Experimental Evidence on Distributional Effects of Head Start

Marianne P. Bitler
University of California, Irvine and NBER

Hilary W. Hoynes
University of California, Berkeley and NBER

Thurston Domina
University of California, Irvine[*]

This version: November 2013

[*]Correspondence to Bitler at UC Irvine Department of Economics, 3151 Social Science Plaza, Irvine, CA 92697, phone (949) 824-5606, fax (949) 824-2182, or mbitler@uci.edu; Hoynes at hoynes@berkeley.edu; or Domina at tdomina@uci.edu. This research was also supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number PO1HD065704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are grateful to Child Care & Early Education Research Connections for providing access to the experimental data used here. We would also like to thank Josh Angrist, Peg Burchinal, Greg Duncan, George Farkas, Avi Feller, Jesse Rothstein, Adam Isen, Doug Miller, Luke Miratrix, Deborah Vandell, 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, UCLA, UC Davis, UC Irvine, UC Santa Cruz, APPAM, the Institute for Research on Poverty, the NBER Summer Institute, and SREE for helpful comments. Kevin Williams provided expert research assistance.

Abstract

The federal Head Start program funds public preschools for eligible low-income families. In so doing it aims to raise educational attainment levels and narrow educational inequalities. Existing research demonstrates that Head Start has short run positive impacts on cognitive outcomes that fade out for many groups by elementary school (e.g., Currie & Thomas (1995)). However, there appear to be some lasting positive impacts of Head Start on educational and other outcomes (Deming (2009); Garces, Currie & Thomas (2002); Ludwig & Miller (2007)). In this study, we provide the first comprehensive analysis of the distributional effects of Head Start. We do so using the first national randomized experiment of the program (the Head Start Impact Study). We examine effects on cognitive and non-cognitive outcomes and explore the heterogeneous effects of the program through 1st grade by estimating quantile treatment effects and various types of subgroup mean treatment effects. We find that an offer of a Head Start slot leads to large and statistically significant gains in cognitive skills in the preschool period. However, once the children enter school, the cognitive gains fade out. Importantly, we find that some groups—including Spanish speakers or those entering preschool with low baseline scores—have cognitive gains that persist through 1st grade. We find little effect of an offer of a Head Start slot on non-cognitive outcomes. Our results provide strong evidence in favor of a compensatory model of educational process in this context.

1 Introduction

Created in 1965, the federal Head Start program is among the more prominent educational initiatives in the US. By giving matching grants to programs providing comprehensive early education, health care, and nutritional services to poor children; and parenting training to their parents; Head Start aims to raise educational attainment levels and narrow educational inequalities. Head Start now enrolls more than 900,000 children and has an annual operating budget of nearly $7 billion (US DHHS, Office of Head Start 2008).

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

Decades of research have documented that Head Start generates important and statistically significant impacts on cognitive outcomes (e.g., see Currie & Thomas (1995) and the review in Currie (2001)). However, concerns about the 'fadeout' of Head Start's impact on cognitive test scores appeared early in the program history, and evidence of fade-out has been shown in more recent cohorts (e.g., Currie & Thomas (1995)). Interestingly, despite this documented fade-out, other work shows positive longer-term impacts of Head Start on educational attainment and young adult outcomes (Garces et al. (2002), Ludwig & Miller (2007), Deming (2009)).

In this study, we provide the first comprehensive analysis of the distributional effects of Head Start. In particular, we use quantile treatment effects and estimate the effect of

Head Start on the distribution of cognitive and noncognitive outcomes. We augment this with various overall and subgroup mean treatment effect estimates. Our analysis is based on data from the Head Start Impact Study, a randomized control trial that follows nearly 5,000 children in two cohorts–three and four years old at the time of Head Start application–through first grade, collecting detailed outcomes on academic and social-emotional measures.

Our analysis makes three contributions. First, by moving beyond the analysis of mean impacts, we can test two competing hypotheses concerning how Head Start impacts vary across the skill distribution. Observational studies indicate that low-achieving children stand to gain the most by enrolling in early education (Manguson, Meyers, Ruhm & Waldfogel (2004); NICHD Early Child Care Research Network (2004)). This effect may be particularly pronounced in the context of Head Start, since the program's curricula are explicitly geared toward both remedying the skills' deficits that often disadvantage poor students at the beginning of elementary school and improving the parenting practices of their parents (Puma, Bell, Cook, Heid & Lopez (2005)). On the other hand, research on learning trajectories in elementary school and beyond indicates that since academic skills are cumulative, achievement inequalities tend to widen as children progress through school (Stanovich (1986)). Applied to early education, this suggests that students who have basic language and numeric competencies are more able to participate fully in the Head Start curriculum than their academically or developmentally-delayed peers. Thus, our analysis will test the 'compensatory' hypothesis (predicting largest gains at the bottom of the skill distribution) against the 'skills-begets-skills' hypothesis (predicting largest gains at the top of the skill distribution).

Second, recent evidence suggests that varying educational interventions such as smaller class sizes (Dynarski, Hyman & Schanzenbach (2011)), intensive preschools (Heckman, Moon, Pinto, Savelyev & Yavitz (2010), Anderson (2008), Schweinhart, Montie, Xiang, Barnett, Belfield & Nores (2005)), and Head Start show a 'fade-out' in cognitive test scores yet yield significant positive longer-term educational and other young adult outcomes. It may be that by focusing on the mean impacts of cognitive outcomes, we have missed some persistent positive effects of test scores. Third, we provide new evidence on the impacts on non-cognitive outcomes, which are a potential channel for the already documented positive longer-term outcomes (Heckman (2007), Garces et al. (2002)).

The Head Start Impact Study, mandated by Congress in 1998, is the first national randomized evaluation of the Head Start program. The study is designed to determine "the impact of Head Start on children's school readiness and parental practices" as well as "under what circumstances Head Start achieves its greatest impact and for which children" (Puma et al. (2005)). The study randomized children applying to oversubscribed Head Start centers for the first time to either an offer of a slot or denial of a slot for one year. These data follow cohorts of children who started out being 3- or 4-years old through first grade, collecting detailed outcomes on academic and social-emotional measures.

In this paper, we use the 3-year old cohort and richly examine the impacts of an offer of Head Start slot 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. We examine cognitive tests such as the Peabody Picture Vocabulary Test (PPVT), as well as various measures from the Woodcock Johnson III battery of achievement tests. We also take advantage of the various non-cognitive scores; for example, teacher reports of child behavior using the Adjustment Scales for Preschool Intervention (ASPI).

We find that the offer of a Head Start slot leads to large and statistically significant gains in cognitive skills in the preschool period. However, once the children enter school, the cognitive gains fade out. Importantly, we find that some groups—including Spanish speakers or those entering preschool with low baseline scores—have cognitive gains that persist through 1st grade. We find little effect on an offer of Head Start non-cognitive outcomes.

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 increases in cognitive outcomes (see the review by Currie (2001)). However, it is less clear how long those test score gains persist. For example, Currie & Thomas (1995), using a within-family, sibling-comparisons research design, find that African American participants experience fade out in test score gains while still in elementary grades. They show that white participant gains persist into adolescence.

