Experimental Evidence on Distributional Effects of Head Start

Marianne P. Bitler, UC Irvine and NBER

Hilary W. Hoynes, UC Berkeley and NBER

Thurston Domina, UC Irvine*

August 21, 2014

Abstract

This study provides the first comprehensive analysis of the distributional effects of Head Start, using the first national randomized experiment of the Head Start program (the Head Start Impact Study). We examine program 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 pre-school period and that the gains are largest at the bottom of the distribution. Once the children enter elementary school, the cognitive gains fade out for the full population, but importantly, cognitive gains persist through 1st grade for some Spanish speakers. These results provide strong evidence in favor of a compensatory model of the educational process. Additionally, our findings of large effects at the bottom are consistent with an interpretation that the relatively large gains in the well-studied Perry Preschool Program are in part due to the low baseline skills in the Perry study population. We find no evidence that the counterfactual care setting plays a large role in explaining the differences between the HSIS and Perry findings.

*Correspondence to Bitler at mbitler@uci.edu; Hoynes at hoynes@berkeley.edu; or Domina at tdomina@uci.edu. This research was 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 thank Child Care & Early Education Research Connections for providing access to the experimental data used here. We also thank Josh Angrist, Peg Burchinal, Liz Cascio, Janet Currie, Greg Duncan, George Farkas, Avi Feller, Alex Gelber, Adam Isen, Doug Miller, Luke Miratrix, Jesse Rothstein, Diane Schanzenbach, Gary Solon, 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.

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 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 federal operating budget of nearly $7 billion with additional expenditures by state and local governments (US DHHS, Office of Head Start 2008). Despite the recent growth of state public pre-K programs, Head Start remains a central early childhood education provider for low-income families (Cascio & Schanzenbach (2013)). Additionally, educational inequities—as measured by preschool enrollment—still exist across the income distribution (Duncan & Magnuson (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 from economics, neurobiology, and developmental psychology on the key role that early experiences play in later development. They conclude that early in life is the most promising period for investments in disadvantaged children, since such investments 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 large positive effects on participants' academic achievement and attainment (Schweinhart, Barnes & Weikart (1993); Campbell & Ramey (1995); Barnett (1996); Currie (2001); Heckman, Moon, Pinto, Savelyev & Yavitz (2010)).

In this study, we provide the first comprehensive analysis of the distributional effects of Head Start. We examine the program's 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 experimental evaluation of Head Start. This work adds to a growing body of evidence on the distributional effects of educational interventions (e.g., Neal & Schanzenbach (2010); Heckman, Pinto & Savelyev (2010); Angrist, Dynarski, Kane, Pathak & Walters (2012); Felts & Page (2013)).

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 (Magnuson, 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 towards remedying the skills' deficits that often disadvantage poor students at the beginning of elementary school (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 the largest gains at the bottom of the skill distribution) against the 'skills-begets-skills' hypothesis (predicting the largest gains at the top of the skill distribution), within the context of the Head Start applicant population.

Our work builds on decades of research documenting that Head Start generates important and statistically significant impacts on cognitive outcomes (e.g., see Currie & Thomas (1995) and the review in Currie (2001)). Recent evidence using a wide range of quasi-experimental methods further demonstrates 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, Currie & Thomas (2002); Ludwig & Miller (2007); Deming (2009); and Carneiro & Ginja (Forthcoming)). While it is clear that Head Start 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 multiply across later periods of youth development.

The Head Start Impact Study, mandated by Congress in 1998, was 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 followed nearly 5,000 children in two cohorts that were three or four years old at the time of Head Start application through third grade, collecting detailed outcomes on academic and social-emotional measures.

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. Our main results use an instrumental variables approach, where the first stage outcome is Head Start participation and the instrument is the offer of a Head Start slot. This allows us 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 examine social-emotional outcomes; for example, relationship measures using the Pianta scales and measures of child behavior using the Adjustment Scales for Preschool Intervention (ASPI) and other parent and teacher reports.

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, as well as for Woodcock Johnson III measures of early mathematics and early literacy.[1] Generally, we find that the gains are largest at the bottom of the distributions of each of

[1] Note that the PPVT and WJIII achievement tests have been shown to be correlated with IQ measures in many settings and for many age groups (e.g., Caravajal (1988), Campbell, Bell & Keith (2001), Schrank, Becker & Decker (2001).

these measures of 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 deviations 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. We go on to document that in large part, these differences across groups largely reflect differences in where in the skill distribution these subgroups are concentrated.

Importantly, our results also speak to a possible explanation for the large short run cognitive gains found in the Perry experiment. The Perry program was considerably more intensive than Head Start, providing daily classes and weekly home visits. While we can not assess the consequences of this programmatic difference, we can assess two additional explanations for the gap between the large and positive effects of Perry (and other model programs of the 1960s) and the smaller and less lasting effects of Head Start. First, the Perry program (like the other model programs) was targeted at a very disadvantaged population. Our results of large effects of Head Start at the bottom of the distribution indicate that in a modern setting, we also identify very large gains at low achievement levels, suggesting some consonance between these earlier findings and the more current ones. Second, the counterfactual care setting for the control groups in the early 1960s was likely quite different from the current options for potential HS children. For example, Perry control group members had little access to any center based care; in contrast, currently, we are seeing the rapid growth of state means-tested early childhood education programs (Cascio & Schanzenbach (2013), Duncan & Magnuson (2013)). Indeed, 25% of the control group children in the HSIS attended a non-Head Start child care center. The HSIS experiment, then, may understate the effects of Head Start compared with 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 non-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.

Overall our 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 and the effects we do find are fairly constant across the distribution.

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 short run 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.[2]

More recently, focus has turned to longer-term impacts of Head Start. 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. 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. Deming

[2] Aizer & Cunha (2012), also using a sibling comparison approach, analyze Head Start during the roll-out of the program in the mid 1960s. An alternative approach used by Griffen (2014) uses observational data but a more structural approach.

(2009) uses a family fixed effect approach and finds that Head Start 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 and find that the program leads to reductions in behavioral problems and health improvements in adolescence and reduces crime and idleness in young adulthood.

Long-term Head Start effects on human capital outcomes are difficult to reconcile with evidence of fade-out in cognitive domains. One potential explanation is that Head Start influences parents and their investments in children (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.

Overall, 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) and Bloom & Weiland (2013) also examine heterogeneity in effects across program sites, suggesting some role for center inputs. Learning more about the heterogeneity of Head Start is the focus of our work. We combine a distributional approach with an eye on various subgroups and where in the counterfactual distribution they are located, bringing new insight into where the gains are concentrated.

2.2 Theoretical Expectations and Heterogeneity Elsewhere in Social 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 experiment 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 a given 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

(Magnuson 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 in the 1998 HS 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.[3]

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. Heckman, Pinto & Savelyev (2010) provide supportive evidence in favor of this theory, finding that the Perry Preschool program led to gains at the top of the distribution of cognitive achievement for girls.

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.[4]

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

[4] 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 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.

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.[5] 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.[6]

The HSIS consists of two age cohorts: 3-year olds and 4-year olds.[7] The experiment was intended to determine the effects of being exposed to one year of Head Start. 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. Indeed many control children also attended HS at age 4 while some treatment children left Head Start programs. Most children in the 4-year old cohort would have been expected to transition to Kindergarten in year two (after the one year of HS treatment).

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 data including outcomes through the end of first grade.

[5] The HSIS is representative of oversubscribed centers in the U.S. Puma et al. (2010) reports that such centers account for about 85% of the children enrolled in Head Start. 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.

[6] Severely disabled children were excluded from randomization.

[7] Note that more precisely, the cohorts were individuals eligible for 2-years or 1-year of Head Start before Kindergarten, we use 3-year olds and 4-year olds as shorthand for this longer description.

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 in the short or medium term; most such impacts are small and statistically insignificant.[8]

As is common for this type of intervention, the HSIS 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 in the experimental year, while about 14 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), the mean intent to treat (ITT) or reduced form, and the treatment effect on the treated (TOT) or instrumental variables results below.

3.2 Sample, Means, and Tests of Balance

We limit our analysis to 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 HS serves children during the two years prior to

[8] Any 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.

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 had not participated in HS at age 3. As might be expected, we find the 4-year old cohort to be potentially more disadvantaged compared with the 3-year old cohort, with a higher share of children living in Spanish speaking households and living with lower educated mothers. Additionally, we focus on the 3 year cohort because increasingly many 4-year olds have 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.[9]

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" or first Head Start year which corresponds to the 2002–03 academic year, which is followed by the "Age-4 year" or second Head Start year (2003–04), the Kindergarten year (2004–05), and the 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.

Appendix 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, we use the Peabody Picture Vocabulary Test (PPVT) of vocabulary knowledge and receptive language[10] and the Woodcock Johnson III measure of Pre-Academic

[9] 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 account for baseline test scores and demographics. More detail on this is given below.

[10] 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 Vocabilario en Imagenes Peabody. A very

Skills, which is a composite of three components measuring language, early 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, 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 (e.g., Romano, Babchishin, Pagani & Kohen (2010)) 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 (Duncan, Dowsett, Claessens, Magnuson, Huston, Klebanov, Pagani, Feinstein, Engel, Brooks-Gunn, Sexton, Duckworth & Japel (2007)).[11] 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 have 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 the Adjustment Scales for Preschool Intervention (ASPI), as well as other teacher and parent reports. Each of these measures are indices constructed from a series of questions about the child posed to the teacher or parent, counting affirmative answers (ASPI) or counting responses on a five-point Likert scale (Pianta) or, in some cases, aggregating parent reports of children's behavior.

Table 1 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.[12] 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

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.

[11] Some of the other WJIII tests included on the survey are not very continuously distributed and are thus perhaps less suitable for our distributional methods.

[12] 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 where there was no valid imputed or actual PPVT at baseline because the child could not take the test in English. We discuss the propensity score weighting in the next section.

who were never married, and 21% are in medium or high risk families.[13] 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).

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 1 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.[14]

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 and higher share of high 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).

[13] 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.

[14] 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.

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. The 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," although it is not necessary with experimental data (e.g., Krueger & Zhu (2004)). Furthermore, we want to control for other baseline observables to adjust for the (presumed relatively minor) imbalances in the demographic characteristics between the treatment and control groups as well as the differences in baseline 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 (here enrollment in Head Start). 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 $\widehat{p}$, the predicted probability of being in the treatment group as a function of baseline characteristics using a logit, estimated using the child baseline weight.[15] We then weight each observation by its inverse propensity-score weight ($1/\hat{p}$ if in the treatment group, and $1/(1-\hat{p})$ if a control observation).

In our logit model, we control for all of the child and caregiver variables in Table 1. Additionally, to richly control for baseline test scores, we assign dummies for decile of the baseline (2002) PPVT, separately for the four assessment month groupings in Table 1. One of the assessment-month groups is "missing or imputed PPVT," allowing us to use and account for the observations with imputed baseline values. Thus, we are also 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.[16] 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 treatments and controls, they look extremely similar with trivial non-overlap.[17] The final column of Table 1 presents the "adjusted" difference in means between the treatments and controls. 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 28 total controls (25 demographic variables and 3 timing variables) 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 2 provides detailed

[15] Zanutto (2006) and Dolton & Smith (2011) discuss use of survey weights with propensity-score weighting estimation.

[16] 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.

[17] 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.

information on child care settings, separately for the treatment and control groups (as well as for their 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 during the Head Start year compared to 15% of the control group children. Thus the experimentally manipulated offer of a Head Start slot led to a highly statistically significant 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 2 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 about 60% are in family day care or being cared for by a parent or relative, with the vast bulk being with a parent or relative. As discussed by Duncan & Magnuson (2013), understanding the counterfactual child care environment is important for providing context for the results and expectations for the effects of the intervention. The offer of a Head Start slot mostly displaces use of 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.[18]

In the bottom of the table, we provide similar tabulations for parent reports of child care use (or Kindergarten enrollment) 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 significant 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.

[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).

