Care Research Network (2004)). This effect may be particularly pronounced in the context of Head Start, since the program’s curricula are explicitly geared both towards remedying the skills’ deficits that often disadvantage poor students at the beginning of elementary school and towards improving the parenting practices of their parents (Puma, Bell, Cook, Heid & Lopez (2005)). On the other hand, research on learning trajectories in elementary school and beyond indicates that since academic skills are cumulative, achievement inequalities tend to widen as children progress through school (Stanovich (1986)). Similarly, the theory of dynamic complementarities (Cunha & Heckman (2010)) argues that a higher endowment of human capital in one period raises the productivity of investments in future periods. Our analysis will test the ‘compensatory’ hypothesis (predicting largest gains at the bottom of the skill distribution) against the ‘skills-begets-skills’ hypothesis (predicting largest gains at the top of the skill distribution).
The Head Start Impact Study, mandated by Congress in 1998, is designed to determine “the impact of Head Start on children’s school readiness and parental practices” as well as “under what circumstances Head Start achieves its greatest impact and for which children” (Puma et al. (2005)). The study randomized children applying to oversubscribed Head Start centers for the first time to either an offer of a slot or denial of a slot for one year. The experiment follows nearly 5,000 children in two cohorts that were three or four years old at the time of Head Start application—through first grade, collecting detailed outcomes on academic and social-emotional measures. The HSIS data has been used previously by Gelber & Isen (2013) who study the effect on parent involvement, Walters (2014) and Bloom & Weiland (2013) who examine heterogeneity in effects across program sites, Miller, Farkas, Vandell & Duncan (2014) who look at whether effects vary by the degree of pre-random assignment parental stimulation, and Feller, Grindal, Miratrix & Page (2013) who look at effects for two groups of compliers, those would otherwise have gone to centers if not in the treatment group and those who would otherwise have stayed home.

In this paper, we use the 3-year old HSIS cohort and examine the impacts of Head Start on cognitive and non-cognitive outcomes. We comprehensively explore the heterogeneity of effects from the first year of preschool through 1st grade, using quantile treatment effects as well as mean treatment effects for subgroups. All of our key results use an instrumental variables approach, where the first stage is Head Start participation and the instrument is the offer of a Head Start slot. This allows us way to rescale the results from the reduced form to estimate the effect of the treatment on the treated. We examine cognitive tests such as the Peabody Picture Vocabulary Test (PPVT) as well as various measures from the Woodcock Johnson III (WJIII) battery of achievement tests. We also take advantage of the various social-emotional outcomes; for example, teacher reports of child behavior using the Adjustment Scales for Preschool Intervention (ASPI) and teacher assessments of the relationship with the child using the Pianta scales as well as parent reports of various social-emotional outcomes.

We find that Head Start leads to large and statistically significant gains in cognitive skills in the preschool period. This is evident for the PPVT, which tests receptive vocabulary and oral language development, as well as for Woodcock Johnson III measures of early
mathematics and early literacy. Generally, we find that the gains are largest at the bottom of the distributions of each of these measures of cognitive achievement. The differences in the treatment effect across the distribution are large—ranging from more than a full standard deviation at the bottom of the distribution to about a quarter of a standard deviation at the top for PPVT. The range of effects is substantial for the WJIII measures as well. We also find differences across groups, with larger positive effects for Hispanics, for those with limited English, and for those with low baseline cognitive skills. Overall these results show strong evidence in favor of a compensatory model of educational process. Once the children enter elementary school, however, these cognitive gains fade out. Importantly, though, we find that Spanish speakers have cognitive gains that persist through 1st grade. We find little effect of Head Start on non-cognitive outcomes, either during preschool or through 1st grade.

In order to interpret our findings, it is important to note that in addition to Head Start, many states provide high-quality means-tested early childhood education (Cascio & Schanzenbach (2013), Magnuson & Duncan (2013)). As a result, many poor children in the US have access to state-funded early childhood education as well as to Head Start, and 25% of the control children in the HSIS attended a non-Head Start child care center. As a result, these analyses may understate the effects of Head Start compared to a counterfactual of no formal early childhood education (Shager, Schindler, Magnuson, Duncan, Yoshikawa & Hard (2013)). To address this possibility, we consider the extent to which subgroup-specific estimated program effects correlate with subgroup-specific rates of treatment compliance and center care participation. Interestingly, we find little role for these factors in explaining underlying differences in the treatment effects across the distribution or in how these distributional effects differ across subgroups.

The remainder of our paper proceeds as follows. We begin in section 2 by discussing the literature and theoretical setting for our problem. In section 3, we discuss the HSIS experiment, HSIS data, and our sample. In section 4, we present the mean treatment effects, and in section 5, we discuss the methods. Our main results are in section 6. We provide a discussion in section 7 and conclude in section 8.
2 Background and Context

2.1 Effects of Head Start

Although Head Start has been extensively evaluated over its nearly 50 years of existence, evidence regarding its effectiveness is mixed. It is well established that children enrolled in Head Start experience improvements in cognitive outcomes (see the review by Currie (2001) for a comprehensive treatment). However, it is less clear how long those test score gains persist. For example, Currie & Thomas (1995), use a within-family, sibling-comparison research design—where the key variation comes when one sibling attends Head Start and the other does not—and find that black participants experience fade out in test score gains in elementary grades while white participants’ gains persist into adolescence.¹

More recently, focus has turned to longer-term impacts of Head Start. Four quasi-experimental studies indicate that Head Start has positive long-term effects. Ludwig & Miller (2007) use a regression discontinuity approach based on the initial roll-out of the program in the mid-1960s and find that Head Start leads to significant decreases in child mortality and increases in educational attainment. Garces et al. (2002) use a sibling-comparison research design and find that Head Start significantly increases educational attainment for whites but not blacks; yet it significantly decreases criminality for blacks but not whites. Deming (2009) uses a similar family fixed effect approach and finds that Head Start participation has a positive effect on young adults’ life chances (measured as an index that includes high school completion, college attendance, idleness, crime, teen pregnancy, and self-reported health status). Carneiro & Ginja (Forthcoming) use variation in Head Start income eligibility rules (which vary by state, year, family size and family structure) and find that the program leads to improvements in behavioral problems and health in adolescence and reduces crime and idleness in young adulthood.

What is not yet fully understood is what explains the pattern of fade-out of cognitive impacts yet positive effects on human capital, health and labor market outcomes in the longer-run. One possibility is that parents respond to the treatment and the long-run impacts

¹An alternative approach used by Griffen (2014) uses observational data but a more structural approach which finds effects like those in the HSIS. This paper concludes that increasing access to Head Start improves cognitive effects which exceed those from providing parents with a child care subsidy (an alternative policy).
operate through changes in their behavior (Gelber & Isen (2013)). Another possibility is
that the long-run treatment operates through changes in non-cognitive or social-emotional
outcomes, something that we explore in our study. Additionally, little is known about
heterogeneous impacts of Head Start. The existing work primarily focuses on differences by
race and gender (as described above). Walters (2014) finds substantial heterogeneity in the
effects of Head Start across centers in the HSIS, suggesting some role for center inputs such
as full-day programs and home visiting. Learning more about the heterogeneity of Head
Start is the focus of our work.

2.2 Theoretical Expectations and Heterogeneity Elsewhere in So-
cial and Educational Policy

In this study, we present a comprehensive analysis of the heterogeneous impacts of the Head
Start program within the context of the first nationally representative randomized experi-
ment of the program. In particular, we focus on estimating the effects of Head Start across
the distribution of cognitive achievement. This approach allows us to test two competing
hypotheses advanced in the broader literature on effects of educational interventions across
the life course.

One theory argues for a “compensatory” effect, where the largest gains to an educational
intervention will accrue to those with lower skills ex-ante. Evidence from observational
studies find evidence in support of this theory in the context of enrolling in early education
(Manguson et al. (2004); NICHD Early Child Care Research Network (2004)). This effect
may be particularly pronounced in the context of Head Start. Head Start has traditionally
focused attention on preparing the most disadvantaged students for school entry, which may
lead Head Start programs to emphasize the most basic cognitive skills. The performance
standards that Congress articulated for Head Start in its 1998 reauthorization speak to this
curricular emphasis, calling on the program to ensure that all students recognize a word as
a unit of print and can identify at least 10 letters (DHHS ACF (2003)). While Head Start
is a highly decentralized program and does not have a single coordinated curriculum, these
performance standards were widely publicized within the program, and may have influenced
instructional priorities (DHHS ACF (2000)). Similarly, Head Start’s mission of serving “at-risk” young children may lead Head Start programs to dedicate particular attention to the most socially and emotionally troubled children.\(^2\)

On the other hand, the theory of “dynamic complementarities” (Cunha & Heckman (2010)) argues that higher endowment of human capital in one period raises the productivity of investment in a future period. Additionally, they advance the idea of “critical periods” whereby human capital investments may be particularly productive during certain periods in the life course. Aizer & Cunha (2012) find evidence in favor of this theory in their sibling analysis of Head Start in the context of the roll-out of the program in the mid 1960s.

In sum, our analysis will test the ‘compensatory’ hypothesis (predicting that the largest gains will accrue at the bottom of the skill distribution) against the ‘skills-begets-skills’ hypothesis (predicting that the largest gains will appear at the top of the skill distribution). We should point out that we analyze these predictions within an experiment and population that is highly disadvantaged relative to the nation. We turn to this below in discussing the study and sample.\(^3\)

3 Head Start Impact Study and Data

In this section, we describe the HSIS experiment and the results of the HHS funded evaluation. We also describe the public use data, our sample, and conduct tests of balance across the treatment and control groups.

\(^2\)Duncan & Vandell (2012) consider the ramifications for both the skills beget skills models and a more complicated child-policy fit model for designing interventions.

\(^3\)The final HSIS report provides some evidence on potential heterogeneous effects of Head Start by various subgroups. Puma, Bell, Cook & Heid (2010) report that the offer of a Head Start slot had greater positive short-term cognitive achievement effects for students with special needs, for students who entered into the program with very low cognitive skills, and for black students and English-language learners. While the program effects for black students and English language learners typically decayed by the end of Kindergarten, Puma et al. (2010) find some evidence to suggest that effects of Head Start program offers for special-needs and low-performing students persist through first grade.
3.1 The HSIS Experiment and Evaluation

The HSIS grew out of a Congressional mandate to evaluate the program which was part of the 1998 re-authorization of Head Start. The HSIS sample consists of 4,442 children who were new applicants to oversubscribed Head Start centers. The sample is nationally representative of children at such centers and consists of 84 regional Head Start programs spanning 353 centers.\(^4\) In the Fall of 2002, applicants were randomly assigned to a treatment group that received an offer to enroll in the Head Start center they applied to and a control group that did not.\(^5\)

The HSIS consists of two age cohorts: 3-year olds and 4-year olds. The experiment was intended to determine the effects of one year of Head Start. Most children in the 4-year old cohort would have been expected to transition to Kindergarten in year two and the 3-year old cohort would largely have been expected to have another year of HS after the year of the experiment (the Age-4 year) before entering Kindergarten. While many 3-year olds in the treatment group would have been expected to continue in Head Start, this is not an explicit component of the experimental treatment, and indeed many control children also attended HS at age 4 while some treatment children left Head Start programs.

The evaluation of the HSIS expects to analyze data for children through grade 5. As of this writing, the HHS-funded evaluation has released final reports on outcomes through grade 1 (Puma et al. (2010)), and more recently through grade 3 (Puma, Bell, Cook, Heid, Broene, Jenkins, Mashburn & Downer (2012)). In our analysis we use the most recently released data including outcomes through the end of first grade.