More recently, focus has turned to longer-term impacts of Head Start. Three quasi-experimental studies indicate that Head Start has positive long-term effects. Ludwig & Miller (2007) use a regression discontinuity approach to estimate long-term Head Start effects. In 1965, the federal Office of Economic Opportunity provided technical assistance in writing grants for the first round of Head Start programs to the 300 counties with the highest poverty rates in the U.S. This offer resulted in a sharp discontinuity in Head Start funding between counties that qualified for application assistance and similarly poor counties that did not qualify. Exploiting this discontinuity, Ludwig & Miller (2007) find that an increase from 50% to 100% in Head Start funding decreases child mortality by nearly 50%, increases youth educational attainment by approximately half a year, and increases college attendance rates. Ludwig & Miller (2007) find little evidence to suggest that Head Start effects vary by race.

Garces et al. (2002) use a within-family, sibling comparison research design and also estimate the long-term impacts of Head Start. Using data from the Panel Study of Income Dynamics (PSID), they compare young adults aged 18 and older who had enrolled in Head Start as children with adult siblings who had enrolled in some other form of preschool. They find that Head Start participation significantly boosts educational attainment for whites but not blacks; yet it significantly decreases criminality for blacks but not whites. White youth who participated in Head Start are approximately 22 percentage points more likely to finish high school and 19 percentage points more likely to attend some college than their siblings who did not participate in Head Start. While black youth who participated in Head

Start do not differ from their siblings on these measures of educational attainment, they are approximately 12 percentage points less likely to be charged with a crime than are their siblings who did not participate in the program. They find no significant impacts of Head Start on adult earnings for either group.

Deming (2009) uses a similar approach to estimate effects for children who enrolled in Head Start between 1984 and 1990 and participated in the National Longitudinal Survey of Youth Mother/Child study. Deming (2009) finds that Head Start participation improves student test scores by 0.15 standard deviations, although these effects fade out by the time students reach middle school. Despite this evidence that Head Start's cognitive effects fade out, he also finds that program participation has a positive long-term 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). Deming (2009) finds that the short-term cognitive effects of Head Start are more pronounced and positive for blacks than whites. Further, in contrast to Garces et al. (2002), he finds no evidence to suggest that Head Start's long-term effects vary by race.

Taken together, these studies complement earlier research indicating that early childhood education programs for poor children have substantial positive effects across the life course (Schweinhart et al. (1993); Belfield, Nores, Barnett & Schweinhart (2006); Masse & Barnett (2007)). However, both the internal and external validity of these studies are limited in important ways. The extent to which the Ludwig & Miller's (2007) findings generalize is unclear, since their analyses are based on discontinuities in Head Start funding among very poor counties in 1965. For the Garces et al. (2002) and Deming (2009) studies, important questions exist about why families might send one sibling to Head Start and not another and the extent to which this selection process might bias estimates of the effects of Head Start.

Still to come: add additional cites.

2.2 Theoretical Expectations and Heterogeneity Elsewhere in Social and Educational Policy

To date, analyses of the HSIS data have focused primarily on the program's mean effects across the entire treatment population. But these mean effects may hide heterogeneity in Head Start effects across the distribution of student achievement and socio-emotional outcomes. In this study, by moving beyond the analysis of mean impacts, we can test two competing hypotheses concerning how Head Start impacts vary across the skill distribution. Observational studies indicate that low-achieving children stand to gain the most by enrolling in early education (Manguson et al. (2004); NICHD Early Child Care Research Network (2004)) and this effect may be particularly pronounced in the context of Head Start. Head Start has traditionally focused particular attention on preparing the most disadvantaged students for school entry. This focus may lead Head Start programs to emphasize the most basic cognitive skills. The performance standards that Congress articulated for Head Start in its 1998 reauthorization speak to this curricular emphasis, calling on the program to ensure that all students recognize a word as a unit of print and can identify at least 10 letters (DHHS ACF (2003)). While Head Start is a highly decentralized program and does not have a single coordinated curriculum, these performance standards were widely publicized within the program, and may have influenced instructional priorities (DHHS ACF (2000)). Similarly, the Head Start's mission of serving "at-risk" youth may lead Head Start programs to dedicate particular attention to the most socially and emotionally troubled children.

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)). Applied to early education, this suggests that students who have basic language and numeric competencies are likely more able to participate fully in the Head Start curriculum than their academically or developmentally-delayed peers. Thus, 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).

The final HSIS provides some evidence on potential heterogeneous effects of Head Start. Puma, Bell, Cook & Heid (2010) report that the offer of a Head Start slot had greater positive short-term 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 the first grade.

Still to come: summary from other settings.

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 randomization.

3.1 The HSIS Experiment and Evaluation

The HSIS grew out of a Congressional mandate as part of the 1998 re-authorization of Head Start. The study sample consists of nearly 5,000 new applicants to oversubscribed Head Start centers in a nationally representative sample of oversubscribed locations within 84 programs across the U.S. With some exceptions to account for participation in other evaluations, all over-subscribed centers were at risk of inclusion. Severely disabled children were excluded, but in the fall of 2002, the other applicants were randomly assigned to a treatment group that received an offer to enroll in the Head Start center of application and a control group that did not.

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

The evaluation of the HSIS is expected 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)). As described below, in our analysis we use the most up-to-date public release of the data, which includes outcomes through grade 1.

The reports on the HSIS show modest 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, the HSIS data indicate that most of the cognitive effects of Head Start placement offers decay as students move out of the program and into elementary school. In contrast, there is not consistent evidence for any effects of the offer of a Head Start slot on socio-emotional (non-cognitive) outcomes. Most impacts on the non-cognitive outcomes are small and statistically insignificant. Overall, any positive socio-emotional effects were limited to parent reports for the 3-year old cohort; for example parents of 3-year olds offered the 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 available for all children in Kindergarten (they are not asked outside of center care), show no significant effects of Head Start on socio-emotional outcomes.

As is common for this type of intervention, in the HSIS the offer of treatment did not translate one-for-one into Head Start participation. In particular, there were "no shows" (those who did not participate in HS despite being offered the slot) as well as "crossovers" (those who participated in HS despite not having been offered the slot). About 15 percent of the children in the control group ultimately enrolled in Head Start, while about 20 percent of treatment group children did not. In light of this cross-over (and incomplete take-up in the treatment group), the findings reported above 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). Arguing that the latter TOT is the more policy-relevant estimand, Ludwig & Phillips (2008) use the HSIS results to estimate the short-term effects of Head Start enrollment. Their back-of-the-envelope calculations suggest that the mean effects of enrolling in Head Start are approximately 30 percent larger than the effects of Head Start offers as reported in the final report (Puma et al. (2005), Puma et al. (2010)).

3.2 Sample, Means, and Balance Tests

We limit our analysis in this paper to data from the 3-year cohort. The main reason for doing so is that, as stated above, eligibility for inclusion in the experiment required children to be first time applicants to the Head Start program. Since the program serves children during the two years prior to Kindergarten, this restriction was not limiting for 3-year olds. However, limiting the sample in this way for 4-year olds may lead to external validity concerns without more information on why these 4-year olds have not participated in HS in the past. As might be expected, we find the 4-year old cohort to be potentially more disadvantaged compared to the 3-year old cohort, with a higher share of children living in Spanish speaking households and living with lower educated mothers. Perhaps more relevant given the current policy setting, increasingly 4-year olds have many other options besides Head Start while it is still an important source of preschool for 3-year olds and this is likely to stay the case even if the administrations' preschool for all program is implemented (Cascio & Schanzenbach (Forthcoming)).