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 is nearly 3 percentage points. Much of the change in the treatment-control difference in child care settings between the Head Start year and the Age-4 year comes from changes in child care setting for the control group. Among control children, 14.6% are in Head Start in the HS year compared to 47.3% in the Age-4 year. Many fewer control children are in parent or relative care during the Age-4 Year: 53.6% of the control children are in parent/relative care in the Head-Start year as compared to 10.3% in the Age-4 year. For treatment children, most of the change is an increase in use of other center-based care (and a reduction in use of 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 of 49 percentage points in the probability of using center care in the Head Start year. It is important to note that any reduction in the magnitude of the first stage would lead to larger TOT effects from a Wald estimate. In that case, our results provide a lower bound for the estimated effects of use of center care in the HSIS.

Before examining the impacts of HS across the distribution of outcomes, we first present mean treatment effects for our main cognitive outcome, PPVT, and also show we achieve balance in attrition with our inverse p-score weights. Table 3 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—reported in the top panel—and the probability that the PPVT test score is missing (or for the baseline 2002 period only, imputed or missing)—reported in the bottom panel.[19] Columns 1–3 report the control mean, the mean intent to treat effect (reduced form), and the two stage least squares estimates; each estimated while balancing baseline characteristics (e.g., using inverse propensity score weights). For reference, in columns 4 and 5 we present the control group mean and treatment-control difference (intent to treat) using only the baseline child weights.

The first row of each panel examines the baseline PPVT, providing another balance test. The table shows that the mean of PPVT at baseline was 231 with a standard deviation of 38 (both measures calculated for the control group). There are at most trivial differences, on average, in the baseline scores across the treatment and control groups, either with (column 2) or without (column 5) controls. (Recall the controlling is done with the weights.) This is, of course, important as it indicates balance in the randomization on key baseline measures. Without the use of inverse propensity-score weights, we do find 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 test scores. However, after inverse propensity-score weights are applied, the probability of a missing baseline score is balanced (the difference is a statistically insignificant 0.03 after adjustment). In fact, our inverse propensity-score weights balance the PPVT non-response in each later year from 2003–2006 as well, with small and statistically insignificant differences in non-response in those years.

Next we turn to mean treatment effects (reduced form effects) in Spring 2003 after the experimental year. 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, reflecting the first stage of about 0.70.

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

[19] The PPVT and one Woodcock Johnson III measure were imputed for almost all children missing them at baseline. No imputations were made for scores missing in spring 2003 or later.

to a increase in PPVT of 4.2 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 is 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} (here, outcomes are test scores or indices 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, $\delta = E[\delta_i] = E[Y_1] - E[Y_0]$. 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 effects (QTE), and ultimately a particular implementation of instrumental variables quantile treatment effects (IV-QTE) which we describe in turn. The QTE are simply the distributional analog of the average treatment effect. If $F(y)$ is the cumulative distribution function (CDF) of Y, then the qth 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 qth quantiles of these two distributions $y_q = y_{q1} - y_{q0}$, where y_{qd} is the qth 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, $y_q = y_{q1} - y_{q0}$. 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.[20] 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 *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 affects the distribution of cognitive achievement. In this context we can test to what extent the data are consistent with the compensatory or skill-begets-skills theories of educational development. An alternative approach would be to use baseline skills (in our context, 2002 PPVT) as the measure of underlying skill and to explore how the treatment varies across this spectrum. A third approach would be to estimate how Head Start affects the value added in test scores from 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) for a couple of reasons. First, as discussed above, not all children were assessed at baseline, and there is imbalance in the month of assessment and the rate of missing PPVT across the treatment and control groups. Second, 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 treatment. In short, the baseline test is not ideal in this context.

5.2 IV-QTE

While it is of interest to look at the directly policy relevant outcome "What would happen if we expanded the number of Head Start slots at oversubscribed centers?," it is also of interest to know the underlying parameter, the effect for a child of attending Head Start.

[20] 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); hence the term unconditional QTE.

Thus, instead of simply looking at the reduced form effects of treatment assignment, we also 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 of a slot 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) who develop 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 taking up the treatment across the distribution for compliers.[21] 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 were sampled in a complicated fashion. For inference, we bootstrap, with centers being randomly sampled with replacement (with all children from sampled centers appearing in the data whenever a center is sampled). We construct confidence intervals using the percentile method. With 999 bootstrap resamples, a 90% confidence interval is given by sorting the resulting 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

[21] We make use of Frolich & Melly's (Forthcoming)'s software, using our own estimate of the predicted probability of being in the treatment group (propensity score) 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.

IV estimates.

6 Main Results for the Heterogeneous Effects of HSIS

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

Figure 1a presents the QTE for the Head Start year, using the PPVT test administered in Spring 2003. 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. 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 3).[22] This graph being of the QTE (and not the IV-QTE), the results show the impact of an offer of Head Start, the reduced form or ITT effects. There are several things to note from these results. First, the offer of a HS slot improves cognitive skills (PPVT) 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 the lower end of the skill distribution are very large, ranging from around 30 for quantiles 1–3, to 17 for quantile 13, to 3 at quantile 43 (which is the last QTE which is statistically significantly different from zero at the 10% level). These effects are substantial relative to the control group standard deviation of 38 (see column 1 of Table 3), implying effect sizes upwards of 0.8 SD at the bottom of the distribution to 0.08 SD at the median. (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 control-group standard deviation of the outcome plotted.)

The QTE capture the impact of the treatment on the distribution of outcomes. One limitation of the QTE estimator is that the 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

[22] We also provide one additional balance test. In particular, in Appendix Figure 1, we report the QTE for the baseline PPVT (2002). Notably, the confidence intervals all include 0; thus, we have balance across the two groups in the baseline test score distribution. (In results not shown, there is only modest imbalance in the baseline QTE before controlling for observables using the inverse propensity score weights.)

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 1b 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., the 2SLS estimate of 10.2 test score points from Table 3). 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 quite 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 our presentation of IV-QTE, examining the impacts for the WJIII Pre-academic Skills composite and its components (Applied Problems, Letter-Word, and Spelling). These tests are complementary to the PPVT (a test of receptive vocabulary), with one major advantage being the ability to examine the Applied Problems test, which reflects early numeracy. The other components encompass pre-reading, letter-word recognition, early writing, and spelling. The WJIII results are presented in Figures 2a–2d. Starting with Figure 2a, 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 effects losing statistical significance beginning around the 40th percentile. Notably, though, there is somewhat less variation in the treatment effects across the distribution compared with the PPVT, and the confidence intervals are wider. Interestingly, when we separately examine the three test components of the Pre-

[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.

Academic Skills test, we see that the effects on early numeracy (Applied Problems, Figure 2b) 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 2c) show evidence of positive effects in the middle and top of the distribution (and some non-negative effects at the bottom). The WJIII Spelling test (Figure 2d) shows little variability across the distribution, with some marginally significant positive effects. Overall the WJIII tests findings largely confirm the results using the PPVT—the evidence is in favor of a compensatory effect of HS for those at the bottom of each test scale distribution. In addition, having a measure of the impact of HS on early numeracy (Applied Problems) is particularly useful given Duncan et al.'s (2007) results showing the importance of early numeracy in predicting long term achievement. We do note, however, that the test components capturing literacy (Letter-Word, and to a lesser extent, Spelling) provide some evidence in favor of the skills-beget-skills hypothesis. (We further remind the reader that the bottom of the various test score distributions do not have to be the same individuals, although they may be.)

6.2 Results for Subgroups in the Head Start Year

Another dimension of heterogeneity is captured by examining 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 the baseline PPVT score. Reduced-form estimates for these subgroups are in the second column of Table 4 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 0.34 standard deviations) than for the other subgroup members. 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). 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 care setting. The third column of Table 4 presents the administrative first stage for the various subgroups. It is notable that the magnitude of the first stage does not always track the magnitude of the reduced form; for example Hispanics have lower take-up rates than do non-Hispanic blacks and whites.

Figure 3a displays this information about the first stage and reduced forms by group more compactly for these and other subgroups. Figure 3a is a scatter plot where the administrative first stage is on the x-axis and the reduced form mean treatment effect (ITT) is on the y-axis. Each dot represents a different subgroup, including as above race/ethnicity, Spanish speaker and terciles of baseline distribution plus gender and mother's education, with the size of the dots reflecting the the sum of inverse propensity score weights accruing to the subgroup. 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 do not have larger first stages than do other subgroup members. The first stage effect of an offer of Head Start on Head Start take-up is higher for females compared with males, but their reduced-form effects are similar.

Figure 3b explores the role of the counterfactual care setting with a scatter plot identical to Figure 3a, except that the x-axis now plots the control group mean for use of non-center care (family day care, parent, relative or unknown). Figure 3c 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. Echoing the results above, overall the HSIS data show little relationship between the counterfactual care settings and the reduced form PPVT effects across subgroups.[24]

We further explore this using the approach of Abadie, Chingos & West (2014) by creating subgroups based on the propensity to use non-center care. Of course the problem is that use of non-center care is endogenous (we do not have a "pre" measure of this variable). Following Abadie et al. (2014), we predict non-center propensity for the treatment group members, using coefficients estimated based on the control group. Non-center propensity is predicted similarly for the control group, except we use a "leave one out" approach (dropping the

[24] See Appendix Table 2 for the point estimates behind these two scatter plots.

own control observation from the regression used to predict its value). (The model includes the same covariates we used to predict the propensity scores.) We predict hexiles of the predicted non-center care counterfactual, and produce scatterplots similar to Figure 3, with the reduced form (ITT) effects on the y-axis and the first stage (Figure 4a), control group non-center mean (Figure 4b) and alternative first stage for non-center care (Figure 4c). (We also include a fitted line.) The results echo the findings from Figure 3—we find no evidence that differences in the magnitude of the mean treatment effects across the predicted hexiles of using non-center care are explained by differences in the take-up of Head Start (the first stage) or counterfactual care setting.[25]

Returning to the subgroups in Figure 3 (and Table 4), it is notable that children in the bottom tercile of the baseline PPVT score show evidence of a larger first stage (79% compared with the full sample first stage of 70.5%), a larger reduced form effect (11.2 compared with the full sample effect of 7.2), and high use of non-center care among the controls. To explore the differences across baseline skills further, we take a less parametric approach. In Figure 5a, we examine how the first stage varies with the baseline PPVT score by estimating local linear regressions (by baseline PPVT) of parent-reported HS use in spring of the Head Start year, separately for the treatment and control groups. When seen across the entire distribution of baseline PPVT, the figure shows remarkable uniformity in the first stage effect of an offer of a slot 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 5b, following Duflo, Dupas & Kremer (2011), we estimate local linear regressions of the mean effect on PPVT at the end of the Head Start year, as a function of the baseline PPVT, estimated separately for the treatment and control groups and in Figure 5c, we plot the treatment-minus-control differences from these local linear regressions (along with the bootstrapped confidence intervals).[26] These estimates as a function of baseline PPVT (another version of a subgroup estimate) show large gains at the bottom of the baseline PPVT skills distribution

[25] An alternative explanation of course is that the selection into counterfactual care choice is based on unobservables which are orthogonal to these observables used to predict non-center care use.

[26] The bandwidth here is 1/20 the range of scores, and we use a rectangular kernel. Results are robust to deviations from this bandwidth.

with treatment group assignment, with some evidence of a positive effect near the top third 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 1a). We prefer the QTE (and IV-QTE) for several reasons. First, note that we only have baseline scores for the full sample for PPVT (most of the WJIII tests were not administered at baseline to Spanish speakers). Second, we worry about 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). Further, and more importantly, for many policy interventions such as Head Start, knowing about effects on the (ex-post) distribution—the QTE or IV-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).

Overall, this analysis suggests that the reduced-form effects of the HSIS vary across demographic subgroups and by baseline test score. We find little evidence that this is explained by differences in either program take-up or counterfactual care settings. Returning to Table 4, 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%. 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, compared to others.