The reports on the HSIS show positive mean effects on students’ cognitive development. At the end of the evaluation’s first year, both 3-year-olds and 4-year-olds in the treatment group scored between 0.10 and 0.30 standard deviations higher than did their peers in the control group on a wide range of cognitive tests. However, most of the cognitive effects decayed as students moved into elementary school. There is little consistent evidence for effects on social-emotional (non-cognitive) outcomes; most such impacts are small and statistically

---

\(^4\)The HSIS is representative of oversubscribed centers in the U.S. With some exceptions to account for participation in other evaluations, all over-subscribed centers were at risk of inclusion. The bulk of centers are oversubscribed.

\(^5\)Severely disabled children were excluded from randomization.
As is common for this type of intervention, in the HSIS the offer of treatment did not translate one-for-one into Head Start participation. In particular, there were “no shows” (those who did not participate in HS despite being offered a HS slot) as well as “crossovers” (those who participated in HS despite not having been offered a HS slot). About 15 percent of the children in the control group ultimately enrolled in Head Start, while about 20 percent of treatment group children did not. In light of this cross-over and incomplete take-up, it is important to point out that the findings described above (from the final report) represent the effects of Head Start enrollment offers (intent to treat or ITT), rather than the effects of Head Start enrollment itself (treatment on the treated or TOT). Given the lack of complete take-up and crossovers in this experiment, we follow Ludwig & Phillips (2008) and Walters (2014) and use an instrumental variable approach in our analysis. We discuss the first stage, the counterfactual child care setting environment (where children would have gone absent the option of the offer of a Head Start slot), and the mean ITT (reduced form) and TOT (IV) results below.

3.2 Sample, Means, and Tests of Balance

We limit our analysis in this paper to data for the 3-year old cohort. The main reason for doing so is that, as stated above, eligibility for inclusion in the experiment required children to be first time applicants to the Head Start program. Since the program serves children during the two years prior to Kindergarten, this restriction was not limiting for 3-year olds. However, limiting the sample in this way for 4-year olds may lead to external validity concerns without more information on why these 4-year olds have not participated in HS in the past. As might be expected, we find the 4-year old cohort to be potentially more disadvantaged compared to the 3-year old cohort, with a higher share of children living in Spanish speaking households and living with lower educated mothers. Perhaps more relevant given the current

---

6Any positive social-emotional effects were limited to parent reports for the 3-year old cohort; for example parents of 3-year olds offered a Head Start slot reported closer and more positive relationships with their children at the end of 1st grade than did parents in the control group (Puma et al. (2010)). The teacher reports, which are universally available for all children beginning in Kindergarten, show no significant effects of Head Start on social-emotional outcomes.
policy setting, increasingly 4-year olds have many other options besides Head Start with the growth of the state pre-K movement (Cascio & Schanzenbach (2013)).

The 3-year old sample consists of 2,449 children, with 1,464 in the treatment group and 985 in the control group. Due to a complicated sample design and adjustments for attrition, there are a variety of sampling and non-response weights available in the data. We use the baseline weights which we augment in order to balance non-response as discussed below.\(^7\)

Randomization occurred in the summer and fall of 2002, and we use HSIS data that span the period from the summer after application to a Head Start center through first grade (if the child is on track). In our analysis, we refer to the main intervention year for our 3-year olds as the “Head Start year” which corresponds to the 2002–03 academic year, which is followed by the “Age-4 year” (2003–04), the Kindergarten year (2004–05), and first grade year (2005–06). Thus, these labels refer to the normative outcome for the children in each year. The HSIS data consist of the results of interviews with parents, teachers, and center directors as well as cognitive and social-emotional tests. We also make use of baseline data from parent interviews and baseline tests given to the children in the Fall of 2002.

Table 1 summarizes the cognitive and social-emotional (non-cognitive) measures that we use, as well as the years during which they are available for all children. For cognitive outcomes, the tests we focus on examined early language and literacy through tests of vocabulary, oral comprehension, pre-writing, and pre-reading as well as testing early numeracy and math skills. Our first important outcome is the Peabody Picture Vocabulary Test (PPVT) of vocabulary knowledge and receptive language.\(^8\) We also use the Woodcock Johnson III measure of Pre-Academic Skills, which is a composite of three components measuring language, literacy, and early numeracy. We focus on these two measures for two reasons. First, both the PPVT and WJIII Pre-Academic Skills index and its subtests (Applied Problems,

\(^7\)The baseline weights adjust for the complex sampling. The HSIS also makes available non-response adjusted weights, however, we do not make use of these adjustments. Our bootstrapping inference procedure requires that we be able to replicate the process of obtaining our inverse propensity-score adjusted weights to adjust for baseline test scores and demographics. More detail on this is given below.

\(^8\)English test takers were administered the PPVT III, while Spanish speakers were administered both the PPVT III if they had sufficient English as well as the Test de Vocabulario en Imagenes Peabody. A very small number (67) of the children were not eligible to take the PPVT III in English while the vast bulk of the Spanish speakers took this test in English. Thus, the PPVT is a useful pre-test for some outcomes given its score is available for almost all children in the fall of the Head Start year.
Letter-Word, and Spelling) are available for all years of the study. Second, the PPVT has been shown to be a good predictor of later life skills while the WJIII Applied Problems (one of the components of the Pre-Academic skills composite) is the only assessment capturing early math skills, which are also highly predictive of later life skills.\(^9\) Social-emotional outcomes include parents’ reports of child behavior and parent and child relationships as well as teachers’ reports of children’s classroom behavior. Parent-reported measures are provided each year while teachers’ reports only become uniformly available in Kindergarten. (Children who are at home do not get teacher reports of their behavior; this includes a large share of the control group in the “Head Start” year.) Social-emotional skills are measured using the Pianta scale (Pianta (1992), Pianta (1996)) and Adjustment Scales for Preschool Intervention (ASPI). Each of these measures are indices of answers constructed from a series of questions to the teacher or parent, counting affirmative answers (ASPI) or counting responses on a five-point Likert scale (Pianta) or, in some cases, other parent reports of children’s behavior.

Table 2 reports summary statistics for child, parent, and caregiver variables at baseline (Fall 2002). The first column reports means for the control group, weighted using the baseline child weights.\(^10\) As expected given their eligibility for and application to Head Start, these children (and their treatment group counterparts) are fairly disadvantaged. A little less than half of these 3-year olds are living with both biological parents, most children have mothers (or caregivers) with at most a high school diploma or a GED, 39% have mothers who were never married, and 21% are in medium or high risk families.\(^11\) The 3-year olds are about evenly split across three race/ethnicity groups: Hispanic, non-Hispanic black, and non-Hispanic white or other. A little more than a quarter either took some tests in Spanish or speak Spanish at home (this group is labeled “Spanish speaker” in the table). Also reflecting

---

\(^9\)Some of the other WJIII tests included on the survey are not very continuously distributed and perhaps less suitable for our distributional methods. Others are only measured in later years.

\(^10\)We omit observations for children where either there is no propensity score weight, which can occur if there was only 1 child in the center or if there was no valid imputed or actual PPVT at baseline as they could not take the test in English. We discuss the propensity score weighting in the next section.

\(^11\)Household risk is assessed by the number of affirmative reports at baseline to the following five conditions: receipt of TANF or Food Stamps, neither parent has a high school diploma or a GED, neither parent is employed or in school, the child’s biological mother/caregiver is a single parent, and the child’s biological mother was age 19 or younger when child was born. Children with 0–2 risk factors are assigned to be low risk, those with 3 risk factors are assigned moderate risk, and those with 4–5 are denoted high risk.
Head Start eligibility criteria, 10% of the 3-year olds are special needs.

As shown in the bottom of the table, the timing and presence of the baseline test assessment varied across children. Recall that these baseline tests were administered during the Fall of 2002. Unfortunately, it was infeasible to administer these pre-tests to many children before Fall 2002 and thus the assessments for many children took place during the school year. In addition, a further complication is that children were assessed in different months (and thus likely at different ages and developmental points). Table 2 shows that about 16% of the control sample children were assessed before November, a third were assessed in November, another quarter were assessed after November, and 26% have no baseline score. The children with missing baseline tests had their test scores (and other subgroup variables) imputed so these pre-tests could be included in later year analyses.\footnote{We have explored also treating these pre-tests as partial post-tests and excluding them from our propensity score weights. This makes little difference for our findings.}

The second column provides the difference in means of the variables between the treatment and control groups (using the baseline child weights). These differences are also adjusted for arbitrary correlation within Head Start center of random assignment. Consistent with the random assignment, almost none of the demographic characteristics are significantly different between the treatment and control groups (22 of the 25 variables are not statistically different at the 5% level), with these measures collected mostly at baseline. These column 2 results show that the treatment group contains a statistically significantly lower share of children born to teen mothers (significant at the 10% level), a lower share of low risk children (significant the 10% level), a higher share of children with special needs (significant at the 5% level), and a lower share of children with younger caregivers (significant at the 5% level).

Perhaps of more concern is the fact that the timing and presence of baseline test assessments vary significantly between the treatment and control groups. The results show that children in the treatment group were somewhat more likely (10 percentage points more so) to be assessed earlier than the control children (before November) while those in the control were 12% less likely to be assessed at all and more likely to be assessed later (if they were assessed). The control group’s later test assessment will, if anything, be expected to lead to...
a downward bias in the estimated effects of Head Start. This differential attrition might also lead to bias. Thus, to account for the differences in attrition, the month of test assessment, and the other observables, in our results throughout the paper, we construct a weight that adjusts for the difference in selection into the sample. This is discussed below.

### 3.3 Weighting and Adjusting for Differences in Observables

Often we are interested in controlling for baseline characteristics of groups in settings such as this. In particular, in the context of educational interventions, it is typical to adjust for pre-treatment baseline test scores if they are available to account for where “kids came in”. Furthermore, we want to control for other baseline observables to adjust for the (relatively minor) imbalances on the demographic characteristics between the treatment and control groups as well as the differences in baseline test assessment date and differential attrition.

Given our interest in distributional estimates and in particular quantile treatment estimates (QTE) and QTE under endogeneity (IV-QTE), there is a natural way to control for observables in this context with experimental data. Firpo (2007) shows that with selection on observables, one can obtain efficient estimates of the unconditional quantile treatment effects by weighting with inverse propensity-score weights, obtained by predicting treatment status with the observables. Frolich & Melly (Forthcoming) extend this to the case of endogeneity in the key right hand side variable. We use these approaches in our context, where the list of observables incorporates both attrition in the baseline test and baseline test scores (thus richly accounting for baseline scores). In particular, we estimate the propensity score \( \hat{p} \) which is the predicted probability of being in the treatment group \((D_i = 1)\) as a function of baseline characteristics using a logit, estimated using the child baseline weight.\(^{13}\) We then weight each observation by its inverse propensity-score weight:

\[
\hat{\omega}_i = D_i \cdot \frac{1}{\hat{p}_i} + 1 - D_i \cdot \frac{1}{1 - \hat{p}_i}.
\]  

(1)

In our logit model, we control for all of the child and caregiver variables in Table 2. Additionally, to richly control for baseline test scores, we assign dummies for decile of the

\(^{13}\)Zanutto (2006) and Dolton & Smith (2011) discuss use of survey weights with propensity-score weighting estimation.
baseline (2002) PPVT, separately for the four assessment month groupings in Table 2. One of the assessment-month groups is “missing PPVT”, allowing us to use the observations with imputed baseline values of PPVT but account for this potentially important information. We also are able to control for differences in attrition. Finally, we include a full set of fixed effects for the Head Start center to which the child applied. In the estimated propensity score model, of the 39 dummies for assessment-month-group-by-decile of test score only 2 are significant at the 5% level, and there is no consistent pattern in the sign of the coefficients among these baseline test score identifiers. Looking at the overlap of the propensity scores between the treatment and controls, they look extremely similar with trivial non-overlap.