The 3-year old sample consists of 2449 children, with 1464 in the treatment arm and 985 in the control arm. There are a variety of sampling and non-response weights available; we use the baseline weights which we augment in order to balance non-response as discussed below.[1]

Randomization occurred in the summer and fall of 2002, and we use the most recent version of the public-use data, which include outcomes through Grade 1. In our analysis, we refer to the main intervention year for our 3-year olds as the "Head Start year", which is followed by the "Age-4 year", Kindergarten year, and first grade year. (Thus, these labels

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

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, and cognitive and social emotional tests. Importantly, we can make use of baseline data from parent interviews as well as baseline tests given to the children in the Fall of 2002.

Table 1 summarizes the cognitive and non-cognitive measures that we use, as well as the years during which they are available. For cognitive outcomes, the tests we focus on examined language and literacy through tests of vocabulary, oral comprehension, pre-writing, and pre-reading as well as testing early numeracy and math skills. Among these tests, we primarily use the Peabody Picture Vocabulary Test (PPVT) for vocabulary knowledge and receptive language which is available at baseline and each year through Grade 1.[2] We also use a number of the Woodcock Johnson III battery of achievement tests. In particular, we use the WJ III measures of Pre-Academic Skills (a composite index measuring language and literacy) and Applied Problems (a composite index measuring math skills), which are available for the Head Start Year and through Grade 1.[3] Socio-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 theory) during the Kindergarten year. Socio-emotional skills are measured using the Pianta scale (Pianta (1992),Pianta (1996)) as well as using the Adjustment Scales for Preschool Intervention (ASPI). Each of these measures are indices of answers constructed from a series of questions to the teacher or parent, sometimes counting affirmative answers (ASPI) and sometimes counting responses on a five-point Likert scale (Pianta).

Table 2 reports summary statistics for child, parent, and caregiver variables at baseline (Fall 2002). The first column reports means for the control group, weighted using the baseline weights. 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

[2] 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 small number (67) of the children were not eligible for the PPVT III in English, 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.

[3] Some of the other WJIII tests included on the survey are not very continuously distributed and perhaps less suitable for our distributional methods. We have, however, examined many of them.

these 3-year olds are living with both biological parents. These 3-year olds are about evenly split across race/ethnic groups, with about 1/3 being non-Hispanic black, 1/3 being non-Hispanic white or other, and 1/3 being Hispanic. A little more than a quarter either took some tests in Spanish or speak Spanish at home (labeled "Spanish speaker" in the table). The bulk of the children have mothers or caregivers with at most a high school diploma or a GED, about 42% have mothers who were never married, and 26% are in medium or high risk families.[4] Also reflecting Head Start eligibility criteria, 11% of the 3-year olds are special needs.

As shown in the bottom of the table, the timing and presence of the baseline test assessment varied across children. Recall that these baseline tests were administered during the Fall of 2002. Unfortunately, it was infeasible to administer these pre-tests to all 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 developmental points). Table 2 shows that about a quarter of the control sample children were assessed before November, a third were assessed in November, another quarter were assessed after November, and 21% 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.[5]

The second column provides the difference in means of the variables between the treatment and control groups (using the baseline weights, thus it is labeled "unadjusted difference"). Consistent with the random assignment, almost none of the demographic characteristics are significantly different between the treatment and control (20 of the 25 variables are not statistically different at the 5% level), with these measures collected mostly at baseline. The results show that the treatment group contains statistically significantly lower shares of whites and those born to teen mothers, higher shares with special needs, and lower shares of children with caregivers aged 20-24.

[4] 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 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.

[5] 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 varies significantly between the treatment and control groups. The results show that children in the treatment group were somewhat more likely to be assessed earlier than the treatment children (before November) while those in the control were 13% 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. However, to account for the differences in the month of test assessment, as well as the other observables, in our results throughout the paper, we construct a weight that adjusts for the difference in selection into the sample (attrition). This is discussed below.

3.3 Weighting and Adjusting for Observables

Often we are interested in controlling for baseline characteristics of groups in settings such as this. In particular, in the context of educational experiments, it is typical to adjust for pre-treatment baseline test scores. Furthermore, we want to control for other baseline observables to adjust for the (relatively minor) imbalances between the treatment and control groups.

Given our interest in distributional estimates and in particular quantile treatment estimates (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. We use this approach in our context. In particular, we estimate the propensity score $\widehat{p}$ which is the predicted probability of being in the treatment group ($D_i = 1$) as a function of baseline characteristics using a logit. We then weight each observation by its inverse propensity-score weight:

$$\hat{\omega}_i \equiv D_i \cdot \frac{1}{\widehat{p}_i} + 1 - D_i \cdot \frac{1}{1 - \widehat{p}_i}. \tag{1}$$

In our logit model, we control for all of the child and caregiver variables in Table 2.

Additionally, to richly control for baseline test scores, we assign 2002 PPVT deciles for each of the four assessment month groupings in Table 2. We use the observations with imputed baseline values of PPVT but control for not having a baseline assessment in the propensity score. The children in this category are treated as having a separate assessment month (missing). We also include a full set of fixed effects for the Head Start center to which the child applied. Note that the fixed effects for the center of application help control for the fact that assignment occurred at the center level and also adjust for important potential geographic heterogeneity.[6]

The results of the propensity score weighting model are available in request. Of the 39 dummies for assessment month group by decile of test score[7] only 2 are significant at the 5% level, and there is no consistent pattern in the sign of the coefficients among these baseline test score identifiers. Looking at the overlap of the propensity scores between the treatment and controls, they look quite similar with trivial non-overlap.[8]

The final column of Table 2 presents the "adjusted" difference in means between the treatment and control. This is the difference in the weighted means, using our inverse propensity score weights. As expected, the treatment and control samples look quite balanced after adjusting for observable differences. Only 2 of 28 variables are statistically significantly different at the 5% level. In the remainder of the paper, we use the inverse propensity score weights in 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-to-one into Head Start participation (either at the center of random assignment or at another Head Start center). Table 3 provides a summary of the most common child care settings children experienced at the end of the Head Start year,

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

[7] There are four assessment-month groups and 10 deciles for a total of 40 decile x assessment-month groups. With the constant term, we are left with 39 coefficients.

[8] We also explored a number of alternative propensity-score adjustments, and our findings are robust to these.

separately for the treatment and control groups. As is evident in the table, the HSIS leads to a large change in child care settings. 82% of those in the treatment group are in Head Start in 2003 compared to only 15% in the control. The biggest change comes from those in parent or relative care—only 10% of children in the treatment group are in parent or relative care compared to over half in the control group.

However, it is also clear from Table 3 that a sizeable increase in Head Start use comes from an overall decline in the use of other non-HS centers. Only 7% of the treatment group is in other center-based care in Spring 2003 (at the end of the Head Start year), compared to 25% of the control group. 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 which case Table 3 shows that the treatment led to a 68 percentage point increase in Head Start attendance. If, however, one considers use of any center-based care—Head Start or other centers—to be the first stage, then the first stage treatment is smaller but still large at 50 percentage points. Overall, the results in table 3 are quite useful for learning about the relevant counterfactual.