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 6a. (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—they are above 20 or 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 4). 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 6b.[27] These translated IV-QTE stretch out the conditional QTE by a different amount depending how uniformly the subgroup in question is laid out across the overall unconditional distribution in the control group. To give some guidance on the amount of stretching this produces, in the translated graphs, there is a symbol at each decile of the subgroup's own (conditional) distribution. Thus, one can see that in the lower half of the (overall unconditional) PPVT distribution, the symbols for deciles for the Spanish speakers are more compressed while in the upper half the opposite is true. Figure 6b shows that once the subgroups' IV-QTE are put on the same absolute scale, the differences between groups become greatly attenuated. We see similarly-sized, very large gains throughout the bottom deciles 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. However, this divergence at the top of the Spanish and English speaking distributions is driven by a small share of the Spanish speaking distribution, and the differences are smaller where the overlap is more substantive.

Figures 7a and 7b provide similar analyses by race/ethnicity; for non-Hispanic whites, non-Hispanic blacks, and Hispanics. The results are quite dramatic—while the conditional

[27] Essentially what this is doing is stretching the conditional distributions in some places and shrinking 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 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 to the full range of the x-axis, indicating the relative lack of scores in some parts of the pooled distribution for each subgroup.

IV-QTE (Figure 7a) show widespread gains for Hispanics (relative to the two non-Hispanic groups), on a common scale, the translated IV-QTE (Figure 7b) shows much more similarity in effects across race/ethnicity groups throughout the distribution. Thus, by race, there is little heterogeneity in effects across groups given that we are looking at a common quantile in the overall control group distribution. The variation in the conditional IV-QTE is driven more by the differences in where the bulk of the subgroups are located within the control group. To put it more simply, the heterogeneity by race appears to be all driven by where in the overall control group the race groups are situated and not by heterogeneous effects given a control percentile.

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.[28]

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 8a 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 the effect of an offer 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 the preschool years (end of 2004). However, once this cohort transitions into elementary school, the control group catches up, and the differences substantially fade.

As with the demographic subgroups, the x-axis scale for each of the years corresponds

[28] 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).

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 8b; the translated IV-QTE show much more alignment in the effects across years at given PPVT score values in the control group. The results suggest that Head Start brings children up to some level but once they are achieving at that level there are no additional gains to be had.

To highlight further the effects in later preschool, Appendix Figures 2a and 2b present results for the end of the 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 around the 80th percentile 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 9a presents the translated IV-QTE for PPVT for Spanish versus English speakers for the first grade year. Notably, 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 quite large, measuring 0.38 to 0.58 standard deviations. Additionally, Figure 9b presents the translated IV-QTE for PPVT 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 do not persist though first grade.

7 Discussion

Our analysis of the cognitive effects of the Head Start Impact Study shows 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 than for others.

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, both when measured by PPVT and when measured by WJIII achievement tests. 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, Grindal, Miratrix & Page (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, language, and a host of others including center of random assignment) do not explain who is in which group of compliers.

Our work also sheds light on the large effects of 1960s intensive interventions, such as the Perry Preschool Program, relative to Head Start. Perry targeted 123 Black children in Ypsilanti, Michigan and randomly assigned them either to a intensive pre-school or to the control condition. The Perry treatment led to an increase of 12 points in mean IQ after 1 or 2 years of treatment and 6 points after the end of Kindergarten (Schweinhart & Weikart (1981)).[29] These are large effects (i.e., a 0.8 effect size) but are not dissimilar from the large effect sizes we find for HSIS at the bottom of the distribution. Further, the Perry children were targeted based on parents' low completed education and occupational status as well as for being of low IQ—the sample average IQ was 79 and participants had scores 1–2 standard deviations below the population mean (Schweinhart & Weikart (1981)). Thus, our results show that HS can generate meaningful increases in cognitive outcomes and suggest that the large Perry effects may derive, in part, from the study participant's low baseline skills being well suited to the intervention.

Taking into account these findings, along with the fadeout in cognitive gains, we suggest that the gains in preschool may not persist if the elementary schooling environment is not of

[29] Take-up in Perry was nearly 100% so the ITT are also TOT estimates.

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 can not test directly for this, given that we do not observe much about the schools the HSIS participants attend. However, if Head Start teachers teach to some (low) proficiency standards (e.g., knowing the ABCs, counting to 10), then the HS setting may be insufficient to yield 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.

These large and persistent Head Start effects on students with limited English skills may also help to reconcile long term human capital effects with evidence of Head Start cognitive effect fade out in early elementary school. It would be interesting to know if the experience of these groups can account for the positive long term outcomes others have found.

Additionally, 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—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). We estimated IV models for 9 parent-reported measures in the Head Start year and 9 parent-reported and 14 teacher-reported measures in Grade 1.[30] We standardize the social-emotional measures to have a mean of 0 and standard deviation of 1 (using the control group mean and SD) and to be aligned so a higher score is always worse than a lower score. As shown in Appendix Table 3, in the Head Start year, the IV estimates are negative—implying an improvement in parent-reported non-cognitive outcomes—for seven of the nine non-cognitive outcomes. However,

[30] 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.

only 2 of the 9 are statistically significant at the 5% level (significance found for hyperactivity and externalizing behavior). The first grade effects are uniformly small and only 2 of the 23 outcomes show statistically significant improvement (at the 5% level), and both only for the parent-reported measures. In sum, while there are a few positive findings, the overall result 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.

8 Conclusion

In this study, we provide the first comprehensive analysis of the distributional effects of Head Start. We use data from the Head Start Impact Study, 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. We find that the gains are largest at the bottom of the distribution of achievement, with treatment on the treated effects sizes upwards of a full standard deviation at the lowest achievement levels. 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 various subgroups of the population including ones defined by race/ethnicity, by the child speaking Spanish or English, by baseline levels of cognitive skills, by gender, and by mother's education. 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. We find little role for differences in Head Start take-up or the underlying counterfactual child care setting in explaining these heterogeneous effects. Importantly, we also find that for Spanish speakers, the cognitive gains persist through 1st grade.

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 achievement. This is revealed using quantile treatment effects and IV-QTE applied to the ex-post distribution of achievement, local linear regression estimates based on baseline skills, and analyses across demographic groups. There is some limited evidence of the effects predicted by "skills beget skills" theories.

These findings serve to shed light on a potential source for the relatively large short run cognitive gains in the Perry experiment. Previous explanations include the intensity of the Perry intervention and/or the counterfactual care environment that operated at that time. The Perry program also targeted very low IQ children; our results of large effects at the bottom of the distribution indicate that in a modern setting we also identify very large gains at low achievement.

References

Abadie, A. (2002), 'Bootstrap tests for distributional treatment effects in instrumental variable models', *Journal of the American Statistical Association* **97**, 284–92.

Abadie, A., Angrist, J. D. & Imbens, G. (2002), 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings', *Econometrica* **70**(1), 91–117.

Abadie, A., Chingos, M. & West, M. (2014), Endogenous stratification in randomized experiments, Working Paper 19742, NBER.

Aizer, A. & Cunha, F. (2012), The production of child human capital: Endowments, investments, and fertility, Working paper.

Angrist, J., Dynarski, S., Kane, T., Pathak, P. & Walters, C. (2012), 'Who benefits from KIPP?', *Journal of Policy Analysis and Management* **31**(4), 837–860.

Barnett, W. (1996), *Lives in the Balance: Age 27 cost-benefit Analysis of the High/Scope Perry Preschool*, High/Scope Press, Ypsilantim MI.

Bloom, H. & Weiland, C. (2013), Moving beyond average impacts: Do Head Start's impacts on children's language, literacy, and math skills vary across individuals, subgroups, and centers?, Working paper.

Campbell, F. & Ramey, C. (1995), 'Cognitive and school outcomes for high-risk African American students at middle adolescence: Positive effects of early intervention', *American Educational Research Journal* **32**(4), 743–772.

Campbell, J., Bell, S. & Keith, L. (2001), 'Concurrent validity of the Peabody Picture Vocabulary Test-Third edition as an intelligence and achievement screener for low-SES African American children', *Assessment* **8**(1), 85–94.

Caravajal, H. (1988), 'Relationship between scores of gifted children on Stanford-Binet IV and Peabody Picture Vocabulary Test Revised', *Assessment for Effective Intervention* **14**(22), 22–25.

Carneiro, P. & Ginja, R. (Forthcoming), 'Long-term impacts of compensatory pre-school on health and behavior: Evidence from Head Start', *AEJ Applied Economics* .

Cascio, E. & Schanzenbach, D. (2013), 'The impacts of expanding access to high-quality preschool education', *Brookings Papers on Economic Activity* pp. 127–179.

Chernozhukov, V. & Hansen, C. (2005), 'An IV model of quantile treatment effects', *Econometrica* **73**(1), 245–261.

Cunha, F. & Heckman, J. J. (2010), Investing in our young people, Working paper, NBER. W16201.

Cunha, F., Heckman, J., Lochner, L. & Masterov, D. (2006), Interpreting the evidence on life-cycle skill formation, *in* E. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education, Volume 1', Elsevier, pp. 697–812.

Currie, J. (2001), 'Early childhood programs', *Journal of Economic Perspectives* **15**(2), 213–238.

Currie, J. & Thomas, D. (1995), 'Does Head Start make a difference?', *American Economic Review* **85**(3), 341–364.

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

DHHS ACF (2000), 'Curriculum in Head Start: Head Start Bulletin 67'.

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'.

Dolton, P. & Smith, J. (2011), The impact of the UK New Deal for lone parents on benefit receipt, Working Paper 5491, IZA.

Duflo, E., Dupas, P. & Kremer, M. (2011), 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya', *American Economic Review* **101**(5), 1739–74.

Duncan, G., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A., Klebanov, P., Pagani, L., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K. & Japel, C. (2007), 'School readiness and later achievement', *Developmental Psychology* **43**(6), 1428–46.

Duncan, G. & Magnuson, K. (2013), 'Investing in preschool programs', *Journal of Economic Perspectives* **27**(2), 109–132.

Duncan, G. & Vandell, D. (2012), Understanding variation in the impacts of human capital interventions on children and youth, Working paper.

Feller, A., Grindal, T., Miratrix, L. & Page, L. (2013), Compared to what? Variation in the impacts of Head Start by alternative child-care setting, Working paper.

Felts, E. & Page, M. (2013), 'Estimating the distributional effects of education reforms: A look at Project STAR', *Economics of Education Review* **32**, 92–103.

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**(1), 259–276.

Frolich, M. & Melly, B. (Forthcoming), 'Unconditional quantile treatment effects under endogeneity', *Journal of Business and Economic Statistics* .

Garces, E., Currie, J. & Thomas, D. (2002), 'Longer-term effects of Head Start', *American Economic Review* **92**(4), 999–1012.

Gelber, A. & Isen, A. (2013), 'Children's schooling and parents' behavior: Evidence from the Head Start Impact Study', *Journal of Public Economics* **101**, 25–38.

Griffen, A. (2014), Evaluating the effects of child care policies on children's cognitive development and maternal labor supply, Working paper.

Heckman, J. (2006), 'Skill formation and the economics of investing in disadvantaged children', *Science* **312**(5782), 1900–1902.

Heckman, J. (2007), The productivity argument for investing in young children, Working Paper 13016, NBER.

Heckman, J. J., Smith, J. & Clements, N. (1997), 'Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts', *Review of Economic Studies* **64**, 487–535.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P. & Yavitz, A. (2010), 'Analyzing social experiments as implemented: A reexamination of the evidence from the High Scope/Perry Preschool Program', *Quantitative Economics* **1**(1), 1–46.

Heckman, J., Pinto, R. & Savelyev, P. (2010), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American Economic Review* **103**(6), 2052–2086.

Holland, P. (1986), 'Statistics and causal inference', *Journal of the American Statistical Association* **81**(396), 945–970.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467 – 75.

Knudsen, E., Heckman, J., Cameron, K. & Shonkoff, J. (2006), 'Economic, neurobiological, and behavioral perspectives on building America's workforce', *Proceedings of the National Academy of Sciences* **103**(27), 10155–10162.

Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**, 33–50.

Krueger, A. & Zhu, P. (2004), 'Another look at the New York City school voucher experiment', *American Behavioral Scientist* **47**(5), 658–698.