The final column of Table 2 presents the “adjusted” difference in means between the treatment and control. This is the difference in the weighted means using our inverse propensity score weights. As expected, the treatment and control samples look even more balanced after adjusting for observable differences with our inverse p-score weight. Only 2 of combined 25 demographic variables and 3 timing variables or 28 controls are statistically significantly different at the 5% level, compared to 5 of the 28 in column 2, using the baseline child weight. In the remainder of the paper, we use the inverse propensity score weights for all of our results.

4 Mean Treatment Effects and the First Stage

As discussed above, an offer of treatment in the HSIS (an offer of a slot in an oversubscribed center) does not translate one-for-one into Head Start participation (either at the center of random assignment or at another Head Start center). Table 3 provides detailed information on child care locations, separately for the treatment and control groups (as well as for their

---

14 With four assessment-month groups and 10 PPVT deciles, this leads to a total of 40 decile x assessment-month groups. With the constant term, we are left with 39 coefficients.

15 The fixed effects for the center of application help control for the fact that random assignment shares varied by center and also adjust for important potential geographic heterogeneity. Note that the public-use data do not reveal anything about the location of the centers, thus we are not able to measure anything about the local area or population.

16 We also explored a number of alternative propensity-score adjustments, and our findings are robust to these (and in fact to not weighting at all). The results of the propensity score weighting model are available in request.

17 As discussed below, in our bootstrap approach for inference, we reestimate the weights with each bootstrap resample.
difference). The first row provides what we call “the Administrative report” of Head Start participation, which is directly provided in the HSIS data and corresponds to the child having participated in a federally funded Head Start program for some time during the first year of the study. The table shows that 86% of the children offered a slot (treatment group children) are enrolled in Head Start compared to 15% of the control group children. Thus the experimentally manipulated offer of a Head Start slot leads to a 70.5 percentage point increase in Head Start enrollment (column 3).

To gain more insight into the counterfactual care setting environment experienced by these children, we also report in Table 3 the modal child care setting, as reported by the parents in Spring 2003 (at the end of the “Head Start” year), and again in Spring 2004. Looking at the control group means for Spring 2003, about 15% are in Head Start, a quarter are in another center, and almost 60% are in family day care or being cared for by a parent or relative, with the bulk being with a parent or relative. As discussed by Magnuson & Duncan (2013), understanding the counterfactual child care environment is important for providing a context for the results and the expectations for the effects of the intervention. The offer of a Head Start slot mostly displaces other centers and parent/relative care: The table shows a 18 percentage point decline in use of other centers with a Head Start offer and a 49 percentage point decline in use of family day care or parent or relative care.\(^1\)

In the bottom of the table, we provide similar tabulations for parent reports of child care for our 3-year old cohort at the end of the second year of the study (the “Age 4 year”). As discussed above, the HSIS treatment is a one-year treatment. As one might expect given this disadvantaged population as well as the aforementioned growing public pre-K options for children aged 4, there is a significant amount of change in care arrangements between ages 3 and 4 for these children. Additionally, the differences in child care settings narrow between the treatment and control, “blunting” the treatment. At the end of the Age-4 year, 61% of the treatment group is in Head Start compared to nearly half of the controls, with a difference of 13.5 percentage points. The difference in staying with a parent or relative

\(^{18}\)Unfortunately, we do not have “administrative” confirmation about children’s use of any centers but Head Start, and must rely on parent reports of the center at which the children spent the most time. However, note the high degree of concordance of the administrative measure and parent reports of Head Start in spring 2003 (the only year for which we have the administrative measure).
has shrunk to 3 percentage points. Much of the change in the treatment-control difference in child care settings between Age-3 year and Age-4 year comes from changes in child care for the control group. Among control children, 14.6% are in Head Start in year 1 compared to 47.3% in year 2. Many fewer control children are in parent or relative care in the Age 4 Year: 53.6% of the control children are in parent/relative care in year 1 as compared to 10.3% in year 2. For treatment children, most of the change is an increase in other center based care (and a reduction in Head Start), with use of other center care increasing from 6.8% to 25.0%.

It is not completely obvious what one might think of as the appropriate “first stage” treatment here. Our position is that the appropriate first stage outcome is use of Head Start in the “Head Start” year. This is the explicit experimental manipulation—the HSIS offered treatment children a slot in Head Start for one year. We use this as the first stage for the analysis of outcomes in all years in the HSIS. Alternatively, if one considers use of any center-based care in the Head Start year—whether Head Start or other centers—to be the first stage outcome, then the first stage treatment is smaller but still quite large, with an offer of a slot leading to an increase 49 percentage points in the probability of using center care. It is important to note that any reduction in the magnitude of the first stage would lead to larger TOT effects. Our results presented here, then, likely provide a lower bound to the estimated effects of use of Head Start in the HSIS.

Before examining the impacts of HS across the distribution of outcomes, we first present mean treatment effects on our main cognitive outcome, PPVT, and also show balance in attrition with our inverse p-score weights. Table 4 contains mean treatment effects and control group means using our inverse propensity-score weights (in the leftmost 3 columns of the table) and using the baseline weights (in the rightmost 2 columns of the table). For each assessment period, we show results for two different outcome variables: PPVT test scores (top panel) and the probability that the PPVT test score is unavailable/probability of attrition (score unavailable or imputed at baseline or unavailable in the later assessments, bottom panel).\footnote{This is also the first stage used by Walters (2014).}

\footnote{PPVT and one of the Woodcock Johnson III measures were imputed for almost all children missing them at baseline.}
Columns 1–3 show for PPVT the control mean, the mean intent to treat effect (reduced form), and the two stage least squares estimates; each estimated controlling for baseline characteristics (e.g., using inverse propensity score weights). For reference, in columns 4 and 5 we present the (unadjusted for non-response) control group mean and treatment-control difference (intent to treat) using only the baseline child weights.

Turning first to the top panel, the table shows that the baseline average test score for PPVT for the 3-year olds was 231 with a standard deviation of 38 (both measures calculated for the control group). Column 2 suggests that there are at most trivial differences, on average, in the baseline scores across the treatment and control groups, either with or without controls. This is, of course, important as it indicates balance in the randomization on key baseline measures.

The lower half of the table shows that without the use of inverse propensity-score weights, there is a statistically significant difference of 12% in the share of observations with missing (imputed or simply not administered) baseline PPVT scores, with more control observations missing. However, after inverse propensity-score weights are applied, the probability of the test score being missing (imputed or not administered) is balanced (the difference in the probability of it being missing is 0.03 after adjustment). In fact, our inverse propensity-score weights balance this PPVT test non-response in each year from 2003–2006 as well, with small and statistically insignificant differences in non-response in those years (all of these showed evidence of statistically significant differential attrition in the control group before adjustment).

Next we turn to mean treatment effects (reduced form effects) in Spring 2003 (end of the Head Start year) after the first year of the experiment. The HSIS’s offer of a Head Start slot led to a statistically significant increase in PPVT of 7.2 points or 0.19 standard deviations (the unadjusted results, using the baseline weights, are nearly identical). The IV estimates (TOT) indicate that Head Start participation led to a 10.2 point or 0.27 standard deviation increase in PPVT. The IV results scale up the ITT results by about 40 percent.

The mean treatment effects for the years after the Head Start year show fade-out in the experiment’s impacts. The IV estimates show that participation in Head Start leads to a increase in PPVT of 422 points at the end of Age-4 Year, 0.3 points at the end of
Kindergarten, and 2.9 points at the end of 1st Grade. None of these effects are statistically significant. The qualitative finding of fade-out and no significant mean effects of the program after the first year hold for both our adjusted and unadjusted estimates, and for the ITT as well as the TOT estimates.

We now move on to our main analysis, examining the heterogeneity of Head Start and the HSIS across groups and across the distribution.

5 Empirical approach

It is useful to begin with the usual potential outcomes model notation (e.g., Rubin (1974); Holland (1986)) for estimation of the effects of a treatment. Each individual $i$ has two potential outcomes, $Y_{1i}$ and $Y_{0i}$ (for our purposes, a test score or index of student behavior). Person $i$ has outcome $Y_{1i}$ if assigned to the treatment group and outcome $Y_{0i}$ if assigned to the control group. $D(i)$ denotes the group that $i$ is assigned to in a randomized experiment ($D(i) = 1$ if in the treatment group and $D(i) = 0$ if in the control group). The treatment effect on person $i$ is then $\delta_i = Y_{1i} - Y_{0i}$. The fundamental evaluation problem is that we do not observe the treatment effect—that is only one potential outcome is observed for each person $i$. With randomization of treatment, however, we can identify the average treatment effect using the difference in means between the treatment and control group, $\delta = E[\delta_i] = E[Y_{1i}] - E[Y_{0i}]$. This is what we presented above.

5.1 Quantile Treatment Effects

Our interest here is in exploring the heterogeneity in impacts of Head Start, and in particular, in examining the impacts across the distribution of cognitive achievement. For this analysis, we use quantile treatment effect (QTE), and ultimately IV-QTE which we describe in turn. The QTE is simply the distributional analog of the average treatment effect. If $F(y)$ be the cumulative distribution function (CDF) of $Y$ then the $q$th quantile of the distribution $F(y)$ is defined as the smallest value $y_q$ such that $F(y_q)$ is at least as large as $q$. Further, if $F_1$ is the CDF if $D = 1$ and $F_0$ the CDF if $D = 0$, then the QTE estimate is the difference between the $q$th quantiles of these two distributions $y_q = y_{q1} - y_{q0}$, where $y_{qd}$ is the $q$th quantile of
distribution $F_d$.

In general, the joint distribution of $(Y_{0i}, Y_{1i})$ is not identified without further assumptions. However, as with the average treatment effect, randomization of treatment implies identification of the marginal quantiles $y_{qd}$, and thus identification of the differences in their quantiles, $y_q = y_{q1} - y_{q0}$. In an experimental setting such as the HSIS, the QTE is the estimator of this difference in the quantiles of the two marginal distributions. For example, given random assignment, one can consistently estimate the QTE at the median (0.50 quantile) simply by subtracting the control group’s sample median from the treatment group’s sample median.\textsuperscript{21} The only adjustment we make to this simple setup for our reduced form estimates of the effect on the distribution is to use the inverse propensity score weights to account for observables, as discussed above.

This QTE approach is an analysis of the \textit{ex post} cognitive achievement (e.g., PPVT at the end of the Head Start year), and the result is an estimate of how Head Start effects the distribution of cognitive achievement. In this context we can test to what extent the data is consistent with the compensatory or skill-begets-skills theory of educational development. An alternative approach would be to use baseline skills (2002 PPVT) as the measure of the underlying skill and to explore how the treatment varies across this spectrum. A third approach would be to estimate how Head Start effects the value added (e.g., PPVT at the end of the Head Start year − PPVT at baseline). We choose to focus on the QTE (although in one set of results we explore estimating mean treatment effects across the distribution of baseline PPVT) and do so for several reasons. First, not all children were assessed at baseline, and there is imbalance in the rate of missing the PPVT across the treatment and control groups. Second, the month of assessment varied across the treatment and control. Third, a significant share of the assessments took place in the mid and late fall of 2002, months into the first year; thus the pre-test for some may reflect some treatment. In short, the baseline test is not ideal in this context. Finally, we note that our propensity score weighting implicitly adjusts for pre-score differences in general by including them (as well as the attrition indicators) as controls in the calculation of the propensity score weights.