Table 3 also presents the child care setting in the second, Age-4 year. As discussed above, the HSIS treatment is a one year long treatment. As one might expect given this disadvantaged population as well as the aforementioned growing public options children aged 4, there is a significant amount of change in care arrangements between ages 3 and 4 for these children. Additionally, the differences in outcomes narrow between the treatment and control, "blunting" the treatment. At the end of the Age-4 year, 63% of the treatment group is in Head Start compared to almost half of the controls. Again, considering the first stage outcome to be use of any center-based care, then the treatment group value is only 3 percentage points larger than the corresponding control value by the end of the second year.

Before examining impacts of HS across the distribution of outcomes, we first present mean treatment effects on our main cognitive outcome, PPVT. Table 4 contains mean treatment effects and control group means using the baseline weights (in the left half of the table) and using our inverse propensity-score weights (in the right right of the table). For each testing period, we show results for two different outcome variables: Test scores and the probability that the test score is unavailable (unavailable or imputed at baseline, unavailable in the later assessments). Several of the test scores are imputed for the vast bulk missing them at baseline.

The table shows that the baseline average test score for PPVT for the 3-year olds was 231 with a standard deviation of 39 (both measures are calculated from the control group). The table also shows that there are at most trivial differences, on average, in the baseline scores across the treatment and control groups, using either set of weights. This is, of course, important as it indicates balance in the randomization.[9]

As discussed earlier, in the context of Table 2, the lower half of the table shows that without the use of inverse propensity-score weights, there are statistically significant differences in the share of observations with missing (imputed or simply not administered) baseline PPVT scores. However, after inverse propensity-score weights are applied, the probability of the test score being missing (imputed or not administered) is balanced (the difference in the probability of it being missing is 0.03 after adjustment). In fact, our inverse propensity-score weights balance this PPVT test non-response in each year from 2003–2006 as well, with small and statistically insignificant differences in non-response in those years.

Next we turn to mean treatment effects in Spring 2003 (Head Start year) after the first year of the experiment. The HSIS led to a statistically significant effect of around 7.4 points or 0.19 standard deviations using the baseline weights. The results are nearly identical using the inverse propensity-score weights. However, the results for the years after the Head Start year show fade-out in the mean impacts, suggesting no significant mean effect of being assigned to the Head Start group. These results are little changed by the use of inverse propensity-score weights.

These mean effects suggest little if any effects of the program after the first year. However, there are developmental theories which suggest that these non-effects could reflect offsetting positive and negative effects for different groups. To explore this, we move to our main results on the effects of Head Start on the distribution of outcomes.

[9] These baseline scores exclude observations with missing values.

5 Empirical approach

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

To define the quantile treatment effect, first let $F(y)$ be the cumulative distribution function (CDF) of Y. 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 (e.g., $y_{0.5}$ is the median). Further, let F_1 be the CDF if $D = 1$ and F_0 be the CDF if $D = 0$. The quantile treatment effects (QTE) estimate is then 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, $y_q = y_{q1} - y_{q0}$. In an experimental setting such as the HSIS, the QTE is the estimator of this difference in the quantiles of the two marginal distributions. For example, given random assignment, one can consistently estimate the QTE at the median (0.50 quantile) simply by subtracting the control group's sample median from the treatment group's sample median.[10] The only adjustment we make to this simple setup is to use the inverse

[10] 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).

propensity score weights to account for observables, as discussed above.

5.1 Inference

The HSIS data are not random samples of children across centers, but rather sampled in a complicated fashion. For inference, we are using bootstrapping, with centers being randomly sampled with replacement (with all children from sampled centers appearing in the data whenever a center is sample). We construct confidence intervals using the percentile method. With 999 bootstrap resamples, a 95% 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.[11]

6 Main Results for the Heterogeneous Effects of HSIS

Here we present our main results for examining the heterogeneous impacts of an offer of Head Start. We begin by focusing on the Head Start year and move on to examine the impacts through Grade 1. We offer two sets of evidence—quantile treatment effects and mean impacts by subgroup.

6.1 QTE Results for the Head Start year, full sample

We begin by analyzing PPVT scores for the full 3-year old sample. As outlined 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.

Before examining the QTE for the Head Start year and beyond, we provide an additional balance test. In particular, in Figure 1, we reported the QTE for the PPVT at baseline.

[11] We plan to later produce uniform confidence intervals.

Notably, the confidence intervals almost all included 0; thus, the weighting has ensured balance across the two groups in the baseline test distribution. (In results not shown, there is only modest imbalance in the baseline QTE before controlling for observables using the inverse propensity score weights.)

Moving on, Figure 2 presents the QTE for the Head Start year, using tests in Spring 2003. The solid line gives the QTE and the long dashed lines give the 90% bootstrapped confidence intervals at each percentile index. The horizontal dotted line presents the average treatment effect (as reported in Table 4). First, note that the offer of a HS slot improves cognitive skills throughout the distribution. Second, the gains associated with treatment are largest at the bottom of the test score distribution. The estimates at lower end of the skill distribution are very large, ranging from 34 for quantiles 1–3, to 17 for quantile 13, and to 3 at quantile 43 (which is the last QTE which is statistically significantly different from zero at the 10% level). These effects are also notably substantial relative to the control group standard deviation (38, see Table 4).

The QTE capture the impact of the treatment on the distribution of outcomes. For example, Figure 2 measures the impact of being offered Head Start on, for example, the median, 25th percentile, and 75th percentile of the PPVT test score distribution. One limitation of the QTE estimator is that QTE at quantile q need not equal the treatment effect for an individual located at quantile q of the control group. While our hypotheses are about effects for individuals of various cognitive abilities, our QTE results will represent effects on the whole distribution. Only with further assumptions, which are perhaps undesirable, can we conclude that the QTE is the treatment effect for a particular individual.[12]

An alternative to the QTE is to use the baseline score to more explicitly "model" the heterogeneity in treatment effects across the distribution of baseline skills. In particular, in Figure 3A, following Duflo, Dupas & Kremer (2011), we estimate local linear regressions of the 2003 PPVT as a function of the baseline score, estimated separately for the treatment and control groups. In Figure 3B, we plot the treatment-minus-control differences from these local

[12] One assumption that allows for treatment heterogeneity is rank preservation (e.g., Heckman, Smith & Clements (1997)), under which a person's location in the distribution is unchanged by the treatment. In this case, the QTE are the same as the distribution of individual treatment effects. Regardless the QTE are useful for assessing overall population gaps.

linear regressions (along with the bootstrapped confidence intervals). This approach yields very similar findings to the QTE qualitatively—positive gains throughout the cognitive skill distribution with the largest gains at the bottom of the distribution. We prefer the QTE in large part because of the prevalence of missing baseline test scores (disproportionately imputed for the controls) as well as the variation in month of assessment for this test (leaking into the treatment period). Furthermore, for many policy interventions such as Head Start, knowing about effects on the distribution—the QTE—are of substantial interest, and the rhetoric about closing test gaps is about ex-post (of some reform or treatment) scores, not ex-ante ones. The QTE captures precisely these ex-post effects while the local linear captures the ex-ante.

Before resuming our analysis of the QTE, the local linear regression in the baseline score also presents a useful way to examine the possible differences in the first stage (Head Start participation as a function of the offer of a slot) across the distribution of skills. Figure 4 plots Head Start participation as a function of the baseline score, using this local linear regression approach, again separately for the treatment and control groups. Both lines in this figure are very flat, showing remarkable balance in the first stage effect of an offer on participation across the baseline skills distribution.