Ludwig, J. & Miller, D. (2007), 'Does Head Start improve children's life chances? Evidence from a regression discontinuity approach', *Quarterly Journal of Economics* **122**(1), 159–208.

Ludwig, J. & Phillips, D. (2008), 'Long-term effects of Head Start on low-income children', *Annals of the New York Academy of Sciences* **1136**, 247–268.

Magnuson, K., Meyers, M., Ruhm, C. & Waldfogel, J. (2004), 'Inequality in preschool education and school readiness', *American Educational Research Journal* **41**(1), 115–157.

Neal, D. & Schanzenbach, D. W. (2010), 'Left behind by design: Proficiency counts and test-based accountability', *Review of Economics and Statistics* **92**(2), 263–283.

NICHD Early Child Care Research Network (2004), 'Modeling the impacts of child care quality on children's preschool cognitive development', *Child Development* **74**, 1454–1475.

Pianta, R. C. (1992), Child-Parent relationship scale, Working paper, University of Virginia. Charlottesville, VA.

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

Puma, M., Bell, S., Cook, R. & Heid, C. (2010), 'Head Start Impact Study: Final report', Prepared for USDHHS, ACF.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A. & Downer, J. (2012), 'Third Grade Follow-Up to the Head Start Impact Study Final Report', Prepared for USDHHS, ACF.

Puma, M., Bell, S., Cook, R., Heid, C. & Lopez, M. (2005), 'Head Start Impact Study: First year findings', Prepared for USDHHS, ACF.

Romano, E., Babchishin, L., Pagani, L. & Kohen, D. (2010), 'School readiness and later achievement: Replication and extension using a nationwide Canadian survey', *Developmental Psychology* **45**(5), 995–1007.

Rubin, D. (1974), 'Estimating causal effects of treatments in randomized and non-randomized studies', *Journal of Educational Psychology* (66), 688–701.

Schrank, F., Becker, K. & Decker, S. (2001), 'Woodcock-Johnson III Asssessment Bulletin Number 4: Calculating ability/achievement discrepancies between the Weschler Intelligence Scale for Children-Third Edition and the Woodcock-Johnson III Tests of Achievement', Riverside Publishing.

Schweinhart, L., Barnes, H. & Weikart, D. (1993), *Significant Benefits: The High/Scope Perry Preschool study through age 27*, High/Scope Press, Ypsilanti, MI.

Schweinhart, L. & Weikart, D. (1981), 'Effects of the Perry Preschool Program on youths through age 15', *Journal of Early Intervention* **3**(29), 29–39.

Shager, H., Schindler, H., Magnuson, K., Duncan, G., Yoshikawa, H. & Hard, C. (2013), 'Can research design explain variation in Head Start research? A meta-analysis of cognitive and achievement outcomes', *Educational Evaluation and Policy Analysis* **35**(1), 76–95.

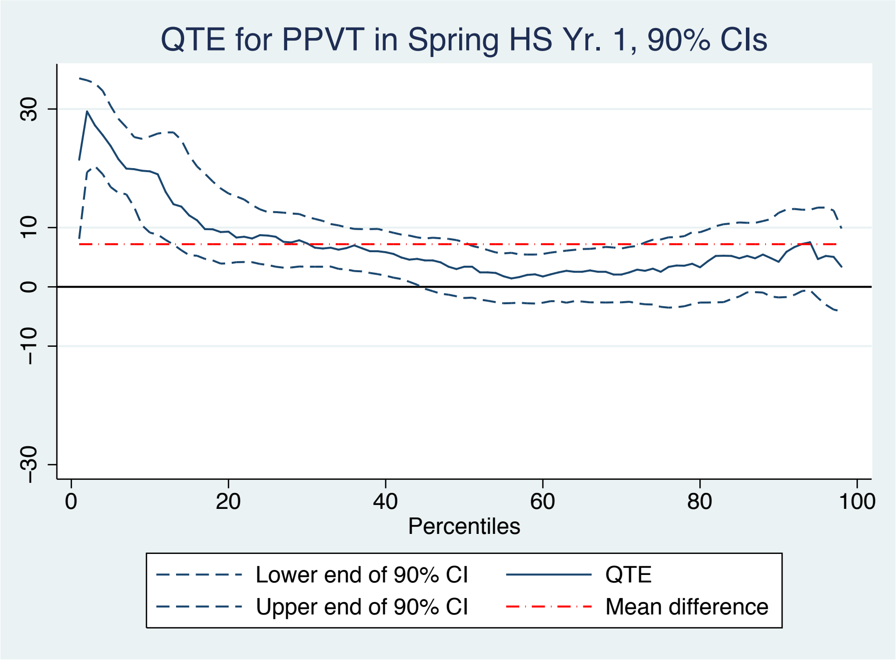
Stanovich, K. (1986), 'Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy', *Reading Research Quarterly* pp. 360–407.

Walters, C. (2014), Inputs in the production of early childhood human capital: Evidence from Head Start, Working paper.

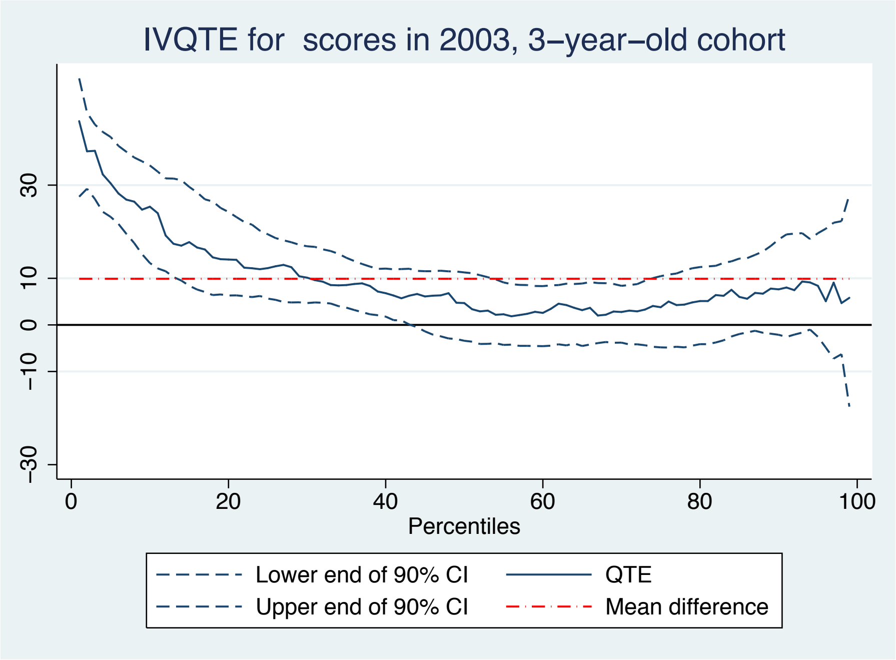
Zanutto, E. (2006), 'A comparison of propensity score and linear regression analysis of complex survey data', *Journal of Data Science* **4**, 67–91.

Figure 1: QTE and IV-QTE for PPVT scores in spring 2003 for the 3-year old cohort, with 90% bootstrapped confidence intervals, using inverse propensity score weights

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

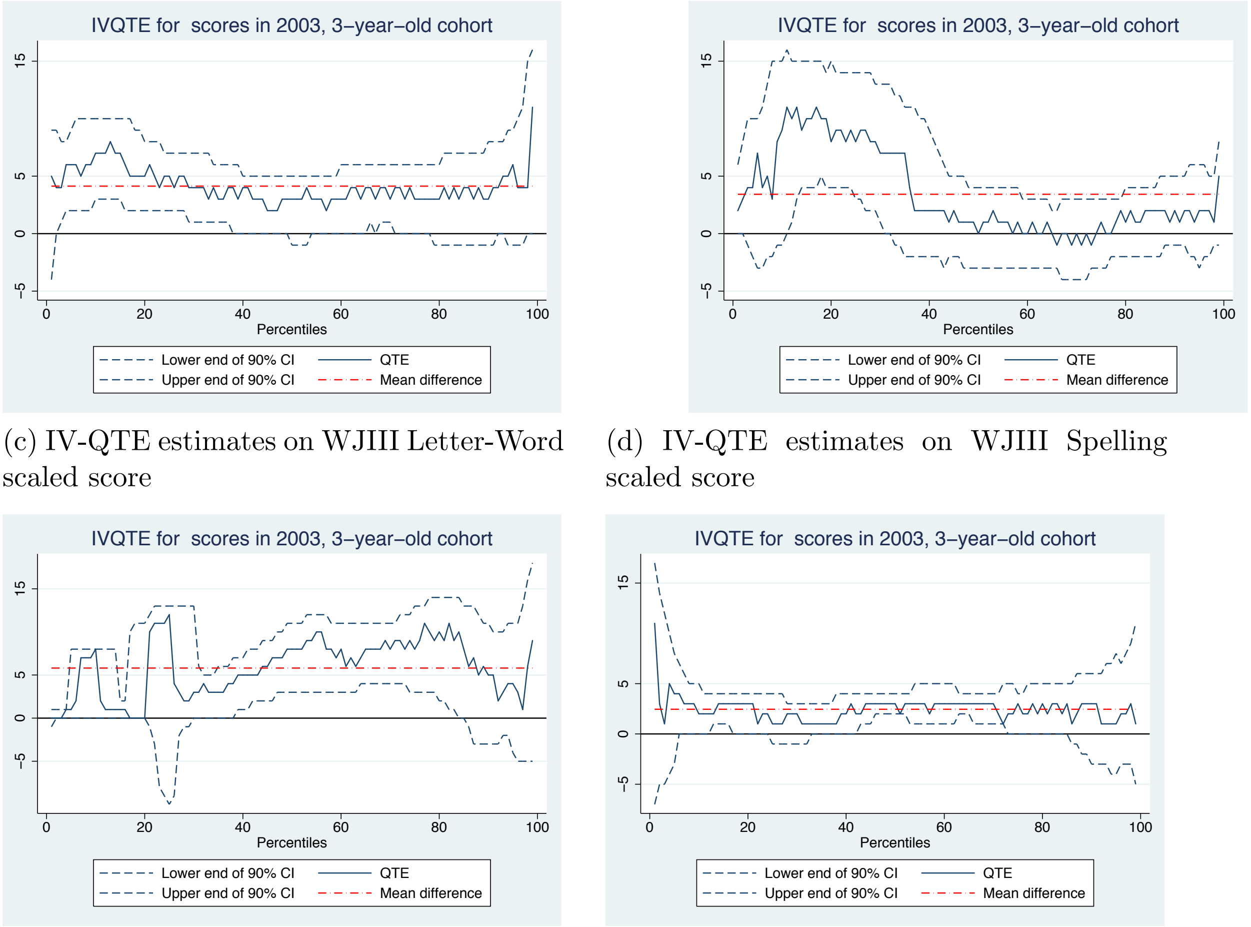


(b) IV-QTE of effect of Head Start on PPVT scores



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 (instrumenting using treatment group status) in the bottom panel, using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center. Dashed horizontal line denotes mean TE (top panel)/mean 2SLS estimate (bottom panel).