\textsuperscript{21}Empirically, with no covariates, this is identical to a set of quantile regressions (Koenker & Bassett (1978)) of the outcome on the constant and a dummy for treatment status at various percentiles (and with no other additional control variables).
5.2 IV-QTE

Of course, while it is of interest to look at the directly policy relevant outcome “What would happen if we expand the number of Head Start slots at oversubscribed centers?”, it is also of interest to know the underlying parameter, the effect of Head Start. Thus, instead of looking at the reduced form effects of treatment assignment, we want to look at the effects of Head Start take-up. In the usual notation, now the endogenous variable is Head Start participation ($D$) and the instrument $Z$ is treatment assignment.

In the mean outcome setting, two-stage least squares gives us the effect of Head Start on the marginal child induced to enter Head Start only if they obtain an offer under some assumptions (Imbens & Angrist (1994)). In the QTE setting, there are various approaches for dealing with endogeneity, such as Abadie, Angrist & Imbens (2002) for conditional IV-QTE or Chernozhukov & Hansen (2005). We use the approach of Frolich & Melly (Forthcoming) which develops estimators for unconditional QTE when treatment (here use of Head Start) is endogenous. They extend the LATE approach of Imbens & Angrist (1994) and Abadie (2002), and estimate effects of participating in Head Start across the distribution for compliers. This approach uses assumptions that are analogous to those used in the usual LATE settings (there are compliers, monotonicity holds, the instrument is independent (excludable and unconfounded), and there is common support). For asymptotics, their approach also requires unique and well-defined quantiles. If these assumptions are satisfied, the distributions of the outcome under either counterfactual are defined non-parametrically.

5.3 Inference

The HSIS data are not random samples of children across centers, but rather sampled in a complicated fashion. For inference, we are using bootstrapping, with centers being randomly sampled with replacement (with all children from sampled centers appearing in the data whenever a center is sample). We construct confidence intervals using the percentile method. With 999 bootstrap resamples, a 90% confidence interval is given by sorting the resulting

[22] We make use of Frolich & Melly’s (Forthcoming)’s software, using our own estimate of the predicted probability of being in the treatment group as an input, but bootstrap our resulting estimates 999 times with replacement and use the resulting distribution of estimates for inference rather than using their asymptotic calculations. This allows us to account for non-response with our inverse propensity score weights.
bootstrap quantile treatment effect estimates for a given quantile $q$ in increasing magnitude, and selecting the 50th and 950th bootstrap estimates. These point-wise confidence intervals are plotted along with the real data QTE in our figures. Within each bootstrap replicate, the propensity score is re-estimated on the bootstrap sample. A similar process is applied to the IV estimates.

6 Main Results for the Heterogeneous Effects of HSIS

Here we present our main results for examining the heterogeneous impacts of Head Start. We begin by focusing on the Head Start year and move on to examine the impacts through Grade 1.

6.1 IV-QTE Results for the Head Start year, full sample

Before examining the QTE for the Head Start year and beyond, we provide one additional balance test. In particular, in Figure 1, we report the QTE for the baseline PPVT (2002). Recall from above, the QTE are the simple difference in the quantiles of the treated and control groups (where the quantiles are calculated using the inverse propensity score weights). We plot the QTE for each centile between 1 and 99 along with the point-wise 90% bootstrapped confidence intervals. Notably, the confidence intervals all include 0; thus, we have balance across the two groups in the baseline test distribution. (In results not shown, there is only modest imbalance in the baseline QTE before controlling for observables using the inverse propensity score weights.)

Moving on, Figure 2a presents the QTE for the Head Start year, using the PPVT test administered in Spring 2003. The solid line gives the QTE and the long dashed lines give the 90% bootstrapped confidence intervals at each percentile index. The horizontal dotted line presents the average treatment effect (as reported in Table 4). This being the QTE (and not the IV-QTE), the results show the impact of an offer of Head Start, the reduced form or ITT effect. There are several things to note from these results. First, the offer of a HS slot improves cognitive skills throughout the distribution (all the point estimates are positive). Second, the gains associated with treatment are largest at the bottom of the test
score distribution. The estimates at lower end of the skill distribution are very large, ranging from around 30 for quantiles 1–3, to 17 for quantile 13, and to 3 at quantile 43 (which is the last QTE which is statistically significantly different from zero at the 10% level). These effects are also notably substantial relative to the control group standard deviation (38, see column 3 of Table 4). (In this figure, and in the rest of the figures in the paper, to help in interpreting the magnitudes we set the y-axis to be approximately plus or minus one standard deviation of the outcome plotted.)

The QTE capture the impact of the treatment on the distribution of outcomes. For example, Figure 2a measures the impact of being offered Head Start on, for example, the median, 25th percentile, and 75th percentile of the PPVT test score distribution. One limitation of the QTE estimator is that QTE at quantile \( q \) need not equal the treatment effect for an individual located at quantile \( q \) of the control group. While our hypotheses are about effects for individuals of various cognitive abilities, our QTE results will represent effects on the whole distribution. Only with further assumptions, which are perhaps undesirable, can we conclude that the QTE is the treatment effect for a particular individual.\(^{23}\)

Having established the existence of a reduced form effect of an offer of Head Start on PPVT at the end of the first year, we now move to the estimates of the treatment-on-the-treated using the IV-QTE approach of Frolich & Melly (Forthcoming). Figure 2b presents these results. The graph is set up as before, except in addition to the IV-QTE (and the bootstrapped 90% confidence intervals) the horizontal line plots the mean treatment-on-the-treated estimate (e.g., 10.2 from Table 4). The results are qualitatively very similar to, but scaled up from, the QTE. The results show that participating in Head Start leads to increases in cognitive achievement across the distribution, with much larger effects at the bottom of the distribution. The magnitudes are large—gains in the bottom quintile range from 12 to 38, representing between 0.32 and 1 SD units. The results for PPVT in the Head Start year provide strong evidence in favor of the compensatory theory.

We now continue the presentation of IV-QTE now examining the impacts for the Wood-

\(^{23}\)One assumption that allows for treatment heterogeneity is rank preservation (e.g., Heckman, Smith & Clements (1997)), under which a person's location in the distribution is unchanged by the treatment. In this case, the QTE are the same as the distribution of individual treatment effects. Regardless the QTE are useful for assessing overall population gaps.
cock Johnson III tests. In particular, we use the WJIII Pre-academic Skills and its components (Applied Problems, Letter-Word, and Spelling). We use these WJIII tests because they are available in all years of the HSIS. The tests are complementary to the PPVT, with one major advantage being the ability to use the Applied Problems test, which reflects early numeracy. The results are presented in Figures 3a–3d. Starting with Figure 3a, the IV-QTE for Pre-Academic Skills indicate results that largely echo the findings for the PPVT. Participating in Head Start leads to an increase in achievement throughout the Pre-Academic Skills distribution, with significant and large effects at the bottom of the distribution and smaller and statistically insignificant effects beginning around the 40th percentile. Notably, though, there is somewhat less variation in the treatment effects across the distribution compared to the PPVT, and the confidence intervals are wider. Interestingly, when we examine the three test components of the Pre-Academic Skills test, we see that the effects on early numeracy (Applied Problems, Figure 3b) show dramatic evidence of large impacts at the bottom of the distribution, with effects above zero until around the 40th percentile. The results for WJIII Letter-Word (Figure 3c) show evidence of positive effects at the top of the distribution (and some non-negative effects at the bottom). The WJIII Spelling test (Figure 3d) shows little variability across the distribution. Overall the WJIII test confirms the results of the PPVT—the evidence is in favor of a compensatory effect of HS. In addition, having a measure of the impact of HS on early numeracy (Applied Problems) is particularly useful given Duncan, Dowsett, Claessens, Magnuson, Huston, Klebanov, Pagani, Feinstein, Engel, Brooks-Gunn, Sexton, Duckworth & Japel’s (2007) results showing the importance of early numeracy in predicting long term achievement. We do note, however, that the component of the test capturing literacy (Letter-Word) provides some evidence in favor of the skills-beget-skills hypothesis.

6.2 Results for Subgroups in the Head Start Year

Another dimension of heterogeneity is to examine differences across subgroups of the HSIS population. In particular, we begin by analyzing reduced form (ITT) mean impacts by race (Hispanic, non-Hispanic Black, non-Hispanic white), language (Spanish speaker, English speaker), and terciles of baseline PPVT scores. These estimates are in the second column.
of Table 5 for PPVT in the Head Start year (along with the control group means in column 1). This table shows dramatic differences across subgroups, with larger reduced form mean treatment effects for Hispanics (11.5 or 0.29 standard deviations), Spanish speakers (15.0 or 0.47 standard deviations) and those in the bottom tercile of the baseline test score (11.2 or 0.33 standard deviations). These are very large effects.

One possible explanation for the differences in treatment effects across groups is differences in the probability of taking up Head Start across subgroups. Additionally, variation across subgroups may reflect differences in counterfactual child care settings across these subgroups, which is viewed by many as an important context for understanding effects of Head Start (and early childhood programs more generally) over time. In fact, a strength of our approach is that analyzing impacts across subgroups gives us some potential insight into the role played by the first stage and the counterfactual. The third column of Table 5 presents the administrative first stage across subgroups. Figure 4a displays the reduced form and administrative first stages by subgroup together, providing a scatter plot where the administrative first stage is on the x-axis, the reduced form mean treatment effect (ITT) is on the y-axis, and each dot represents a subgroup. (We also add other subgroups to the plots including gender and mother’s education.) Perhaps surprisingly, the graph shows little or no systematic relationship between the first stages and the reduced form effects on PPVT by subgroup. For example, while Hispanics and Spanish speakers have the largest reduced form effects, they don’t have larger first stages than other subgroups. The first stage effect of an offer on Head Start take-up is higher for females compared to males, but their reduced form effects are similar. It is also notable that there is relatively little variation in the first stage across groups. Figure 4b explores the role of the counterfactual with a scatter plot identical to Figure 4a, except the x-axis now plots the control group mean by subgroup for use of non-center care (family day care, parent, relative or unknown). Figure 4c provides another lens by plotting on the x axis the “first stage” of the effect of the Head Start offer on use of non-center care. (See Table 6 for the point estimates behind these two scatter plots.) Echoing the results above, overall the HSIS data show little relationship between the counterfactual care settings and the reduced form PPVT effects across subgroups.

Looking at Figures 4a–4c, children in the bottom tercile of the baseline PPVT score show
evidence of a larger first stage (79% compared to the full sample first stage of 70.5%) and a larger reduced form effect (11.2 compared to the full sample effect of 7.2). Additionally, the counterfactual for this group shows very high use of non-center care, with 66 percent of the control group in a family day care, with a parent or relative, or in unknown care. To explore this further, we take a more non-parametric approach. In Figure 5 we first examine how the first stage varies with the baseline PPVT score by estimating local linear regressions (by baseline PPVT) separately for the treatment and control groups. When seen across the entire distribution of baseline PPVT, the figure shows remarkable uniformly in the first stage effect of an offer on Head Start participation across the baseline skills distribution. We can take this further to explore the reduced form impacts of a Head Start offer on PPVT at the end of the Head Start year, across baseline scores. In particular, in Figure 6a, following Duflo, Dupas & Kremer (2011), we estimate local linear regressions of the PPVT at the end of the Head Start year, as a function of the baseline PPVT, estimated separately for the treatment and control groups. In Figure 6b, we plot the treatment-minus-control differences from these local linear regressions (along with the bootstrapped confidence intervals). These estimates show large gains at the bottom of the baseline skills distribution with some evidence of a positive effect at the top of the baseline skills distribution. Interestingly, this approach yields similar findings to the QTE qualitatively (the figure is best compared to the ITT QTE in Figure 2a). We prefer the QTE in part because of the prevalence of missing baseline test scores (disproportionately imputed for the controls) as well as the variation in month of assessment for this test (leaking into the treatment period). Importantly, for many policy interventions such as Head Start, knowing about effects on the distribution—the QTE—are of substantial interest, and the rhetoric about closing test gaps is about ex-post (of some reform or treatment) scores, not ex-ante ones. The QTE captures precisely these ex-post effects while the local linear results captures the ex-ante effects (in a very unrestricted way). Finally, we note that our inverse propensity score weights implicitly control for baseline score (separately by decile and period of test administration, which allows for a separate set of deciles for the imputed scores).