Figure 5 presents the QTE for two of the Woodcock Johnson III battery of achievement tests, including the composite Pre-Academic Skills measure (Figure 5A), which measures early literacy; and the composite Applied Problems measure (Figure 5B), which measures early numeracy. The results largely echo the findings for the PPVT. The effect of an offer of Head Start is positive throughout the Pre-Academic Skills distribution, with modestly larger and statistically significant effects at the bottom. The results for Applied Problems show a large concentration of gains at the bottom of the distribution which is nonetheless insignficant. Generally, the estimates are much less precise for the WJIII tests than for the PPVT. That said, having a measure of the impact of a HS offer on early numeracy is particularly useful given Duncan, Dowsett, Claessens, Magnuson, Huston, Klebanov, , Pagani, Feinstein, Engel, Brooks-Gunn, Sexton, Duckworth & Japel's (2007) results showing the importance of early numeracy in predicting long term achievement.

6.2 Results for Subgroups in the Head Start Year

Another dimension of heterogeneity is to examine differences across subgroups of the HSIS population. In particular, here we present results by race, language, and and terciles of the baseline PPVT. In Table 5, we present mean treatment effects for these subgroups for PPVT in the Head Start year (along with the control group means). This table shows dramatic differences across groups, with larger mean treatment effects for Hispanics (11.5 or 0.29 standard deviations), Spanish speakers (15.0 or 0.47 standard deviations) and those in the bottom tercile of the baseline test score (11.2 or 0.33 standard deviations). These are very large effects.

One possible explanation for the differences in treatment effects across groups is either a difference in the probability of Head Start take-up, or a difference in the counterfactual child care settings across groups. Table 6, however, shows that there are relatively small differences across the groups in the "first stage" of Head Start take-up, or in the probability of the child being in an other center or home based care (typically with parents or relatives).

We can explore the differences across subgroups by estimating QTE separately for each subgroup (these are referred to as "conditional QTEs" in the literature as they are conditioned on being in the subgroup). 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 QTEs for Spanish speakers are very large (above 10, which is a third of a standard deviation) and extend through the 60th percentile.

While these conditional QTEs 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 test score distribution. Given that these conditional QTE plot each line on a common percentile scale (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 are the percentiles of the full sample of the control group). These "translated" QTE are presented in Figure 6B. This shows that the subgroups' QTE are put on the same absolute scale, the differences between groups become attenuated but are still notable. We see similarly-sized, very large gains throughout the bottom decile of PPVT scores for both groups; yet we see larger effects of the treatment on Spanish speakers (compared to English speakers) throughout the rest of the distribution.[13]

Figures 7A and 7B provide similar analyses by race: non-Hispanic whites, non-Hispanic blacks, and Hispanics. The results are quite dramatic–while the conditional QTEs (Figure 7A) show widespread gains for Hispanics, relative to the two non-Hispanic groups, on a common scale the translated QTE (Figure 7B) shows much more similarity across the distribution.

Taking these results as a whole, the effects of a HS offer on cognitive outcomes in the Head Start year reveal compelling evidence of heterogeneity. Based on the analysis across the distribution as well as across demographic groups, we find evidence in favor of the "compensatory" theory. That is, we find larger gains in the lower end of the skill distribution. 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 ratings from direct classroom observation (Arnett and ECERS-R score).

[13] Essentially what this is doing is stretching the conditional distribution in some places and and shrinking it in others in order to put them both on the same scale. We use the PPVT score of the control group (for the full sample) at each percentile of the full sample as the anchor. We take 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 QTE for subgroup A would be relocated at the overall control group's 25th percentile. To give some guidance on the amount of stretching this produces, in the translated graphs, there is a symbol at each 10th percentile of the subgroup's own distribution. Thus, one can see that in the lower half of the (overall) PPVT distribution the symbols for deciles for the Spanish speakers are more compressed while in the upper half the opposite is true.

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. Returning to the PPVT, Figure 8A shows the QTE for each year: Head Start year (2003), Age-4 year (2004), Kindergarten year (2005), and first grade year (2006). The positive (and significant, although not shown here) effects at the bottom remain for the lower end of the distribution through both of the preschool years. However, once this cohort transitions into elementary school, the gains substantially fade. Figure 8B zeros in on the last year in our data, grade 1, showing no significant effects of having been assigned to an offer of Head Start.

Figures 9A and 9B present results for grade 1 for the WJIII Pre-Academic Skills and Applied Problems composites. The results for Pre-Academic Skills have faded out by grade 1. 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 10A presents the translated QTE for Spanish versus English speakers for the first grade year. Here we see encouraging evidence of persistent gains for Spanish speakers, throughout the bottom half of the (overall) distribution. These gains, at 8–10 points, are substantial, measuring 0.2 to 0.25 standard deviations. Additionally, Table 10B presents the translated QTE for the three terciles of baseline scores, for Grade 1. These results suggest lasting gains for those in the lowest tercile of baseline scores.

7 Discussion

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

Given the broad discussion of “compensatory” versus “skills-beget-skills” theories of education, our work provides new and compelling evidence in favor of the compensatory theory. In particular, those with low baseline scores, and those with limited English, gain the most from the intervention. Our analysis shows that these differences can not be explained by differences in take-up of the program.

The results also show that once the children enter elementary school these gains diminish substantially, for the population as a whole. However, if we focus on those with the greatest deficits entering preschool, we find persistent effects through Grade 1. For example, we showed this for Spanish speakers as well as for those in the lowest tercile of baseline scores.

What can we conclude from these results? First, the gains in preschool may not persist if the elementary schooling environment is not of high quality. We are limited in what we can say about this given that we do not observe much about the schools the study participants attend. Second, the Head Start teachers may be teaching to some proficiency standards (e.g., knowing the ABCs, counting to 10) and the quality of the settings are insufficient to achieve gains beyond that point. As described in Cascio & Schanzenbach (Forthcoming), 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)).

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

Second, many have argued that preschool (or other investments during this crucial period) may lead to improvements in non-cognitive outcomes and these may facilitate gains in elementary school and beyond. This is testable in the HSIS. The data contain a host of non-cognitive or socio-emotional measures; here we focus on the Pianta scales of student-teacher and child-parent relations and the Adjustment Scales for Preschool Intervention (a measure of emotional and behavioral adjustment to preschool). (Recall we only have the teacher reports for everyone when they are all in school, thus our reliance on parent reports in the Head Start year.) In Table 7, we present mean treatment effects for these socio-emotional outcomes. The table presents effects of the Head Start offer on both parent-reported measures (during the Head Start and first grade years) as well as teacher-reported measures during first grade, when all the children should be in a group setting. The measures have been standardized to a mean 0 and standard deviation of 1 (although the signs are not aligned such that a higher score is positive). The table presents the mean treatment effects and standard errors for each year. During the Head Start year, there is a statistically significant *decrease* in hyperactivity, but the seven other point estimates are generally favorable but insignificant. The first grade effects are uniformly small and insignificant for the parent-reported measures, and particularly for the teacher-reported measures. While there are a few positive findings, the overall finding is one of very small and statistically insignificant effects in the socio-emotional domain.