Figure 2: 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 & inverse propensity score weights



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

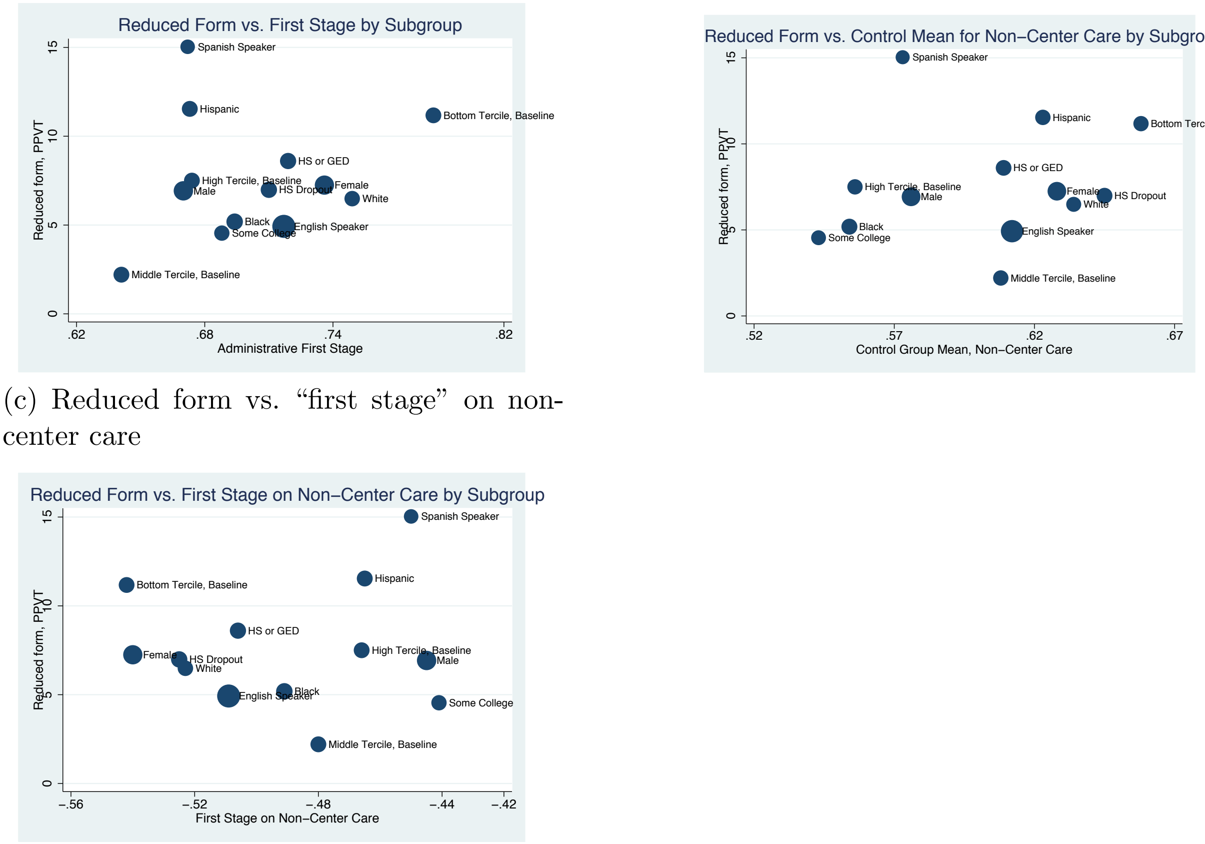
(b) IV-QTE estimates on WJIII Applied Problems scaled score

(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 top left panel presents results for the Pre-Academic Composite score, the top right 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. Dashed horizontal lines denote mean 2SLS estimates.

Figure 3: Reduced-form effects of a Head Start offer on PPVT in 2003 by subgroup compared to the administrative first stage, compared to the control group mean for non-center care, and compared to the "first stage" on non-center care, using inverse propensity score weights

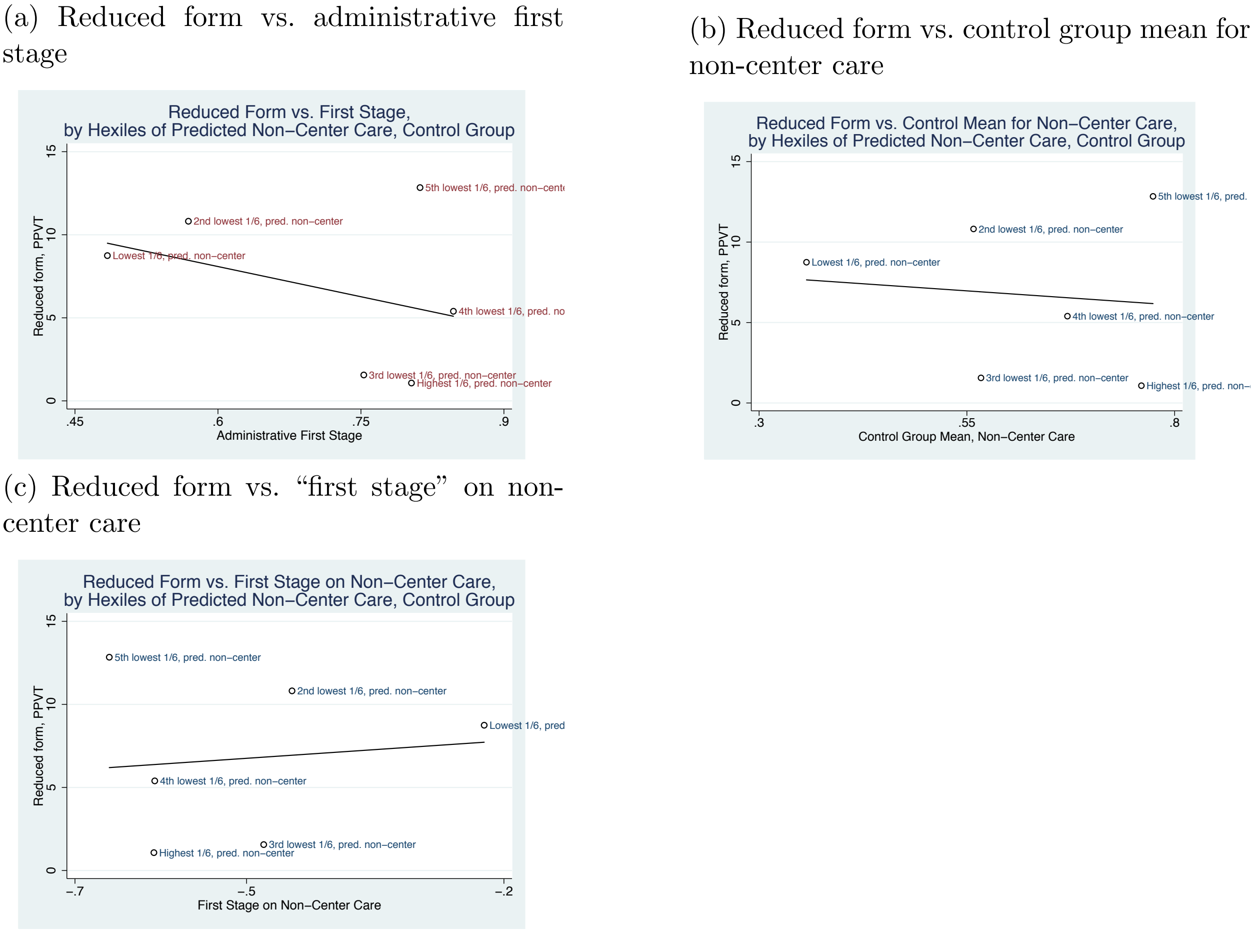

(a) Reduced form vs. administrative first stage

(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 subgroup-specific control group mean by subgroup for non-center care (4b), and versus the subgroup-specific "first stage" effect of an offer of Head Start on take-up of non-center care by subgroup (4c). Subgroups are not mutually exclusive. Size of points reflects the sum of the inverse propensity score weights for the observations in each subgroup. All estimates use the HSIS and inverse propensity score weights.

Figure 4: Reduced-form effects of a Head Start offer on PPVT in 2003 by predicted hexiles of non-center care using the control group compared to the administrative first stage, compared to the control group mean for non-center care, and compared to the “first stage” on non-center care, using inverse propensity score weights



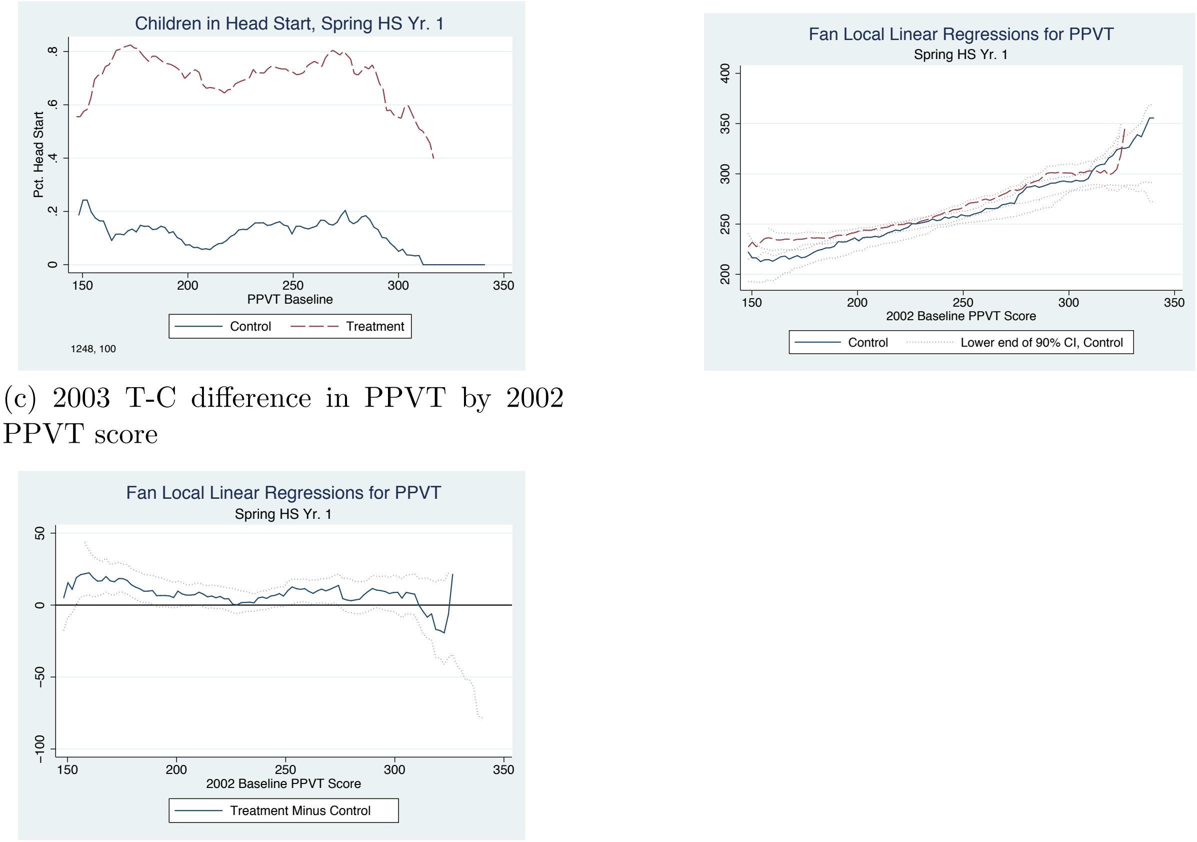
(a) Reduced form vs. administrative first stage

(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 hexiles of predicted non-center care (y-axis) versus the hexile-specific administrative first stage (4a), versus the hexile-specific control group mean by hexile for non-center care (4b), and versus the hexile-specific “first stage” effect of an offer of Head Start on take-up of non-center care by hexile (4c). Hexiles of the predicted value of being in non-center care are constructed as follows (following Abadie, Chingos, and West, 2014). Control group observations’ predicted values use regression estimates that leave the control observation out of the sample (leave one out). Treatment group observations’ predicted values use regressions estimated with the entire control group. Non-center care is predicted using the same Xs as are used to create the inverse propensity score weights. The fitted line is also plotted, weighting each observation by the sum of the inverse propensity score weights for each hexile. All estimates use the HSIS and inverse propensity score weights.