This analysis suggests that the reduced form effects of the HSIS vary across demographic subgroups. We find little evidence that this is explained by differences in either program...
take-up or counterfactual care settings. Returning to Table 5, we present the IV estimates for our key subgroups in column 4. Not surprisingly given the above discussion, these results are qualitatively similar to the reduced form, but scaled up by 40-50%. Overall, the analysis of mean effects (ITT and TOT) across subgroups again shows evidence of the compensatory hypothesis, with larger effects for those with lower baseline scores, and for Hispanics and those who speak Spanish at home.

We can further explore the differences across subgroups by estimating the IV-QTE separately for each subgroup (these are referred to as “conditional IV QTE” as they are conditioned on being in the subgroup of interest). The results for language subgroups are presented in Figure 7a. (To make for easier viewing, we drop the confidence intervals and the mean treatment effect from these graphs.) The results are dramatic—while both Spanish and English speakers have the largest gains at the bottom of the PPVT distribution (as in the full sample), the IV-QTE for Spanish speakers are very large at the bottom of the distribution (for example they are above 20, which is two-thirds of a standard deviation, through the 60th percentile).

While these conditional IV-QTE are useful, making comparisons across the two groups is complicated by the fact that it is not an “all held constant” situation. In particular, a much larger share of the Spanish-speaking students have PPVT scores in the lower end of the (pooled) test score distribution (for example, see the differences in the control group means by subgroup in Table 5). Given that these conditional IV-QTE plot each line on a common percentile scale (the subgroup-specific percentiles), making conclusions by looking across the subgroup QTE at specific percentiles is problematic. We address this by performing a simple translation to put each subgroup’s QTE on the same absolute scale (the scale we use here is given by the percentiles of the full sample of the control group). We term these “translated” IV-QTE and they are presented in Figure 7b. The figure shows that once the subgroups’ IV-QTE are put on the same absolute scale, the differences between groups become attenuated but are still notable. We see similarly-sized, very large gains throughout the bottom decile of PPVT scores for both groups; yet we see larger effects of the treatment on Spanish speakers (compared to English speakers) throughout the rest of the distribution.24

---

24 Essentially what this is doing is stretching the conditional distributions in some places and shrinking...
Figures 8a and 8b provide similar analyses by race: For non-Hispanic whites, non-Hispanic blacks, and Hispanics. The results are quite dramatic—while the conditional IV-QTE (Figure 8a) show widespread gains for Hispanics, relative to the two non-Hispanic groups, on a common scale, the translated IV-QTE (Figure 8b) shows much more similarity across race/ethnicity groups throughout the distribution.

Taking these results as a whole, we find compelling evidence in the heterogeneity of Head Start on cognitive achievement at the end of the Head Start year. Based on the analysis across the distribution as well as across demographic groups, we find evidence largely in favor of the “compensatory” theory. That is, we find larger gains in the lower end of the skill distribution. We also find evidence that students entering preschool with low English language skills stand to gain more from the experience. To complement this work, we also explored differences across characteristics of the center of random assignment. We found larger effects for centers whose directors cited having a significant amounts of competition from other preschool centers in their area. We also explored, but found little difference in treatment effects, based on variation across the centers of random assignment in teacher credentials, curriculum, staffing, and teacher ratings from direct classroom observation (Arnett and ECERS-R scores).

### 6.3 Results for beyond the Head Start year

Having established the results for the Head Start year, we now move on to examine impacts through grade 1. Figure 9a shows the IV-QTE for PPVT for each year: Head Start year (2003), Age-4 year (2004), Kindergarten year (2005), and first grade year (2006). As we discussed above, the first stage for each of these outcomes is the same—it is Head Start participation in the Head Start year. The figure shows positive (and significant, although CIs are not shown here) effects at the bottom remain for the lower end of the distribution through them in others in order to put them both on the same scale. We use the PPVT score of the control group (for the full sample) at each percentile of the full sample as the anchor. We take the IV-QTE at each percentile of each subgroup, and find the location of that percentile value of each subgroup’s control group in the overall control group distribution. For example, if subgroup A’s median were the 25th percentile in the overall control group, the median IV-QTE for subgroup A would be relocated to the overall control group’s 25th percentile. You can easily see that neither line extends the full range of the x axis indicating the lack of scores in that part of the pooled distribution for each outcome.
the preschool years (end of 2004). However, once this cohort transitions into elementary school, the gains substantially fade.

As with the demographic subgroups, the x axis scale for each of the years corresponds to the percentiles of the PPVT distribution for that year. This is a meaningful scale—it reveals, for example, the effect of participation in Head Start on median PPVT scores at the end of the Head Start year, Age-4 year, Kindergarten and 1st grade. However, it is also useful to translate the IV-QTE to a common scale, as we did with the demographic groups. This is presented in Figure 9b; the translated IV-QTE show much more alignment across years. The results suggest that Head Start brings children up to some level but once they are achieving at that level there is no additional gain to be had.

To highlight further the effects in later preschool, Figures 10a and 10b present results for the end of Age-4 year for WJIII Pre-Academic Skills and Applied Problems tests. The results show persistence of gains into the end of the second year, especially at the bottom of the distribution for the Applied Problems test (marginally significant). By the end of grade 1 (in results not shown here) the effects of Head Start on Pre-Academic Skills have faded out. There is a hint of gains at the bottom of the distribution of the Applied Problems score, but the results are not statistically significant.

Given the large gains experienced by some demographic subgroups in the Head Start Year, it is natural to return to those groups to look at effects on this longer-term outcome. Figure 11a presents the translated IV-QTE for Spanish versus English speakers for the first grade year. Interestingly, here we see some evidence of persistent gains for Spanish speakers, throughout the bottom half of the (overall) distribution. These gains, at 10–15 points, are substantial, measuring 0.2 to 0.25 standard deviations. Additionally, Figure 11b presents the translated IV-QTE for the three terciles of baseline scores, for Grade 1. Importantly, these results suggest that the larger effects of Head Start in preschool for those with low baseline skills does not persist though first grade.
7 Discussion

Our analysis of the cognitive effects of the Head Start Impact Study show that Head Start participation led to significant gains in cognitive skills in the preschool years. Additionally, these gains are largest at the bottom of the cognitive skills distribution. Further, these gains are larger for Spanish speakers as well as those who at baseline are scoring in the bottom tercile of the PPVT distribution.

Given the broad discussion of “compensatory” versus “skills-beget-skills” theories of education, our work provides new and compelling evidence in favor of the compensatory theory. In particular, those with low baseline scores, and those with limited English, gain the most from the intervention. Our analysis shows that these differences cannot be explained by differences in take-up of the program or differences in the counterfactual care setting. Interestingly, Feller et al. (2013) find—using a principal stratification approach—that effects are concentrated among those compliers who in the absence of an offer of a slot would have stayed home with lesser effects among those who would have gone to a center no matter what. Yet, our breakdowns of the first stage by subgroup suggest that the observables we have looked at (mother’s education, gender, race/ethnicity, baseline test score tercile, and language) do not explain who is in which group of compliers.

The results also show that once the children enter elementary school these gains diminish substantially, for the population as a whole. Importantly, though, we do find evidence of persistent effects through grade 1 for those with low English skills.

What can we conclude from these results? First, the gains in preschool may not persist if the elementary schooling environment is not of high quality. As stated in Duncan and Magnuson (2013, p. 118), “If little learning occurs in low-quality schools, then early advantages imparted by programs such as Head Start might be lost. In this case, preschool does not ‘immunize’ against the adverse effects of subsequent low-quality schooling.” We are limited in what we can say about this given that we do not observe much about the schools the study participants attend. Second, the Head Start teachers may be teaching to some proficiency standards (e.g., knowing the ABCs, counting to 10) and the quality of the settings are insufficient to achieve gains beyond that point. As described in Cascio & Schanzenbach
(2013), Head Start programs score a 5 (on a 10 point scale) in the NIEER scale, compared to higher scores for many state-funded preschool programs. In these settings there may not be the capacity for dynamic complementarities (Cunha & Heckman (2010)). Finally, our results suggest that those with limited English skills in early childhood stand to gain from Head Start.

How do our results speak to the broader finding in the Head Start literature of a fadeout in effects on cognitive scores in early elementary school but a rebound in positive outcomes on labor market, human capital and so on? First, our results suggest that there may be subgroups of the population, who enter Head Start without proficiency in English language, who experience lasting effects of Head Start on cognitive outcomes. It would be interesting to know if the experience of these groups can account for the positive long term outcomes others have found.

Second, many have argued that preschool (or other investments during this crucial period) may lead to improvements in non-cognitive outcomes and these may facilitate gains in elementary school and beyond. This is testable in the HSIS. The data contain a host of non-cognitive or social-emotional measures; here we focus on the Pianta scales of student-teacher and child-parent relations and the Adjustment Scales for Preschool Intervention [ASPI] (a measure of emotional and behavioral adjustment to preschool). (Recall we only have the teacher reports for everyone when they are all in school, thus our reliance on parent reports in the Head Start year.) In Table 7, we present IV results capturing the effect of Head Start attendance on social-emotional outcomes. The table presents effects of Head Start attendance during the Head Start year on parent-reported measures and teacher-reported measures. The social-emotional measures have been standardized to have a mean of 0 and standard deviation of 1 (using the control group mean and SD) and aligned so a higher score is always bad. Starting with the Head Start year, the IV estimates for seven of the nine non-cognitive outcomes have a negative sign—a negative sign implying an improvement in parent reported non-cognitive outcomes. However, only 2 of the coefficients are statistically significant at the 5% level. Such as they are, these two coefficients indicate large declines in parent reports of hyperactivity and externalizing behavior. By contrast, the first grade effects are uniformly small and mostly insignificant for the parent-reported measures, and
particularly for the teacher-reported measures. The sole exceptions are the Pianta measures as reported by the parents—at the end of first grade, children in Head Start have a statistically significant decline in conflict and a decline in the lack of a positive relationship (increase in the probability of a positive relationship). In sum, while there are a few positive findings, the overall finding is one of very small and statistically insignificant effects in the social-emotional domain on average, and this lack of an effect is mirrored in the distributional results.

Finally, Cunha & Heckman (2010) talk about early childhood being both a sensitive and critical period for child development. Duncan et al. (2007) argue that early cognitive scores are predictive of later cognitive scores. Similarly, Magnuson, Ruhm & Waldfogel (2007) provide evidence of what they call “sleeper” effects of preschool on cognitive outcomes. While we can not test these ideas in the HSIS, we explore these issues using the data from Deming’s (2009) children of the NLSY 1979 sample. Using Deming’s observational data, we regress his young adult outcomes on cognitive test scores measured at two points in time: one during the preschool years (PPVT, at ages 3–5) and another taken during middle elementary school (PIAT, at ages 9–11). With this we can explore whether cognitive achievement in preschool provides predictive power for later outcomes (conditional on controlling for the cognitive score at ages 9–11). We restrict the sample to individuals in the top quartile of predicted Head Start attendance (to mimic our HSIS sample). We find some predictive power of preschool cognitive skills in the expected direction—higher scores associated with better outcomes—but only two of these effects (those on repeating a grade, being idle) reach statistical significance. The magnitude of the coefficients on these early PPVT scores, compared to those on the later cognitive measure, are nonetheless suggestive of a role for early cognitive skills in young adult outcomes.\(^\text{25}\)

---

\(^{25}\)We use Deming’s combined sample of siblings and non-siblings and predict Head Start participation using a detailed list of demographics, family characteristics, mother’s AFQT, birth order, and so on. We then limit our sample to individuals in the top quarter of predicted HS participation. The PIAT is a math assessment, chosen because it was available for a larger sample of children; the PPVT was not administered to many middle school children.
8 Conclusion

In this study, we provide a comprehensive evaluation of the first national randomized experiment of Head Start. We focus on the 3-year old cohort and examine impacts on cognitive and non-cognitive outcomes in the Head Start year and through Grade 1. We find that Head Start participation leads to large and statistically significant gains in cognitive skills in the preschool period in receptive vocabulary, early literacy, and early numeracy. However, once the children enter school, the overall cognitive gains fade out. We find little effect of the experiment on non-cognitive outcomes.