Finally, Cunha & Heckman (2010) talk about early childhood being both a sensitive and critical period for child development. Duncan et al. (2007) argue that early cognitive scores are predictive of later cognitive scores. Similarly, Magnuson, Ruhm & Waldfogel (2007) provide evidence of what they call "sleeper" effects of preschool on cognitive outcomes. While we can not test these ideas in the HSIS, we explore these issues in Deming's (2009) NLSY sample. In particular, we take Deming's sample and outcomes in young adulthood, but do not restrict ourselves to the siblings he studies. Using this observational data, we regress these outcomes on cognitive test scores measured during the preschool years (labeled PPVT 3–5) as well as on another cognitive test score taken during middle elementary school (labeled PIAT 9–11).[14] We restrict the sample to individuals in the top quartile of predicted Head Start attendance (to mimic our experimental sample).[15] We are essentially looking to see if PPVT during preschool provides predictive power for later outcomes in the presence of a measure of cognitive score at ages 9-11. We find some predictive power of preschool cognitive skills in the expected directions—higher scores mean better outcomes—but only

[14] PIAT is a math assessment.

[15] We predict Head Start participation using a detailed list of demographics, family characteristics, mother's AFQT, birth order, and so on. We then use these estimates to predict Head Start participation. We limit our sample to individuals in the top quarter of this predicted index.

two of these effects (those on repeating a grade, being idle) reach statistical significance. The magnitude of the coefficients on these early PPVT scores, compared to those on the later cognitive measures, are nonetheless suggestive of a role for early cognitive skills in young adult outcomes.

8 Conclusion

In this study, we provide a comprehensive evaluation of the first national randomized experiment of Head Start. We focus on the 3-year old cohort and examine impacts on cognitive and non-cognitive outcome in the Head Start year and through Grade 1. We find that the offer of a Head Start slot leads to large and statistically significant gains in cognitive skills in the preschool period. However, once the children enter school, the overall cognitive gains fade out. Importantly, we find that some groups including Spanish speakers as well as those entering preschool with low baseline scores, have cognitive gains that start out being large during the Head Start year and persist through 1st grade. We find little effect of the experiment on non-cognitive outcomes.

These results still provide an incomplete picture of the potential benefits of Head Start. Prior work finds that despite a fadeout in effects on cognitive outcomes, there are positive effects on the long term on labor market and human capital outcomes (Deming, 2009). Future work will hopefully follow these participants into adulthood.

References

Anderson, M. (2008), 'Multiple inference and gender differences in the effects of early inteven-tion: A reevaluation of the Abecedarian, Perry Preschool and Early Training projects', *Journal of the American Statistical Association* **103**(484), 1481–1495.

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

Belfield, C. R., Nores, M., Barnett, W. S. & Schweinhart, L. J. (2006), 'The High/Scope Perry Preschool Program: Cost-benefit analysis using data from the age-40 followup.', *Journal of Human Resources* **16**, 162–190.

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.

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

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 monitoriing 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.

Dynarski, S., Hyman, J. & Schanzenbach, D. (2011), Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion, Working Paper 17533, NBER.

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

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

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.

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

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.

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., Ruhm, C. & Waldfogel, J. (2007), 'The persistence of preschool effects: Do subsequent classroom experiences matter?', *Early Childhood Research Quarterly* **22**(1), 18–38.

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

Masse, L. N. & Barnett, W. S. (2007), 'Comparative benefit-cost analysis of the Abecedarian Program and its policy implications', *Economics of Education Review* **26**, 113–125.

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, Working paper. 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, Working Paper OPRE Report # 2012-45. Prepared for USDHHS, ACF.

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

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

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., Montie, J., Xiang, Z., Barnett, W., Belfield, C. & Nores, M. (2005), *Lifetime effects: The High/Scope Perry preschool study through age 40*, High/Scope Press, Ypsilanti, MI.

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

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

Table 1: Tests administered in the HSIS data, by year

	Head Start year			
	Age 3	Age 4	K	1st Grade
Language, Literacy				
PPVT	X	X	X	X
Pre-Academic skills (WJIII)	X	X	X	X
Oral Comprehension (WJIII)	X	X	X	X
Math				
Applied Problems (WJIII)	X	X	X	X
Socio-Emotional parent reports				
Aggressive behavior (ASPI)	X	X	X	X
Hyperactive (ASPI)	X	X	X	X
Withdrawn (ASPI)	X	X	X	X
Social competencies (ASPI)	X	X	X	X
Social skills/pos. learning (ASPI)	X	X	X	X
Conflict (Pianta)	X	X	X	X
Closeness (Pianta)	X	X	X	X
Positive relationships (Pianta)	X	X	X	X
Socio-Emotional 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 int. (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

Source: HSIS Final Report.

Table 2: Summary statistics at baseline

	Control mean	Unadjusted T-C	Adjusted T-C
Child Characteristics			
Non-Hispanic white	0.31	-0.041**	-0.001
Non-Hispanic black	0.36	0.025	0.001
Hispanic	0.33	0.016	0.000
Female	0.49	0.018	0.017
Spanish speaker	0.26	0.012	0.016
Child risk index, low	0.74	-0.025	0.044
Child risk index, medium	0.17	0.012	-0.019
Child risk index, high	0.09	0.013	-0.025**
Special needs	0.11	0.033**	0.008
Lives with both biological parents	0.48	0.008	0.024
Lives in urban area	0.83	0.011	0.006
Mother/caregiver characteristics			
White	0.33	-0.044**	-0.008
Black	0.35	0.027	0.010
Hispanic	0.31	0.017	-0.001
Teen mother	0.16	-0.040***	0.023
Less than high school	0.35	-0.023	0.001
High school / GED	0.36	0.016	-0.024
More than high school	0.29	0.006	0.023
Married	0.43	-0.007	0.022
Divorced	0.15	-0.001	0.017
Never married	0.42	0.007	-0.038
Age 20-24	0.28	-0.045**	0.036
Age 25-29	0.31	0.006	0.021
Age 20-29	0.27	0.015	-0.016
Age 40+	0.11	0.015	-0.042***
Fall 2002 test month			
Before November	0.22	0.142***	0.015
November	0.33	0.040**	-0.022
After November	0.25	-0.048***	-0.025
No fall assessment (imputed)	0.21	-0.134***	0.031
Number of observations	2,378		

Notes: Table reports means for baseline characteristics for the 3-year old cohort control group. Column 2 provides the unadjusted difference in means, using the baseline weights. Column 3 provides the adjusted difference in means, using the propensity score weights. *, **, and *** denote significant at the 10, 5, and 1 percent levels.

Table 3: Child care setting, by treatment and control status

	Head Start year, spring 2003		Age 4 year, spring 2004	
	Treatment Group	Control Group	Treatment Group	Control Group
Head Start	0.82	0.15	0.63	0.49
Other Center	0.07	0.25	0.26	0.37
Non-center	0.01	0.07	0.02	0.02
Parent/Relative	0.10	0.54	0.09	0.12

Notes: Means for age-3 cohort, spring 2003 and spring 2004. The data on modal child care center (center where the child spent the most time) come from the parent reports and exclude missing values. Statistics weighted using inverse propensity score weights.