Figure 5: Local linear regression of spring 2003 Head Start participation and 2003 PPVT scores on 2002 baseline scores for treatment and control groups, and T-C difference in 2003 PPVT scores by 2002 score



(a) 2003 HS participation by 2002 PPVT score for T and C

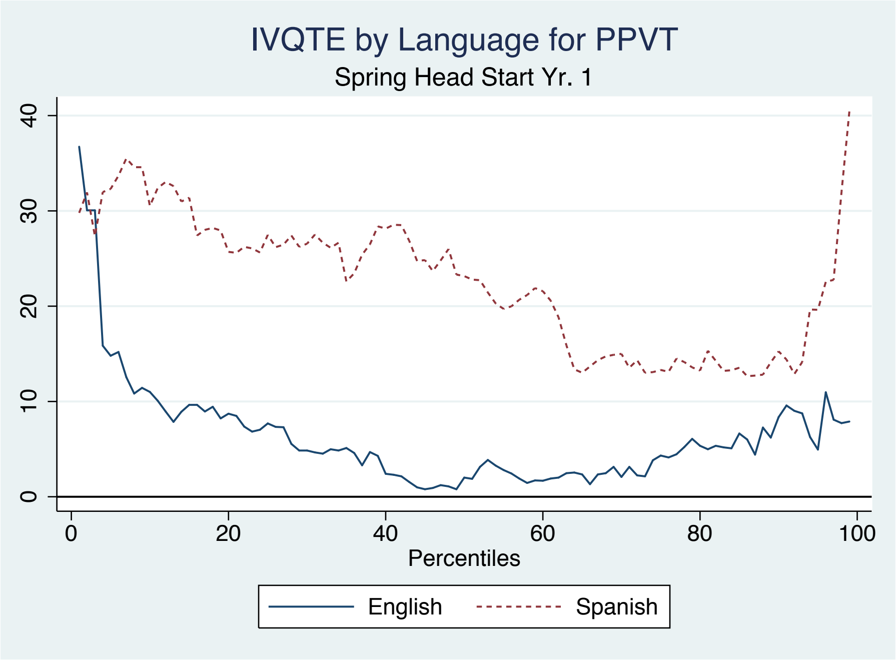
(b) 2003 PPVT score by 2002 PPVT score for T and C

(c) 2003 T-C difference in PPVT by 2002 PPVT score

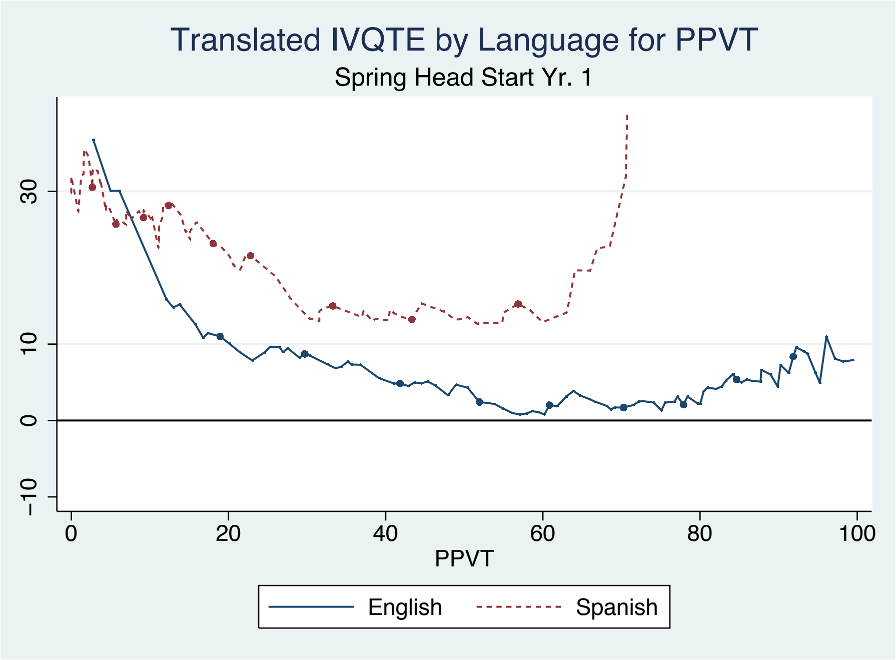
Notes: Figure shows local linear regression of parent-reported Head Start participation in spring 2003 on baseline PPVT test scores, local linear regression of 2003 PPVT score on baseline PPVT test scores, and 2003 treatment minus control difference in local linear regressions of 2003 PPVT score on baseline PPVT test scores, using inverse propensity score weights. The kernel is rectangular and the bandwidth 1/20th of the score range. 90% confidence intervals for panels b and c are obtained by bootstrapping by Head Start center.

Figure 6: Conditional IV-QTE 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



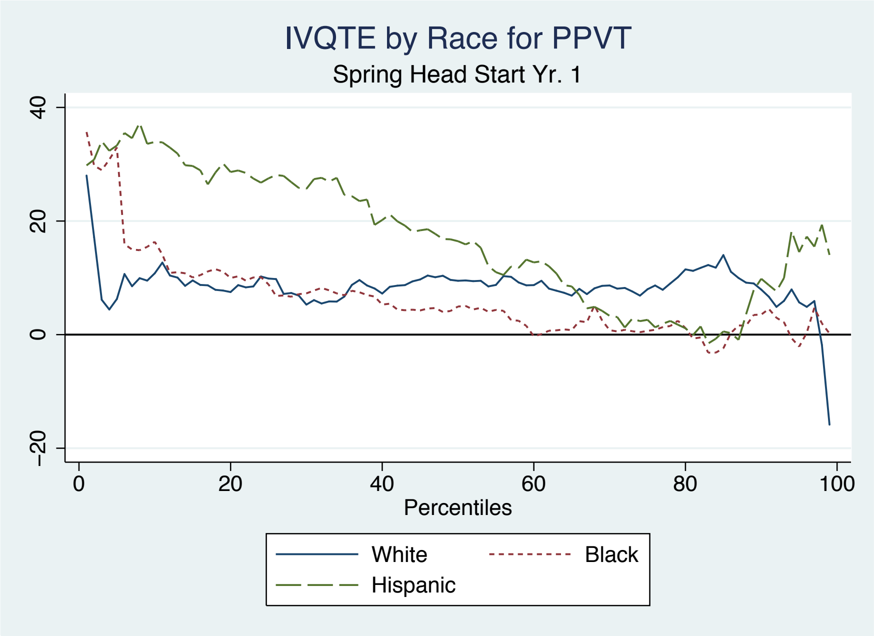
(b) Translated IV-QTE of effect of HS on PPVT scores, by language



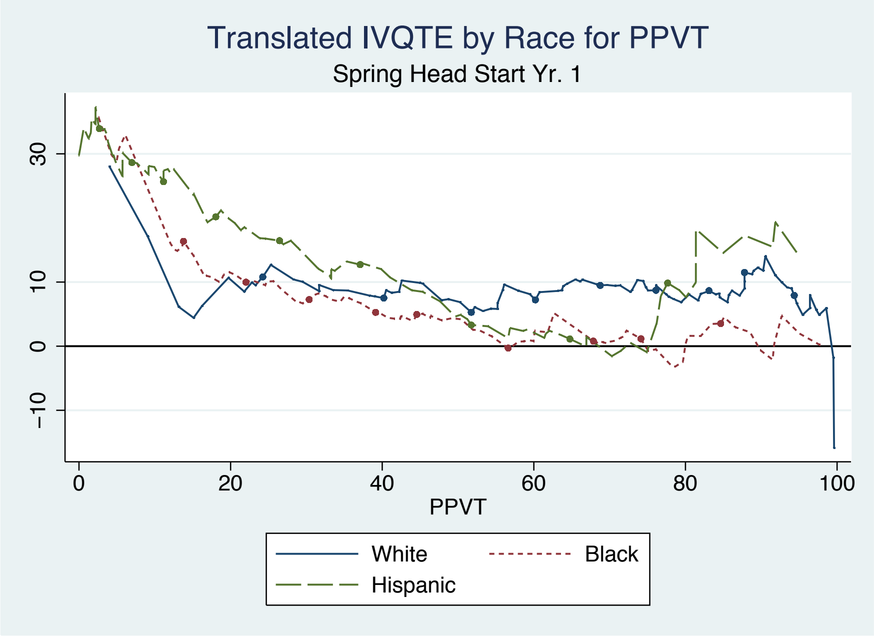
Notes: Figure shows conditional IV-QTE (top panel) and translated IV-QTE (bottom panel) for effect of Head Start participation during the Head Start year on PPVT IRT scores at end of the first year in spring 2003 (at the end of the Head Start year) for the 3-year old cohort, by language, using inverse 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. Dots on translated IV-QTE graphs denote the deciles of the unconditional control group.

Figure 7: Conditional IV-QTE 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



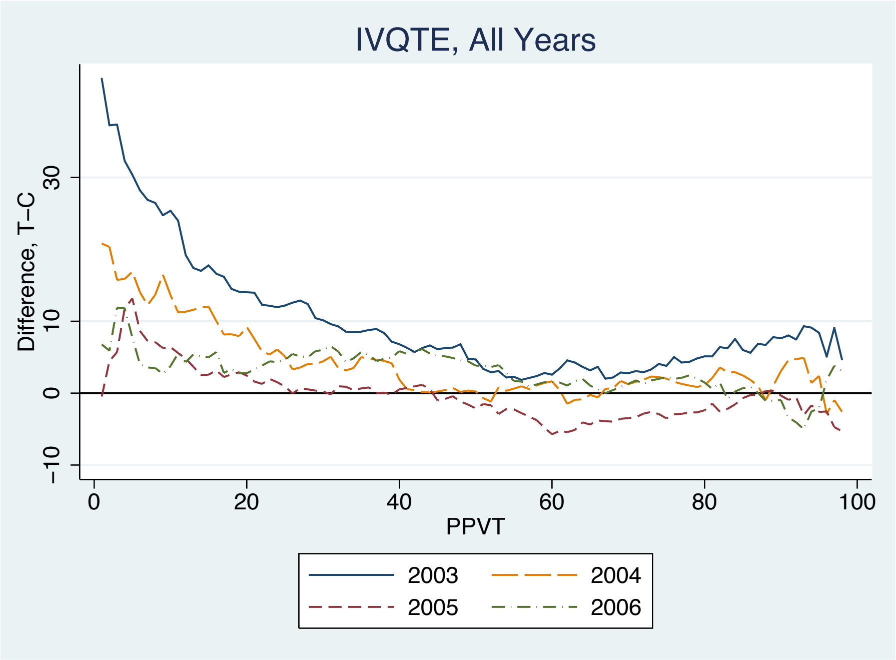
(b) Translated IV-QTE of effect of HS on PPVT scores, by race



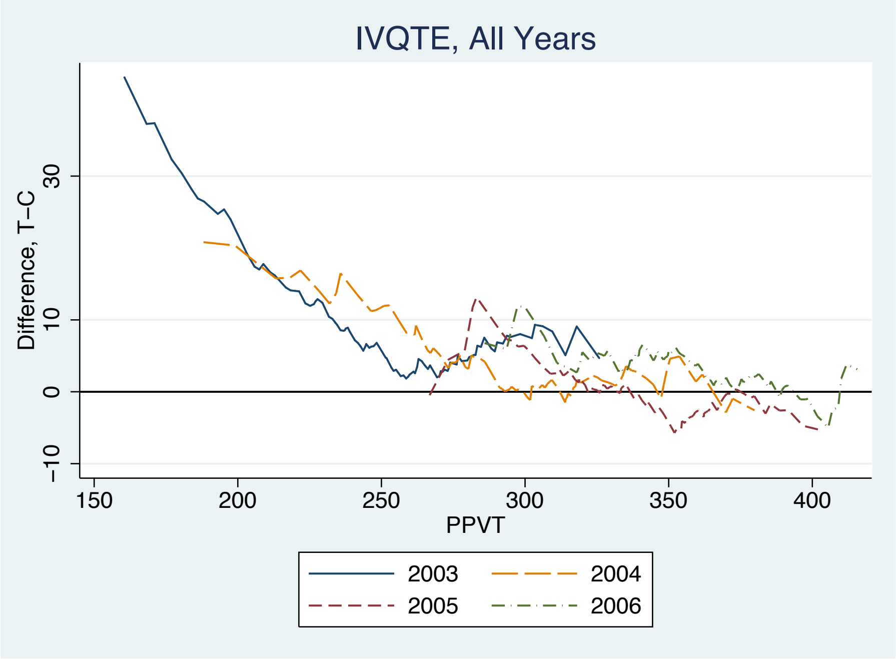
Notes: Figure shows conditional IV-QTE (top panel) and translated IV-QTE (bottom panel) for effect of Head Start participation during the Head Start year on PPVT IRT scores at end of the first year in spring 2003 (at the end of the Head Start year) for the 3-year old cohort, by race/ethnicity (non-Hispanic 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. Dots on translated IV-QTE graphs denote the deciles of the unconditional control group.