We explore variation in the effects across subgroups of the population including race/ethnicity, child speaks English or Spanish, baseline levels of cognitive skills, gender, and education of the mother. We analyze these subgroups in part because they provide a framework through which we can learn about the role played by differences in Head Start take-up and counterfactual child care setting across groups. We find significant variation in the effect of Head Start during the preschool years, with larger effects for Hispanics, Spanish speakers, and those with lower baseline cognitive skills. Interestingly, we find no role for differences in take-up or counterfactual care setting in explaining the results across subgroups. Importantly, we also find that for Spanish speakers, the cognitive gains persist through 1st grade. We find little role for differences in Head Start take-up or the underlying counterfactual child care setting in explaining these heterogeneous effects.

This study provides new evidence that the effects of Head Start are largely consistent with a compensatory theory of education. The cognitive gains are largest at the bottom of the distribution of skills. This is revealed using quantile treatment effects and IV-QTE applied to the ex-post distribution of skills, local linear regression estimates based on baseline skills, and analyses across demographic groups. There is limited evidence of the effects predicted by “skills beget skills” theories.

These results still provide an incomplete picture of the potential benefits of Head Start. Prior work finds that despite a fadeout in effects on cognitive outcomes, there are positive effects on the long term on labor market and human capital outcomes (Deming, 2009). Much more could be learned over time by following these participants into adulthood.
References


DHHS ACF (2003), ‘Initial guidance on new legislative provisions on performance standards, performance measures, program self assessment and program monitoring ACYF-IM-HS-00-03’.


Pianta, R. C. (1996), Student-Teacher relationship scale, Working paper, University of Virginia. Charlottesville, VA.


Table 1: Cognitive and socio-emotional outcomes in the HSIS, by normative year in school

<table>
<thead>
<tr>
<th>Language, literacy, vocabulary</th>
<th>Age 3</th>
<th>Age 4</th>
<th>Kindergarten</th>
<th>1st grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPVT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Letter Word (WJIII)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Spelling (WJIII)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Mathematics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applied Problems (WJIII)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Composite</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Academic Skills (WJIII)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Socio-emotional outcomes, parent reports</th>
<th>Age 3</th>
<th>Age 4</th>
<th>Kindergarten</th>
<th>1st grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive behavior</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hyperactive</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Withdrawn</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Social competencies</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Social skills/positive learning</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Behavioral problems</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Conflict (Pianta)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Closeness (Pianta)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Positive relationships (Pianta)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Socio-emotional outcomes, teacher report</th>
<th>Age 3</th>
<th>Age 4</th>
<th>Kindergarten</th>
<th>1st grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive behavior (ASPI)</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hyperactive (ASPI)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Withdrawn (ASPI)</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Shy (ASPI)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Oppositional (ASPI)</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Problems with peer interactions (ASPI)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Problems with structure (ASPI)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Interaction problems (ASPI)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Conflict (Pianta)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Closeness (Pianta)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Positive relationships (Pianta)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Notes: Table reports whether each of the cognitive and social emotional outcomes are reported in each year for all children. Column 1 reports whether each measure was taken in the experimental Head Start year (when the children are age 3), and columns 2–4 report the same for the second pre-school year (age 4), the normative Kindergarten year, and the normative first grade year. Note that the teacher reports of socio-emotional outcomes are only listed if they were collected from everyone (in pre-school, these were only collected from children who were in organized center care).
<table>
<thead>
<tr>
<th>Table 2: Summary statistics at baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Child characteristics</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Non-Hispanic white: 0.344 -0.022 -0.001</td>
</tr>
<tr>
<td>Non-Hispanic black: 0.338 0.005 0.001</td>
</tr>
<tr>
<td>Hispanic: 0.318 0.017 0.0002</td>
</tr>
<tr>
<td>Female: 0.527 -0.023 0.017</td>
</tr>
<tr>
<td>Spanish speaker at home: 0.257 0.011 0.016</td>
</tr>
<tr>
<td>Low risk: 0.789 -0.041* 0.044**</td>
</tr>
<tr>
<td>Medium risk: 0.156 0.013 -0.019</td>
</tr>
<tr>
<td>High risk: 0.055 0.028* -0.025*</td>
</tr>
<tr>
<td>Special needs: 0.103 0.031** 0.008</td>
</tr>
<tr>
<td>Lives with both biological parents: 0.499 -0.003 0.024</td>
</tr>
<tr>
<td>Urban: 0.793 0.001 0.006</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Mother/caregiver characteristics</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Non-Hispanic white: 0.364 -0.020 -0.008</td>
</tr>
<tr>
<td>Non-Hispanic black: 0.339 0.001 0.010</td>
</tr>
<tr>
<td>Hispanic: 0.298 0.018 -0.001</td>
</tr>
<tr>
<td>Mother was teenager at birth: 0.176 -0.043* 0.023</td>
</tr>
<tr>
<td>High School dropout: 0.346 -0.032 0.001</td>
</tr>
<tr>
<td>Only high school diploma/GED: 0.325 0.033 -0.024</td>
</tr>
<tr>
<td>More than high school: 0.329 -0.001 0.023</td>
</tr>
<tr>
<td>Mom is married: 0.457 -0.018 0.022</td>
</tr>
<tr>
<td>Mom is divorced: 0.149 -0.010 0.017</td>
</tr>
<tr>
<td>Mom is never married: 0.394 0.028 -0.038</td>
</tr>
<tr>
<td>Age 20–24: 0.316 -0.052** 0.036</td>
</tr>
<tr>
<td>Age 25–29: 0.332 -0.016 0.021</td>
</tr>
<tr>
<td>Age 30–39: 0.259 0.029 -0.016</td>
</tr>
<tr>
<td>Age 40 or older: 0.064 0.038*** -0.042***</td>
</tr>
<tr>
<td>Fall 2002 test month/no Fall assessment</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Before November: 0.162 0.099*** 0.015</td>
</tr>
<tr>
<td>November: 0.318 0.029 -0.022</td>
</tr>
<tr>
<td>After November: 0.260 -0.013 -0.025</td>
</tr>
<tr>
<td>No Fall assessment (imputed): 0.259 -0.116*** 0.031</td>
</tr>
</tbody>
</table>

*Notes: Table reports summary statistics for baseline characteristics for the 3-year old cohort in the Head Start Impact Study. Column 1 provides the control group means, using the baseline child weights. Column 2 provides the unadjusted difference in means between the treatment and control groups (using the baseline weights). Column 3 provides the adjusted differences in means using the propensity score weights. The risk index for the child is based on Food Stamp/TANF receipt in Fall 2002, whether both parents are high school dropouts, whether neither is working, whether the child's mother was a teenager at birth, and whether the mother is single. Imputed fall assessments were done for children eligible to take the PPVT in English who did not take it. Children who could not take the Baseline PPVT in English did not have a baseline score imputed. SEs for differences allow for arbitrary correlation within center of random assignment. The sample size for column 1 is 985, for columns 2 and 3 it is 2378. *, **, and *** denote significance at the 10%, 5%, and 1% levels, clustering on Head Start center of random assignment.
Table 3: Child care setting, by treatment or control status

<table>
<thead>
<tr>
<th>Head Start in HS Year</th>
<th>Treatment Mean</th>
<th>Control Mean</th>
<th>Difference T-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head Start</td>
<td>0.857</td>
<td>0.153</td>
<td>0.705</td>
</tr>
<tr>
<td>Other Center</td>
<td>0.068</td>
<td>0.252</td>
<td>-0.183</td>
</tr>
<tr>
<td>Family day care</td>
<td>0.014</td>
<td>0.064</td>
<td>-0.050</td>
</tr>
<tr>
<td>Parent/relative</td>
<td>0.094</td>
<td>0.536</td>
<td>-0.442</td>
</tr>
<tr>
<td>Not reported</td>
<td>0.001</td>
<td>0.002</td>
<td>-0.001</td>
</tr>
</tbody>
</table>

**Notes:** Table reports means for the age-3 cohort treatment and control group child care settings and their difference for the official administrative report of Head Start attendance for the Head Start year (top panel), and for parent reports for Spring 2003 (middle panel) and Spring 2004 (bottom panel). Data are from the Head Start Impact Study. Data for the Spring 2003 and 2004 Parent Reports are for the modal child care setting (setting where the child spent the most time) as reported by parents. Data for the Head Start administrative report are those used in the HSIS reports. Statistics exclude observations missing a valid PPVT score. Statistics weighted using inverse propensity score weights.
Table 4: Reduced form effects of being offered a slot in HS for PPVT and for PPVT test score missing (attrition), by year

<table>
<thead>
<tr>
<th></th>
<th>Inverse P-Score Weights</th>
<th>Baseline Child Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control mean (SD)</td>
<td>Reduced form (SE)</td>
</tr>
<tr>
<td><strong>PPVT scores</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline PPVT, fall 2002</td>
<td>231</td>
<td>-0.003</td>
</tr>
<tr>
<td>(38)</td>
<td>(1.84)</td>
<td>(2.25)</td>
</tr>
<tr>
<td>PPVT, spring 2003</td>
<td>251</td>
<td>7.20***</td>
</tr>
<tr>
<td>(38)</td>
<td>(1.64)</td>
<td>(2.40)</td>
</tr>
<tr>
<td>PPVT, spring 2004</td>
<td>298</td>
<td>2.89</td>
</tr>
<tr>
<td>(40)</td>
<td>(1.81)</td>
<td>(2.60)</td>
</tr>
<tr>
<td>PPVT, spring 2005</td>
<td>339</td>
<td>0.21</td>
</tr>
<tr>
<td>(29)</td>
<td>(1.29)</td>
<td>(1.84)</td>
</tr>
<tr>
<td>PPVT, spring 2006</td>
<td>358</td>
<td>2.00</td>
</tr>
<tr>
<td>(30)</td>
<td>(1.42)</td>
<td>(2.07)</td>
</tr>
<tr>
<td><strong>PPVT missing or imputed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No baseline PPVT or imputed</td>
<td>0.222</td>
<td>0.031</td>
</tr>
<tr>
<td>(0.416)</td>
<td>(0.022)</td>
<td>(0.500)</td>
</tr>
<tr>
<td>No spring 2003 PPVT</td>
<td>0.190</td>
<td>-0.009</td>
</tr>
<tr>
<td>(0.393)</td>
<td>(0.023)</td>
<td>(0.414)</td>
</tr>
<tr>
<td>No spring 2004 PPVT</td>
<td>0.179</td>
<td>0.009</td>
</tr>
<tr>
<td>(0.383)</td>
<td>(0.022)</td>
<td>(0.409)</td>
</tr>
<tr>
<td>No spring 2005 PPVT</td>
<td>0.233</td>
<td>0.005</td>
</tr>
<tr>
<td>(0.423)</td>
<td>(0.022)</td>
<td>(0.424)</td>
</tr>
<tr>
<td>No spring 2006 PPVT</td>
<td>0.251</td>
<td>0.001</td>
</tr>
<tr>
<td>(0.434)</td>
<td>(0.022)</td>
<td>(0.436)</td>
</tr>
</tbody>
</table>