Table 4: Mean treatment effects for PPVT, by year

	Baseline weights				Inv. p-score weights			
	Control mean	[SD]	Mean treatment effect	(SE)	Control mean	[SD]	Mean treatment effect	(SE)
A. PPVT								
Baseline, fall 2002	231	[39]	-0.73	(2.24)	231	[39]	-0.00	(1.91)
Head Start year, spring 2003	251	[38]	7.37***	(2.32)	251	[38]	7.20***	(1.88)
Age 4 year, spring 2004	298	[41]	3.09	(2.39)	298	[41]	2.88	(2.00)
Kindergarten, spring 2005	340	[28]	0.89	(1.83)	340	[28]	0.21	(1.52)
Grade 1, spring 2006	357	[30]	3.29	(1.87)	357	[30]	2.01	(1.55)
B. PPVT missing or imputed								
Baseline, fall 2002	0.28		-0.12***	(0.02)	0.22		0.03	(0.02)
Head Start year, spring 2003	0.22		-0.10***	(0.02)	0.19		-0.01	(0.02)
Age 4 year, spring 2004	0.21		-0.07***	(0.02)	0.18		0.01	(0.02)
Kindergarten, spring 2005	0.24		-0.06**	(0.02)	0.23		0.01	(0.02)
Grade 1, spring 2006	0.26		-0.07***	(0.02)	0.25		0.00	(0.02)

Notes: Table reports control group means and mean treatment effects for the 3-year old cohort in the HSIS. The first set of columns uses the baseline weight and the second set of columns uses the propensity score weights. Panel A reports the mean treatment effects on PPVT scores and Panel B reports the prevalence of a score not having been administered (missing) or being imputed (baseline only). *, **, and *** denote significant at the 10, 5, and 1 percent levels.

Table 5: Mean treatment effects for subgroups, PPVT in Head Start year (spring 2003)

	Control mean	[SD]	Mean treatment effect	(SE)
Hispanic	234	[39]	11.5***	(3.46)
Non-Hispanic black	250	[32]	5.2**	(2.61)
Non-Hispanic white	268	[34]	6.5**	(2.24)
Spanish speaker	223	[32]	15.0***	(3.32)
English speaker	261	[34]	4.9**	(1.99)
Baseline PPVT, top tercile	274	[36]	7.5**	(3.25)
Baseline PPVT, middle tercile	251	[30]	2.2	(2.60)
Baseline PPVT, bottom tercile	229	[33]	11.2***	(2.83)

Notes: Table reports control group means and mean treatment effects for the 3-year old cohort in the HSIS for different demographic subgroups. All calculations use the inverse propensity score weights. *, **, and *** denote significant at the 10, 5, and 1 percent levels.

Table 6: Effect of Head Start offer on child care arrangements, Head Start year (spring 2003) by subgroup

		Head Start	Other Center	Parent/rel. Other
Hispanic	1st stage effect	0.576	-0.395	-0.181
	Control mean	*0.117*	*0.537*	*0.345*
Non-Hispanic black	1st stage effect	0.620	-0.413	-0.207
	Control mean	*0.145*	*0.464*	*0.391*
Non-Hispanic white	1st stage effect	0.607	-0.438	-0.169
	Control mean	*0.101*	*0.536*	*0.363*
Spanish speaker	1st stage effect	0.584	-0.398	-0.186
	Control mean	*0.134*	*0.510*	*0.356*
English speaker	1st stage effect	0.608	-0.421	-0.186
	Control mean	*0.118*	*0.511*	*0.371*
Baseline PPVT, top tercile	1st stage effect	0.611	-0.373	-0.238
	Control mean	*0.125*	*0.451*	*0.425*
Baseline PPVT, middle tercile	1st stage effect	0.544	-0.404	-0.141
	Control mean	*0.144*	*0.509*	*0.347*
Baseline PPVT, bottom tercile	1st stage effect	0.638	-0.470	-0.168
	Control mean	*0.101*	*0.580*	*0.319*

Notes: Table reports control group means and mean treatment effects for child care arrangements in the Head Start year (spring of 2003). The sample includes the 3-year old cohort in the HSIS, and the calculations are by subgroup. All calculations use the inverse propensity score weights.

Table 7: Mean treatment effects for socio-emotional outcomes

		Head Start year (Spr. 2003)		Grade 1 (Spr. 2006)	
Parent Reports	ASPI Aggressive	-0.073	(0.045)	-0.068	(0.045)
	ASPI Hyperactive	-0.197***	(0.046)	-0.077	(0.045)
	ASPI Social Competencies	-0.008	(0.048)	0.047	(0.044)
	ASPI Social Skills	0.011	(0.047)	0.033	(0.047)
	ASPI Withdrawn	0.017	(0.045)	-0.018	(0.047)
	Pianta Conflict	-0.021	(0.044)	-0.090	(0.047)
	Pianta Closeness	0.078	(0.041)	0.060	(0.044)
	Pianta Pos Rel	0.042	(0.044)	0.089	(0.047)
Teacher Reports	Closeness (Pianta)			0.011	(0.050)
	Conflict (Pianta)			-0.002	(0.049)
	Aggressive (ASPI)			-0.019	(0.050)
	Oppositional (ASPI)			0.034	(0.048)
	Inattentive (ASPI)			0.002	(0.049)
	Shy/socially reticent (ASPI)			0.044	(0.050)
	Withdrawn/low energy (ASPI)			0.074	(0.049)
	Combined ASPI index- negativity			0.006	(0.044)
	Combined ASPI index - shy			0.060	(0.044)
	Combined ASPI index - interactive			0.013	(0.042)

Notes: Table reports mean treatment effects for socio-emotional outcomes. Teacher reports are not available for all children until Kindergarten. All outcomes are standardized by subtracting off the control mean and dividing the result by the control standard deviation. Therefore all treatment effects are in standard deviation units. The "combined measures" are equally weighted averages of individual submeasures, also in standard deviation units. Sample includes the 3-year old cohort in the HSIS. All calculations use the inverse propensity score weights. *, **, and *** denote significant at the 10, 5, and 1 percent levels.

Table 8: Effects of test scores in early and middle childhood on adult outcomes, children of NLSY

	Repeat	Learning Disability	Idle	Crime	Poor Health	HS Grad	HS Grad (no ged)	Some College	Teen Preg.
PPVT 3-5	-0.003**	0.00004	-0.002*	-0.002	-0.0004	0.0004	0.002	0.001	-0.0009
	(0.001)	(0.0005)	(0.0008)	(0.001)	(0.0009)	(0.001)	(0.001)	(0.001)	(0.001)
	-0.009	*0.001*	*-0.010*	*-0.007*	*-0.004*	*0.001*	*0.004*	*0.006*	*-0.004*
PIAT 9-11	-0.004***	-0.002***	-0.002*	0.00001	0.0001	0.002**	0.002*	0.002**	-0.001
	(0.001)	(0.0004)	(0.001)	(0.0009)	(0.0007)	(0.001)	(0.001)	(0.001)	(0.001)
	-0.011	*-0.038*	*-0.009*	*-0.015*	*-0.038*	*0.003*	*0.004*	*0.012*	*-0.018*
Mean Dep Var	0.380	0.052	0.204	0.274	0.105	0.607	0.568	0.169	0.226
Observations	596	615	599	599	599	599	549	599	599
R-squared	0.161	0.077	0.093	0.061	0.039	0.118	0.128	0.080	0.081

Notes: Calculations from NLSY sample from Deming (2009) but using singletons and siblings. Sample is further limited to those in the top quartile of predicted Head Start participation. Results weighted. We predict Head Start participation using demographics, family characteristics, mother's AFQT, and birth order. *, **, and *** denote significant at the 10, 5, and 1 percent levels.