Figure 8: Conditional IV-QTE 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



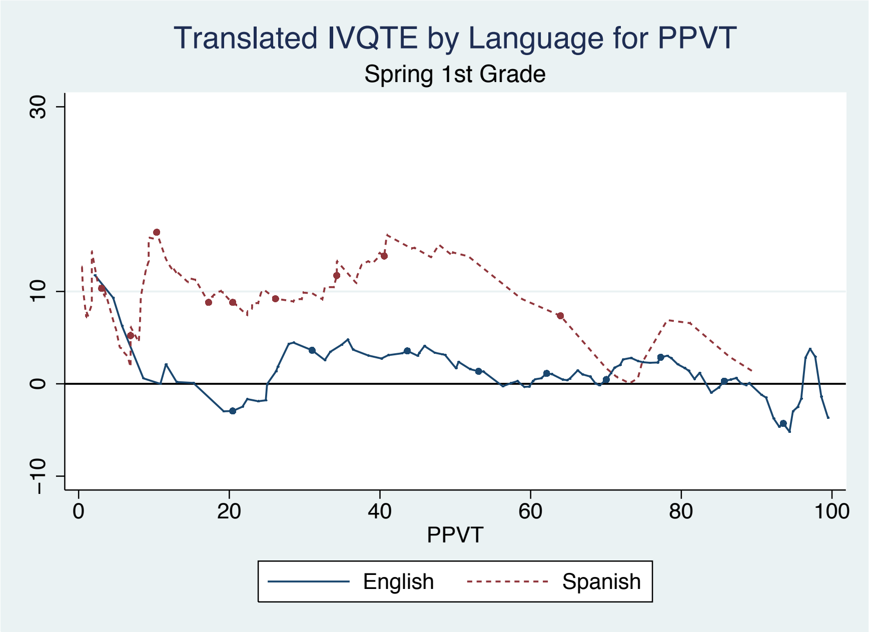
(b) IV-QTE for PPVT scores for all years, translated



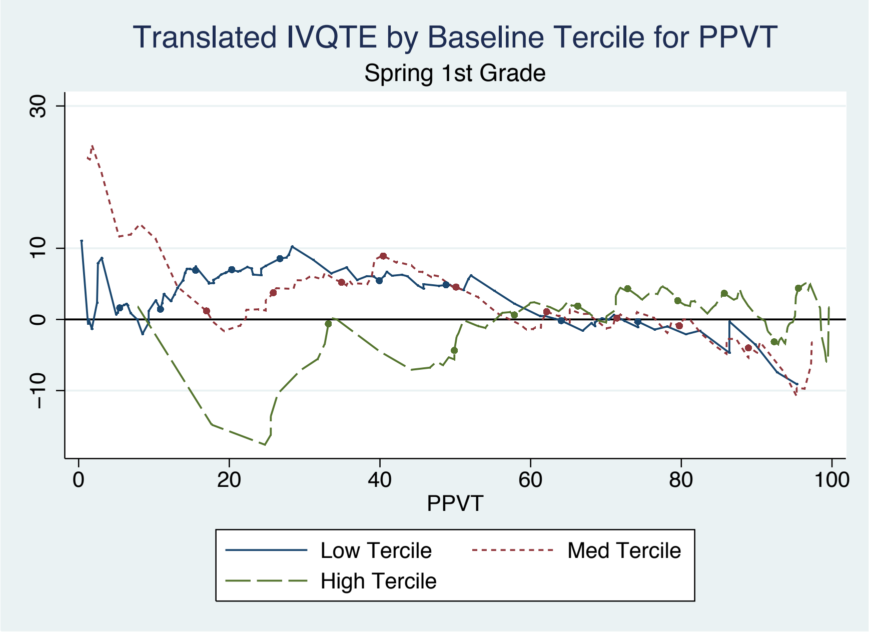
Notes: Figure shows conditional IV-QTE (top panel) and translated IV-QTE (bottom panel) for effect of Head Start participation during the Head Start year on PPVT IRT scores for the 3-year old cohort, for all years, using inverse propensity score weights. Top panel estimates IV-QTE separately for each 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 9: 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 by 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. Dots on translated IV-QTE graphs denote the deciles of the unconditional control group.

Table 1: Summary statistics at baseline

	Baseline Child Weights		Inv. P-Score Weights
	Control Mean	Difference T-C	Difference T-C
Child characteristics			
Non-Hispanic white	0.344	-0.022	-0.001
Non-Hispanic black	0.338	0.005	0.001
Hispanic	0.318	0.017	0.0002
Female	0.527	-0.023	0.017
Spanish speaker at home	0.257	0.001	0.016
Low risk	0.789	-0.041*	0.044**
Medium risk	0.156	0.013	-0.019
High risk	0.055	0.028*	-0.025*
Special needs	0.103	0.031**	0.008
Lives with both biological parents	0.499	-0.003	0.024
Urban	0.793	0.001	0.006
Mother/caregiver characteristics			
Non-Hispanic white	0.364	-0.020	-0.008
Non-Hispanic black	0.339	0.001	0.010
Hispanic	0.298	0.018	-0.001
Mother is teenager	0.176	-0.043*	0.023
High School dropout	0.346	-0.032	0.001
Only high school diploma/GED	0.325	0.033	-0.024
More than high school	0.329	-0.001	0.023
Mom is married	0.457	-0.018	0.022
Mom is divorced	0.149	-0.010	0.017
Mom is never married	0.394	0.028	-0.038
Age 20–24	0.316	-0.052**	0.036
Age 25–29	0.332	-0.016	0.021
Age 30–39	0.259	0.029	-0.016
Age 40 or older	0.064	0.038***	-0.042***
Fall 2002 test month/no Fall assessment			
Before November	0.162	0.099***	0.015
November	0.318	0.029	-0.022
After November	0.260	-0.013	-0.025
No Fall assessment (imputed)	0.259	-0.116***	0.031

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 inverse 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. The sample size for column 1 is 964, for columns 2 and 3 it is 2378. *, **, and *** denote significance at the 10%, 5%, and 1% levels. Significance levels for differences allow for arbitrary correlation within center of random assignment.

Table 2: Child care setting, by treatment or control status

	Treatment Mean	Control Mean	Difference T-C
Head Start in HS Year			
Head Start	0.857	0.153	0.705***
(Administrative report)			
Parent Report, Spring 2003			
Head Start	0.823	0.146	0.677***
Other Center	0.068	0.252	-0.183***
Family day care	0.014	0.064	-0.050***
Parent/relative	0.094	0.536	-0.442***
Not reported	0.001	0.002	-0.001
Parent Report, Spring 2004			
Head Start	0.608	0.473	0.135***
Other Center	0.250	0.355	-0.105***
Family day care	0.018	0.015	0.003
Parent/relative	0.077	0.103	-0.025
Kindergarten	0.016	0.021	-0.005
Not reported	0.031	0.033	-0.002

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 the end of the Head Start year—spring 2003 (middle panel) and for the end of the Age-4 year—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/Kindergarten for children in Kindergarten in spring 2004. 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. *, **, and *** denote significance at the 10%, 5%, and 1% levels. Significance levels for differences allow for arbitrary correlation within center of random assignment.

Table 3: Reduced-form effects of a HS slot offer on PPVT scores and on test score missing (attrition) and 2SLS effects of HS participation on PPVT scores, by year

	Inverse P-Score Weights			Baseline Child Weights	
	Control mean [SD]	Reduced form (SE)	2SLS (SE)	Control mean [SD]	Reduced form (SE)
PPVT scores					
Baseline PPVT, fall HS year	231	-0.003		231	-0.88
(fall 2003)	[38]	(1.84)		[39]	(2.25)
PPVT, spring HS year	251	7.20***	10.20***	252	6.56***
(spring 2003)	[38]	(1.64)	(2.40)	[37]	(2.04)
PPVT, Age-4 year	298	2.89	4.15	299	2.49
(spring 2004)	[40]	(1.81)	(2.60)	[41]	(2.31)
PPVT, Kindergarten year	339	0.21	0.30	340	0.76
(spring 2005)	[29]	(1.29)	(1.84)	[29]	(1.49)
PPVT, first grade year	358	2.00	2.90	358	3.05
(spring 2006)	[30]	(1.42)	(2.07)	[30]	(1.94)
PPVT missing or imputed					
Imputed/no baseline PPVT fall HS year	0.222	0.031		0.281	-0.118***
(fall 2002)	[0.416]	(0.022)		[0.450]	(0.022)
No PPVT spring HS year	0.190	-0.009		0.219	-0.103***
(spring 2003)	[0.393]	(0.023)		[0.414]	(0.018)
No PPVT Age-4 year	0.179	0.009		0.212	-0.072***
(spring 2004)	[0.383]	(0.022)		[0.409]	(0.022)
No PPVT Kindergarten year	0.233	0.005		0.235	-0.054**
(spring 2005)	[0.423]	(0.022)		[0.424]	(0.021)
No PPVT first grade year	0.251	0.001		0.254	-0.065***
(spring 2006)	[0.434]	(0.022)		[0.436]	(0.023)

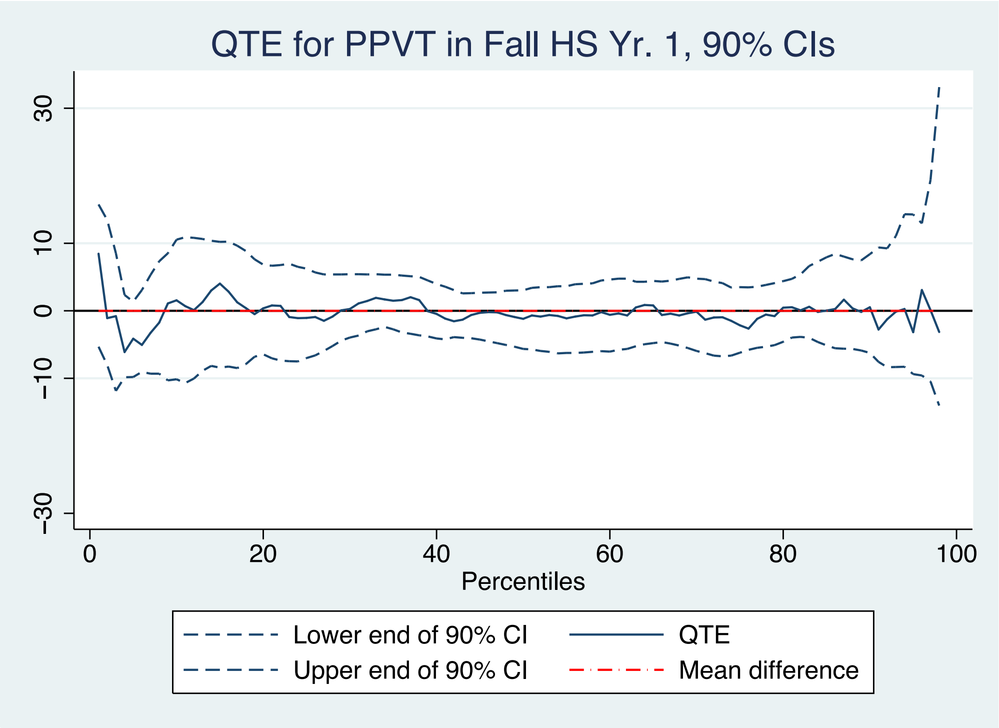
Notes: Table reports means (columns 1 and 4), mean reduced-form estimates (ITT) of the effect of being offered a Head Start slot (columns 2 and 5), and two-stage least squares (TOT) estimates of the effect of attending Head Start in the Head Start Year (column 3), for the 3-year old cohort in the Head 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 is 1 for observations where the test score is missing or imputed). Results reported for 2002 include the imputed values for the tests. 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-stage least squares (TOT) estimates of the effect of attending Head Start during the Head Start year using inverse propensity score weights with the standard errors in parentheses. The first stage for the Head Start year (spring 2003) is 0.705 (SE is 0.029), that for the Age-4 year (spring 2004) is 0.696 (SE is 0.029), that for the Kindergarten year (spring 2005) is 0.700 (SE is 0.028), and that for the first grade year (spring 2006) is 0.694 (SE is 0.029). 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. Inverse 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 4: 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 take-up, 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)