Notes:
Table reports means (columns 1 and 4), mean reduced form estimates (intent to treat) of the effect of being offered a Head Start slot (columns 2 and 5), and two-stage least squares (treatment effect on treated) estimates of the effect of attending Head Start in the Head Start Year, column 3) for the 3-year old cohort in the Head Start Impact Study by assessment period. The top panel reports the results for PPVT test scores and the bottom panel reports the results for indicators for the test score being missing (for 2002 this includes imputed values). Scores for 2002 include the imputed values. Column 1 has the control group means, using our inverse propensity score weights with the standard deviations in brackets. Column 2 has the reduced form effects (ITT) of being offered a Head Start slot using inverse propensity score weights with the standard errors in parentheses. Column 3 has the two-state least squares (treatment effect on the treated) estimates of the effect of attending Head Start during the Head Start year using inverse propensity score weights with the standard errors in parentheses. Column 4 has the control group means, using the baseline weights with the standard deviations in brackets. Column 5 has the reduced form effects (ITT) of being offered a Head Start slot using the baseline score weights with the standard errors in parentheses. SEs for differences allow for arbitrary correlation within center of random assignment. Propensity score weights estimated including baseline demographics and baseline test deciles by approximate month of assessment. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.
Table 5: Reduced form estimates of the effect of an offer of a Head Start slot, first stage estimates of the effect of an offer on administrative Head Start takeup and two-stage least squares estimates (TOT) of the effect of attending Head Start on PPVT by demographic subgroup for the Head Start year (spring 2003)

<table>
<thead>
<tr>
<th></th>
<th>Control mean [SD]</th>
<th>Reduced form (SE)</th>
<th>First stage (SE)</th>
<th>Two-stage least squares (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Race/ethnicity subgroups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>234</td>
<td>11.54***</td>
<td>0.673***</td>
<td>17.13***</td>
</tr>
<tr>
<td></td>
<td>[39]</td>
<td>(3.06)</td>
<td>(0.060)</td>
<td>(4.94)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>250</td>
<td>5.19**</td>
<td>0.694***</td>
<td>7.48*</td>
</tr>
<tr>
<td></td>
<td>[32]</td>
<td>(2.58)</td>
<td>(0.046)</td>
<td>(3.83)</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>268</td>
<td>6.49**</td>
<td>0.749***</td>
<td>8.66**</td>
</tr>
<tr>
<td></td>
<td>[34]</td>
<td>(3.14)</td>
<td>(0.033)</td>
<td>(4.23)</td>
</tr>
<tr>
<td><strong>Language subgroups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish speaker at home</td>
<td>223</td>
<td>15.0***</td>
<td>0.672***</td>
<td>22.37***</td>
</tr>
<tr>
<td></td>
<td>[32]</td>
<td>(3.30)</td>
<td>(0.057)</td>
<td>(5.65)</td>
</tr>
<tr>
<td>English speaker at home</td>
<td>261</td>
<td>4.93***</td>
<td>0.717***</td>
<td>6.87***</td>
</tr>
<tr>
<td></td>
<td>[34]</td>
<td>(1.86)</td>
<td>(0.029)</td>
<td>(2.62)</td>
</tr>
<tr>
<td><strong>Baseline PPVT score tercile subgroups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottom tercile, baseline PPVT</td>
<td>229</td>
<td>11.2***</td>
<td>0.787***</td>
<td>14.2***</td>
</tr>
<tr>
<td></td>
<td>[33]</td>
<td>(2.77)</td>
<td>(0.035)</td>
<td>(3.73)</td>
</tr>
<tr>
<td>Middle tercile, baseline PPVT</td>
<td>251</td>
<td>2.21</td>
<td>0.641***</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td>[30]</td>
<td>(2.45)</td>
<td>(0.047)</td>
<td>(3.84)</td>
</tr>
<tr>
<td>Top tercile, baseline PPVT</td>
<td>274</td>
<td>7.50**</td>
<td>0.674***</td>
<td>11.13**</td>
</tr>
<tr>
<td></td>
<td>[36]</td>
<td>(3.16)</td>
<td>(0.041)</td>
<td>(4.78)</td>
</tr>
</tbody>
</table>

*Notes*: Table reports control group means (column 1), reduced form effects of an offer of Head Start (column 2), first stage effects of an offer of Head Start on Head Start use in the Head Start year (column 3), and two-stage least squares estimates of the effect of attending Head Start in the Head Start year (column 4) for the 3-year old cohort in the Head Start Impact Study by various subgroups for the Head Start (spring 2003) assessment period. Each panel reports results for a set of mutually exclusive subgroups. Column 1 has the control group means, using the inverse propensity score weights with the standard deviations in brackets. Column 2 has the reduced form estimates of the effects of an offer of a Head Start slot on PPVT scores using inverse propensity score weights with the standard errors in parentheses. Column 3 has the first stage effects of an offer of a Head Start slot on use of Head Start during the Head Start year using inverse propensity score weights with the standard errors in parentheses. Column 4 has two-stage least squares estimates of the effect participating in Head Start during the Head Start year on PPVT using inverse propensity score weights with the standard errors in parentheses. Propensity score weights estimated including baseline demographics and baseline test deciles by approximate month of assessment. SEs for differences allow for arbitrary correlation within center of random assignment. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.
Table 6: Effect of Head Start offer on parent report of child care arrangements, by demographic subgroup for the Head Start year (spring 2003)

<table>
<thead>
<tr>
<th>Race/ethnicity subgroups</th>
<th>Head Start</th>
<th>Other Center</th>
<th>Parent/relative/Other care</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>First stage</td>
<td>0.636***</td>
<td>-0.170***</td>
</tr>
<tr>
<td></td>
<td>(SE)</td>
<td>(0.032)</td>
<td>(0.027)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.137</td>
<td>0.240</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.623</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>First stage</td>
<td>0.683***</td>
<td>-0.191***</td>
</tr>
<tr>
<td></td>
<td>(SE)</td>
<td>(0.027)</td>
<td>(0.026)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.178</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.554</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>First stage</td>
<td>0.712***</td>
<td>-0.189***</td>
</tr>
<tr>
<td></td>
<td>(SE)</td>
<td>(0.029)</td>
<td>(0.027)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.121</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.634</td>
<td></td>
</tr>
<tr>
<td>Language subgroups</td>
<td>First stage</td>
<td>0.641***</td>
<td>-0.191***</td>
</tr>
<tr>
<td>Spanish speaker at home</td>
<td>(SE)</td>
<td>(0.035)</td>
<td>(0.032)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.152</td>
<td>0.275</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.573</td>
<td></td>
</tr>
<tr>
<td>English speaker at home</td>
<td>First stage</td>
<td>0.690***</td>
<td>-0.181***</td>
</tr>
<tr>
<td></td>
<td>(SE)</td>
<td>(0.019)</td>
<td>(0.018)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.144</td>
<td>0.243</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.612</td>
<td></td>
</tr>
<tr>
<td>Baseline PPVT score tercile subgroups</td>
<td>First Stage</td>
<td>0.721***</td>
<td>-0.179***</td>
</tr>
<tr>
<td>Bottom tercile, baseline PPVT</td>
<td>(SE)</td>
<td>(0.028)</td>
<td>(0.024)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.123</td>
<td>0.219</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.658</td>
<td></td>
</tr>
<tr>
<td>Middle tercile, baseline PPVT</td>
<td>First Stage</td>
<td>0.638***</td>
<td>-0.158***</td>
</tr>
<tr>
<td></td>
<td>(SE)</td>
<td>(0.032)</td>
<td>(0.027)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.165</td>
<td>0.227</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.608</td>
<td></td>
</tr>
<tr>
<td>Top tercile, baseline PPVT</td>
<td>First Stage</td>
<td>0.656***</td>
<td>-0.190***</td>
</tr>
<tr>
<td></td>
<td>(SE)</td>
<td>(0.030)</td>
<td>(0.029)</td>
</tr>
<tr>
<td></td>
<td>Control mean</td>
<td>0.155</td>
<td>0.289</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.556</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Table reports the coefficient on treatment status for parent reports of the child using Head Start, another center, or some other source of care in the Head Start year using Head Start Impact Study data. The parent report is for the modal child care setting for the child. For each subgroup, the first row has the coefficient on the treatment group dummy for each of the three outcomes, the second row has the SEs, and the third row has the control group mean. The column 1 outcome is that the parent reported the child attended a Head Start center, the column 2 outcome is that the parent reported the child attended a non-Head Start center, and the column 3 outcome is that the child was at home with a parent or relative or attended a non-center family day care or did not report a child care setting. All regressions estimated with the propensity score weights. Propensity score weights estimated including baseline demographics and baseline test deciles by approximate month of assessment. SEs for differences allow for arbitrary correlation within center of random assignment. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.
Table 7: Two-stage least squares estimates of the effect of attending Head Start on socio-emotional outcomes for spring 2003 and 2006

<table>
<thead>
<tr>
<th>Parent reports</th>
<th>Head Start year</th>
<th>Grade 1 year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Spring 2003)</td>
<td>(Spring 2006)</td>
</tr>
<tr>
<td>Aggressive (ASPI)</td>
<td>-0.115*</td>
<td>-0.084</td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td>(0.078)</td>
</tr>
<tr>
<td>Hyperactive (ASPI)</td>
<td>-0.274***</td>
<td>-0.111</td>
</tr>
<tr>
<td></td>
<td>(0.074)</td>
<td>(0.074)</td>
</tr>
<tr>
<td>Lack of Social Competencies (ASPI)</td>
<td>0.022</td>
<td>-0.067</td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td>(0.079)</td>
</tr>
<tr>
<td>Lack of Social Skills (ASPI)</td>
<td>-0.032</td>
<td>-0.014</td>
</tr>
<tr>
<td></td>
<td>(0.075)</td>
<td>(0.076)</td>
</tr>
<tr>
<td>Withdrawn (ASPI)</td>
<td>0.027</td>
<td>-0.069</td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td>(0.076)</td>
</tr>
<tr>
<td>Conflict (Pianta)</td>
<td>-0.012</td>
<td>-0.169**</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
<td>(0.075)</td>
</tr>
<tr>
<td>Lack of closeness (Pianta)</td>
<td>-0.121*</td>
<td>-0.099</td>
</tr>
<tr>
<td></td>
<td>(0.064)</td>
<td>(0.076)</td>
</tr>
<tr>
<td>Lack of positive relationship (Pianta)</td>
<td>-0.048</td>
<td>-0.166**</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
<td>(0.078)</td>
</tr>
<tr>
<td>Externalizing behavior problems</td>
<td>-0.169**</td>
<td>-0.106</td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.072)</td>
</tr>
<tr>
<td>Teacher reports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggressive (ASPI)</td>
<td>-0.059</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td></td>
</tr>
<tr>
<td>Oppositional (ASPI)</td>
<td>-0.009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.083)</td>
<td></td>
</tr>
<tr>
<td>Inattentive (ASPI)</td>
<td>-0.074</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.079)</td>
<td></td>
</tr>
<tr>
<td>Shy/socially reticent (ASPI)</td>
<td>0.068</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.079)</td>
<td></td>
</tr>
<tr>
<td>Withdrawn/low energy (ASPI)</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.081)</td>
<td></td>
</tr>
<tr>
<td>Problems with structured learning (ASPI)</td>
<td>-0.021</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.082)</td>
<td></td>
</tr>
<tr>
<td>Problems with interactions (ASPI)</td>
<td>-0.019</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.077)</td>
<td></td>
</tr>
<tr>
<td>Combined ASPI index—negativity</td>
<td>-0.048</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.074)</td>
<td></td>
</tr>
<tr>
<td>Combined ASPI index—shy</td>
<td>0.057</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td></td>
</tr>
<tr>
<td>Combined ASPI index—interactive</td>
<td>-0.031</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td></td>
</tr>
<tr>
<td>Lack of closeness (Pianta)</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.087)</td>
<td></td>
</tr>
<tr>
<td>Lack of positive relationship (Pianta)</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.089)</td>
<td></td>
</tr>
<tr>
<td>Conflict (Pianta)</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.088)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Table reports two-stage least squares estimates of the effect of attending Head Start at age 3 on social-emotional outcomes. Outcomes reported for parent reports from 2003 and 2006 and teacher reports for 2006, using Head Start Impact Study data. Teacher reports are not available for all children until Kindergarten. Variables are standardized to first be “bad” outcomes, and then to have mean 0 and standard deviation 1 except for the combined ASPI indices. Propensity score weights estimated including baseline demographics and baseline test deciles by approximate month of assessment. SEs for differences allow for arbitrary correlation within center of random assignment. ∗, ∗∗, and ∗∗∗ denote significance at the 10%, 5%, and 1% levels. See text for more details.
Figure 1: QTE for PPVT scores at baseline for the 3-year old cohort, with 90% bootstrapped confidence intervals, propensity score weights.

Notes: Figure shows QTE for PPVT IRT score at baseline in Fall 2002 for the 3-year old cohort, using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center.
Figure 2: QTE and IV QTE for PPVT scores in spring 2003 for the 3-year old cohort, with 90% bootstrapped confidence intervals, propensity score weights

(a) QTE of effect of a Head Start Offer on PPVT scores

(b) IV-QTE of effect of Head Start on PPVT scores

Notes: Figure shows QTE for effect of an offer of a Head Start slot on PPVT IRT scores at end of the first year in spring 2003 (at the end of the Head Start year) for the 3-year old cohort in the top panel and IV-QTE for effects of Head Start participation during the Head Start year on PPVT scores at the end of the Head Start year, using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center.
Figure 3: IV-QTE for effect of Head Start participation during the Head Start year on various Woodcock Johnson-III tests as measured in the Spring of 2003, with 90% bootstrapped confidence intervals & propensity score weights

(a) IV-QTE estimates on WJIII Pre-Academic scaled score

(b) IV-QTE estimates on WJIII Applied Problems scaled score

(c) IV-QTE estimates on WJIII Letter-Word scaled score

(d) IV-QTE estimates on WJIII Spelling scaled score

Notes: Figure shows IV-QTE for the effect of Head Start participation during the Head Start year on various Woodcock Johnson III scaled scores for Spring 2003 for the 3-year old cohort, using the randomization as an instrument. The left top panel presents results for the Pre-Academic Composite score, the top left panel presents results for the Applied Problems score, the bottom left panel presents results for the Letter-Word score, and the bottom right presents results using the Spelling score, all using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center.
Figure 4: Reduced form effects of a Head Start offer on PPVT in 2003 by subgroup compared to administrative first stage, control group mean, and “first stage” on non-center care, with propensity score weights

(a) Reduced form vs. administrative first stage

(b) Reduced form vs. control group mean for non-center care

(c) Reduced form vs. “first stage” on non-center care

Notes: Figure plots reduced form effects of an offer of Head Start on PPVT for 2003 (the Head Start year) by subgroup (y-axis) versus the subgroup specific administrative first stage (4a), versus the control group mean by subgroup for non-center care (4b), and versus the “first stage” effect of an offer of Head Start on take-up of non-center care by subgroup. All estimates use the HSIS and propensity score weights.
Figure 5: Local linear regression of spring 2003 Head Start participation on 2002 baseline scores, for treatment and control groups

Notes: Figure shows local linear regression of parent-reported Head Start participation in spring 2003 on baseline PPVT test scores, using inverse propensity score weights.
Figure 6: Local linear estimates of treatment and control 2003 PPVT scores on baseline PPVT scores and of treatment-control differences

(a) Local linear estimates

(b) Treatment-control differences in local linear estimates

Notes: Figure in top panel shows local linear regression of 2003 PPVT on baseline PPVT separately for the treatment and control groups, using inverse propensity score weights. Figure in bottom panel shows treatment minus control differences of means in top figure from local linear regression of 2003 PPVT on baseline PPVT separately for the treatment and control groups, using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center.
Figure 7: Conditional and translated IV-QTE for PPVT scores in spring 2003 for the 3-year old cohort, by language

(a) Conditional IV-QTE of effect of HS on PPVT scores, by language

(b) Translated IV-QTE of effect of HS on PPVT scores, by language

Notes: Figure shows conditional IV-QTE (top panel) and translated IV-QTE (bottom panel) for effect of Head Start participation during the Head Start year on PPVT IRT scores at end of the first year in spring 2003 (at the end of the Head Start year) for the 3-year old cohort, by language, using inverse propensity score weights. Top panel estimates IV-QTE separately for English and Spanish speakers. Bottom panel presents “translated” IV-QTE, which position the IV-QTE from the top-panel at the relevant percentile of the unconditional control group distribution to which they correspond. 90% confidence intervals are obtained by bootstrapping by Head Start center.
Figure 8: Conditional and translated IV-QTE for 2003 PPVT scores, by race/ethnicity (non-Hispanic black and white and Hispanic, any race)

(a) Conditional IV-QTE of effect of HS on PPVT, by race

(b) Translated IV-QTE of effect of HS on PPVT scores, by race

Notes: Figure shows conditional IV-QTE (top panel) and translated IV-QTE (bottom panel) for effect of Head Start participation during the Head Start year on PPVT IRT scores at end of the first year in spring 2003 (at the end of the Head Start year) for the 3-year old cohort, by race/ethnicity (non-Hispanic black, non-Hispanic white, and Hispanics of any race), using inverse propensity score weights. Top panel estimates IV-QTE separately by race/ethnicity. Bottom panel presents “translated” IV-QTE, which position the IV-QTE from the top-panel at the relevant percentile of the unconditional control group distribution to which they correspond. 90% confidence intervals are obtained by bootstrapping by Head Start center.
Figure 9: Conditional and translated IV-QTE for PPVT scores for the 3-year old cohort, all years

(a) Conditional IV-QTE of effect of HS on PPVT scores, all years

(b) IV-QTE for PPVT scores for all years, translated

Notes: Figure shows conditional IV-QTE (top panel) and translated IV-QTE (bottom panel) for effect of Head Start participation during the Head Start year on PPVT IRT scores for the 3-year old cohort, for all years, using inverse propensity score weights. Top panel estimates IV-QTE separately for each year. Bottom panel presents “translated” IV-QTE, which position the IV-QTE from the top-panel at the absolute level of the PPVT score from the control group distribution to which they correspond.
Figure 10: IV QTE of Woodcock Johnson III Pre-Academic and Applied Problems tests for spring 2004

(a) Pre-Academic Skills

Notes: Figure shows IV-QTE for the effect of Head Start participation during the Head Start year on various Woodcock Johnson III scaled scores for Spring 2004 (Age 4 year) for the 3-year old cohort, using the randomization as an instrument. The left top panel presents results for the Pre-Academic Composite score, the top left panel presents results for the Applied Problems score, the bottom left presents results for the Letter-Word score, and the bottom right presents results using the Spelling score, all using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center.
Figure 11: Translated IV-QTE for PPVT scores in spring 2006 (first grade year) for the 3-year old cohort, by language and baseline score tercile

(a) Translated IV-QTE of effect of HS on PPVT scores, by language

(b) Translated IV-QTE of effect of HS on PPVT scores, by baseline score tercile

Notes: Figure shows translated IV-QTE by language (top panel) and baseline score tercile (bottom panel) for effect of Head Start participation during the Head Start year on PPVT IRT scores at end of the first grade year in spring 2006, for the 3-year old cohort, using inverse propensity score weights. Each panel presents “translated” IV-QTE, which position the conditional IV-QTE by subgroup at the relevant percentile of the unconditional control group distribution to which they correspond. 90% confidence intervals are obtained by bootstrapping by Head Start center.
### Appendix Table 1: Two-stage least squares estimates (TOT) of the effect of attending Head Start on Woodcock Johnson III outcomes for spring 2003 and 2006

<table>
<thead>
<tr>
<th></th>
<th>Head Start year (Spring 2003)</th>
<th>Grade 1 year (Spring 2006)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language, literacy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Letter Word</td>
<td>5.810***</td>
<td>0.226</td>
</tr>
<tr>
<td></td>
<td>(1.551)</td>
<td>(0.806)</td>
</tr>
<tr>
<td>Spelling</td>
<td>1.637***</td>
<td>-0.584</td>
</tr>
<tr>
<td></td>
<td>(0.574)</td>
<td>(0.887)</td>
</tr>
<tr>
<td><strong>Mathematics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applied Problems</td>
<td>3.414**</td>
<td>1.610</td>
</tr>
<tr>
<td></td>
<td>(0.050)</td>
<td>(1.185)</td>
</tr>
<tr>
<td><strong>Composite</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Academic Skills</td>
<td>3.966***</td>
<td>0.308</td>
</tr>
<tr>
<td></td>
<td>(1.113)</td>
<td>(1.270)</td>
</tr>
</tbody>
</table>

**Notes:** Table reports two-stage least squares estimates of the effect of attending Head Start at age 3 on Woodcock Johnson III outcomes. Outcomes reported for 2003 and 2006, using Head Start Impact Study data. Propensity score weights estimated including baseline demographics and baseline test deciles by approximate month of assessment. SEs for differences allow for arbitrary correlation within center of random assignment. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.
### Appendix Table 2: Effects of test scores in early and middle childhood on adult outcomes, children of the NLSY

<table>
<thead>
<tr>
<th></th>
<th>Repeat a Grade</th>
<th>Learning Any Poor HS HS grad. Some Pregnant As teen</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPVT at 3–5</td>
<td>-0.003**</td>
<td>0.00004 -0.002* -0.002 0.0004 0.002 0.001 -0.001</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.0005) (0.0008) (0.001) (0.0009) (0.001) (0.001) (0.001)</td>
</tr>
<tr>
<td>% effect</td>
<td>-0.009</td>
<td>0.001 -0.010 -0.007 -0.004 0.001 0.004 0.006 -0.004</td>
</tr>
<tr>
<td>PIAT at 9–11</td>
<td>-0.004***</td>
<td>-0.002*** 0.00001 0.0001 0.002** 0.002* 0.002** -0.001</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.0004) (0.001) (0.0009) (0.0007) (0.001) (0.001) (0.001)</td>
</tr>
<tr>
<td>% effect</td>
<td>-0.011</td>
<td>-0.038 -0.009 -0.015 -0.038 0.003 0.004 0.012 -0.018</td>
</tr>
<tr>
<td>Mean of dep. var</td>
<td>0.380</td>
<td>0.052 0.204 0.274 0.105 0.607 0.568 0.169 0.226</td>
</tr>
<tr>
<td>N</td>
<td>596</td>
<td>615 599 599 599 599 549 599 599 599</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.161</td>
<td>0.077 0.093 0.093 0.061 0.039 0.118 0.128 0.080 0.081</td>
</tr>
</tbody>
</table>

Notes: Table reports coefficients on test scores in the children of the NLSY 1979 sample of Deming (2009) but including singletons as well as siblings. Sample is further limited to those in the top quartile of predicted Head Start participation using observables. Results are weighted. Head Start participation while age 3–5 is predicted using demographics, family characteristics, mother’s AFQT score, and birth order. The PPVT score is measured at ages 3–5, and the PIAT score is measured at ages 9–11. Outcomes are measured when children are young adults. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.