	Control mean [SD]	Reduced form (SE)	First stage (SE)	Two-stage least squares (SE)
Race/ethnicity subgroups				
Hispanic	234	11.54***	0.673***	17.13***
	[39]	(3.06)	(0.060)	(4.94)
Non-Hispanic Black	250	5.19**	0.694***	7.48*
	[32]	(2.58)	(0.046)	(3.83)
Non-Hispanic White	268	6.49**	0.749***	8.66**
	[34]	(3.14)	(0.033)	(4.23)
Language subgroups				
Spanish speaker at home	223	15.0***	0.672***	22.37***
	[32]	(3.30)	(0.057)	(5.65)
English speaker at home	261	4.93***	0.717***	6.87***
	[34]	(1.86)	(0.029)	(2.62)
Baseline PPVT score tercile subgroups				
Bottom tercile, baseline PPVT	229	11.2***	0.787***	14.2***
	[33]	(2.77)	(0.035)	(3.72)
Middle tercile, baseline PPVT	251	2.21	0.641***	3.44
	[30]	(2.45)	(0.047)	(3.84)
Top tercile, baseline PPVT	274	7.50**	0.674***	11.13**
	[36]	(3.16)	(0.041)	(4.78)

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. Inverse propensity score weights estimated including baseline demographics and baseline test deciles by approximate month of assessment. SEs allow for arbitrary correlation within center of random assignment. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.

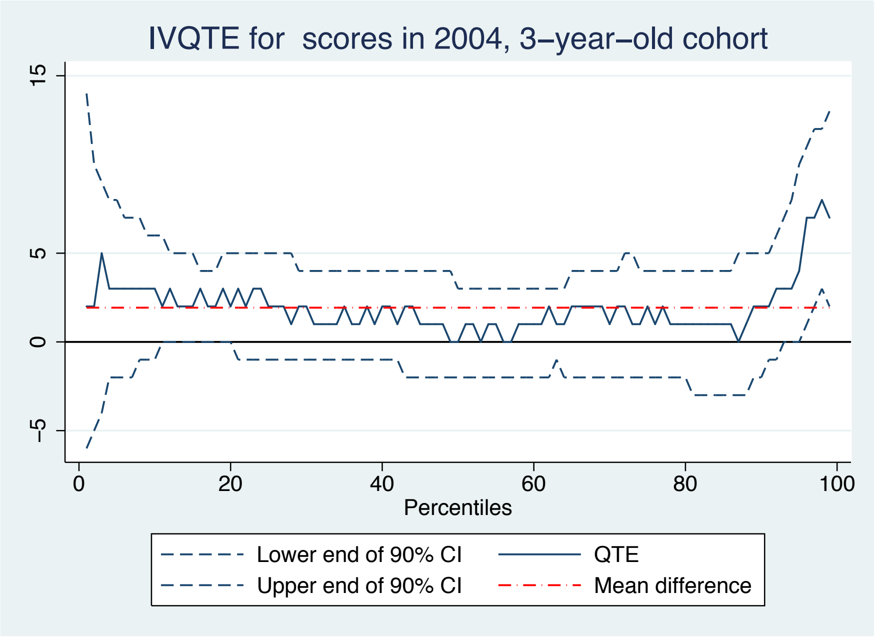
For online publication only: Appendix Figure 1: QTE for PPVT scores at baseline for the 3-year old cohort, with 90% bootstrapped confidence intervals, using inverse propensity score weights



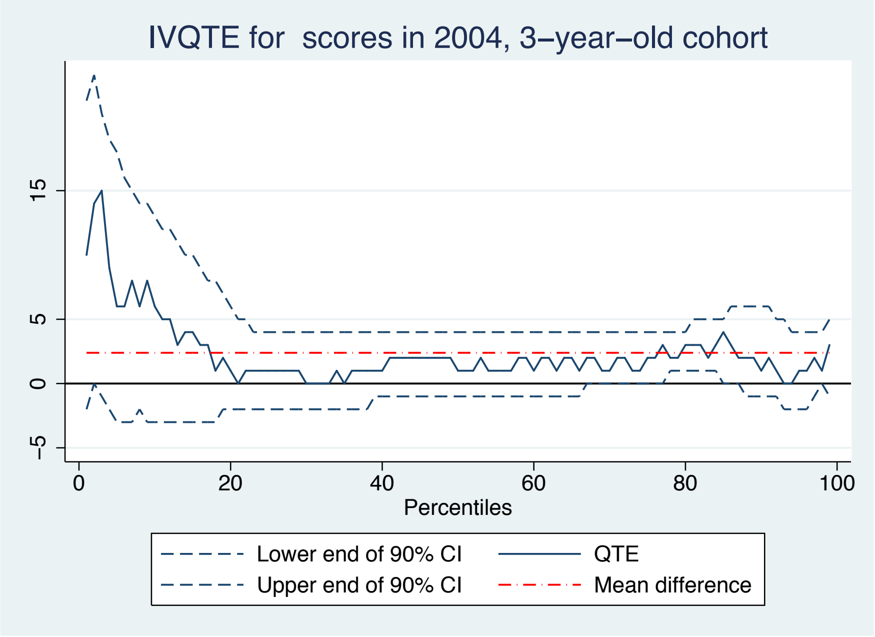
Notes: Figure shows QTE for PPVT IRT score at baseline in Fall 2002 for the 3-year old cohort, using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center.

For online publication only: Appendix Figure 2: IV-QTE of Woodcock Johnson III Pre-Academic and Applied Problems tests for spring 2004

(a) Pre-Academic Skills



(b) Applied Problems



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 top panel presents results for the Pre-Academic Composite score and the bottom panel presents results for the Applied Problems score, both using inverse propensity score weights. 90% confidence intervals are obtained by bootstrapping by Head Start center. Dashed horizontal lines denote mean 2SLS estimates.

For online publication only: Appendix Table 1: Cognitive and social-emotional outcomes in the HSIS, by normative year in school

	Head Start year			
	Age 3	Age 4	Kindergarten	1st grade
Language, literacy, vocabulary				
PPVT	X	X	X	X
Letter Word (WJIII)	X	X	X	X
Spelling (WJIII)	X	X	X	X
Mathematics				
Applied Problems (WJIII)	X	X	X	X
Composite				
Pre-Academic Skills (WJIII)	X	X	X	X
Social-emotional outcomes, parent reports				
Aggressive behavior	X	X	X	X
Hyperactive	X	X	X	X
Withdrawn	X	X	X	X
Social competencies	X	X	X	X
Social skills/positive learning	X	X	X	X
Behavioral problems	X	X	X	X
Conflict (Pianta)	X	X	X	X
Closeness (Pianta)	X	X	X	X
Positive relationships (Pianta)	X	X	X	X
Social-emotional outcomes, teacher report				
Aggressive behavior (ASPI)			X	X
Hyperactive (ASPI)			X	X
Withdrawn (ASPI)			X	X
Shy (ASPI)			X	X
Oppositional (ASPI)			X	X
Problems with peer interactions (ASPI)			X	X
Problems with structure (ASPI)			X	X
Interaction problems (ASPI)			X	X
Conflict (Pianta)			X	X
Closeness (Pianta)			X	X
Positive relationships (Pianta)			X	X

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 social-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).

For online publication only: Appendix Table 2: Effect of Head Start offer on parent report of child care arrangements, by demographic subgroup for the Head Start year (spring 2003)

		Head Start	Other Center	Parent/relative/ Other care
Race/ethnicity subgroups				
Hispanic	First stage	0.636***	-0.170***	-0.465***
	(SE)	(0.060)	(0.038)	(0.056)
	Control mean	0.137	0.240	0.623
Non-Hispanic Black	First stage	0.683***	-0.191***	-0.491***
	(SE)	(0.047)	(0.034)	(0.045)
	Control mean	0.178	0.269	0.554
Non-Hispanic White	First stage	0.712***	-0.189***	-0.523***
	(SE)	(0.034)	(0.033)	(0.041)
	Control mean	0.121	0.245	0.634
Language subgroups				
Spanish speaker at home	First stage	0.641***	-0.191***	-0.450***
	(SE)	(0.057)	(0.049)	(0.055)
	Control mean	0.152	0.275	0.573
English speaker at home	First stage	0.690***	-0.181***	-0.509***
	(SE)	(0.029)	(0.023)	(0.030)
	Control mean	0.144	0.243	0.612
Baseline PPVT score tercile subgroups				
Bottom tercile, baseline PPVT	First Stage	0.721***	-0.179***	-0.542***
	(SE)	(0.036)	(0.035)	(0.044)
	Control mean	0.123	0.219	0.658
Middle tercile, baseline PPVT	First Stage	0.638***	-0.158***	-0.480***
	(SE)	(0.047)	(0.035)	(0.048)
	Control mean	0.165	0.227	0.608
Top tercile, baseline PPVT	First Stage	0.656***	-0.190***	-0.466***
	(SE)	(0.040)	(0.032)	(0.039)
	Control mean	0.155	0.289	0.556

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 inverse propensity score weights. Inverse propensity score weights estimated including baseline demographics and baseline test deciles by approximate month of assessment. SEs allow for arbitrary correlation within center of random assignment. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.

For online publication only: Appendix Table 3: Two-stage least squares estimates of the effect of attending Head Start on social-emotional outcomes for spring 2003 and 2006

	Head Start year	Grade 1 year
	(Spring 2003)	(Spring 2006)
Parent reports		
Aggressive (ASPI)	-0.115*	-0.084
	(0.068)	(0.078)
Hyperactive (ASPI)	-0.274***	-0.111
	(0.074)	(0.074)
Lack of Social Competencies (ASPI)	0.022	-0.067
	(0.071)	(0.079)
Lack of Social Skills (ASPI)	-0.032	-0.014
	(0.075)	(0.076)
Withdrawn (ASPI)	0.027	-0.069
	(0.071	(0.076)
Conflict (Pianta)	-0.012	-0.169**
	(0.066)	(0.075)
Lack of closeness (Pianta)	-0.121*	-0.099
	(0.064)	(0.076)
Lack of positive relationship (Pianta)	-0.048	-0.166**
	(0.066)	(0.078)
Externalizing behavior problems	-0.169**	-0.106
	(0.070)	(0.073)
Teacher reports		
Aggressive (ASPI)		-0.059
		(0.085)
Oppositional (ASPI)		-0.009
		(0.083)
Inattentive (ASPI)		-0.074
		(0.079)
Shy/socially reticent (ASPI)		0.068
		(0.079)
Withdrawn/low energy (ASPI)		0.043
		(0.081)
Problems with structured learning (ASPI)		-0.021
		(0.082)
Problems with peer interactions (ASPI)		-0.055
		(0.082)
Problems with teacher interactions (ASPI)		-0.019
		(0.077)
Combined ASPI index—negativity		-0.048
		(0.074)
Combined ASPI index—shy		0.057
		(0.071)
Combined ASPI index—interactive		-0.032
		(0.068)
Lack of closeness (Pianta)		0.010
		(0.087)
Lack of positive relationship (Pianta)		0.016
		(0.089)
Conflict (Pianta)		0.023
		(0.088)

Notes: Table reports two-stage least squares estimates of the effect of attending Head Start at age 3 on social-emotional outcomes. Outcomes are 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 standardized to be "bad" outcomes, and then to have mean 0 and standard deviation 1 except for the combined ASPI indices. SEs allow for arbitrary correlation within center of random assignment. *, **, and *** denote significance at the 10%, 5%, and 1% levels. See text for more details.