Figure 1: QTE for PPVT scores at baseline

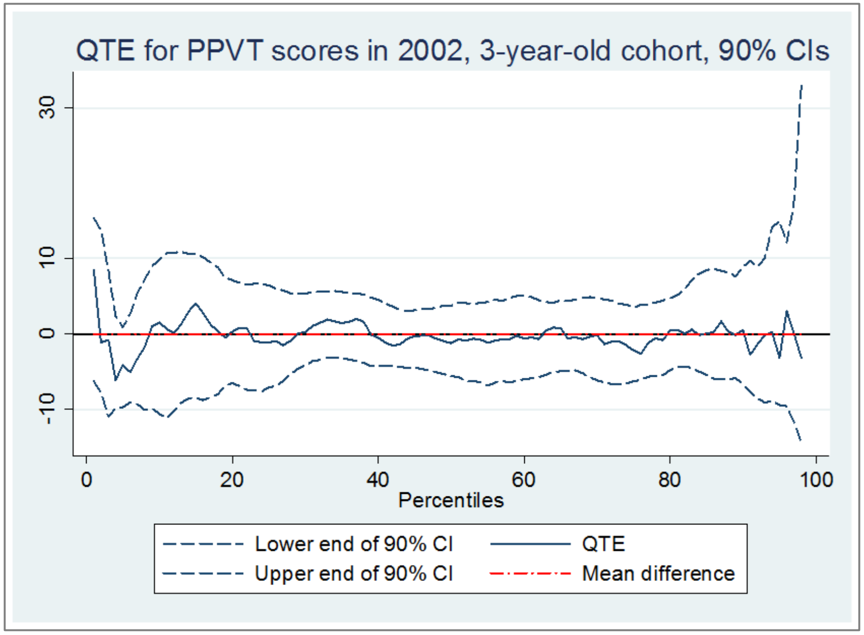
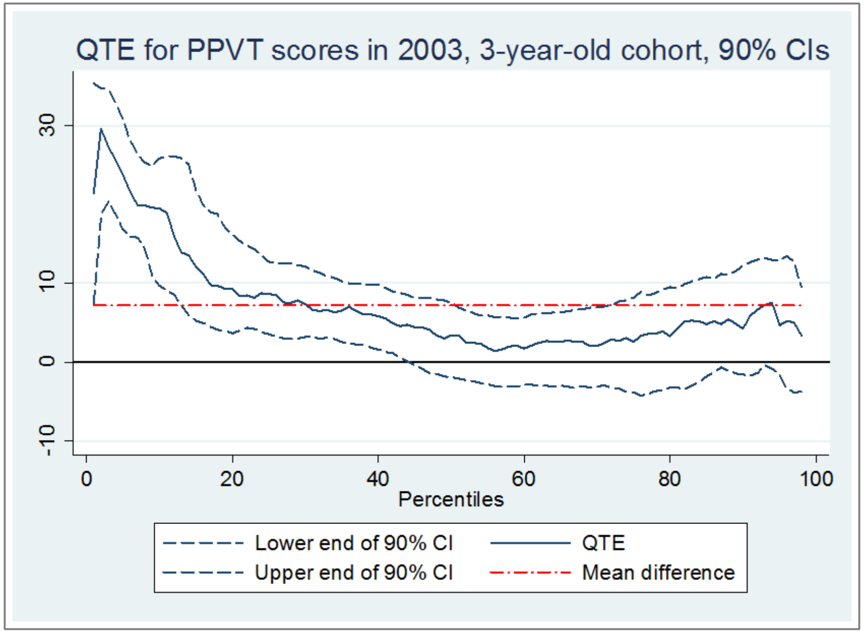


Figure 2: QTE for PPVT scores at end of Head Start year



Notes: Calculations based on 3-year old cohort of HSIS using inverse propensity score weights. 90% confidence intervals obtained by bootstrapping Head Start center.

Figure 3A: Local linear estimates of 2003 PPVT on baseline score, by treatment and control

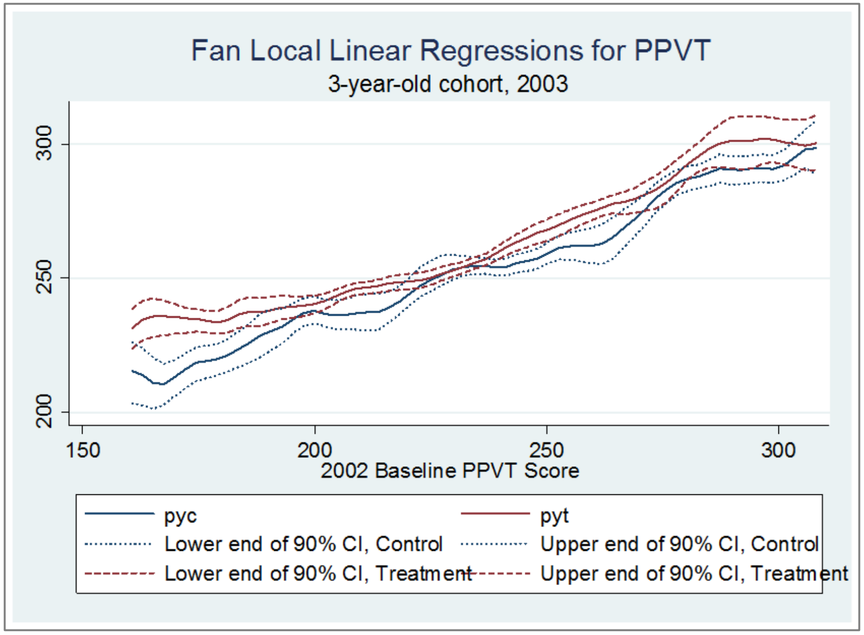
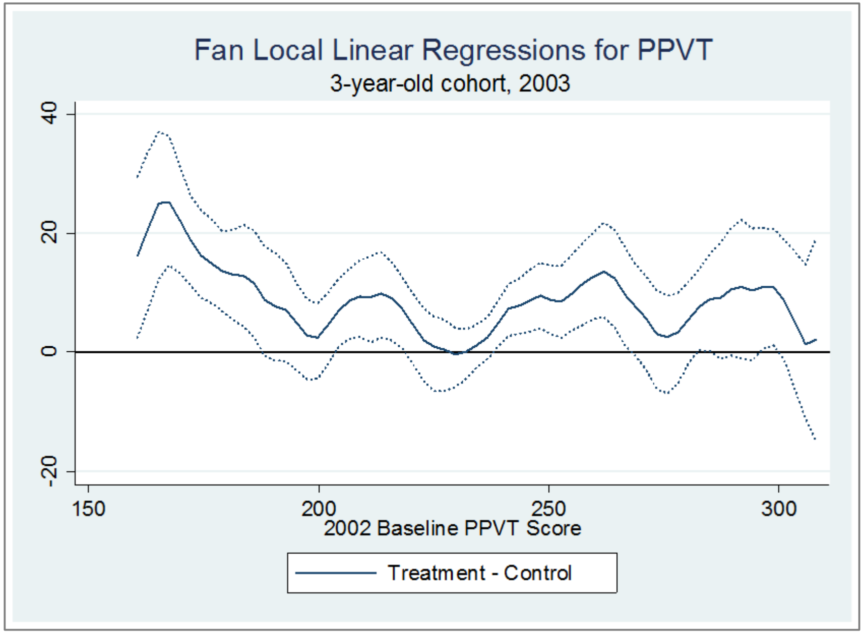
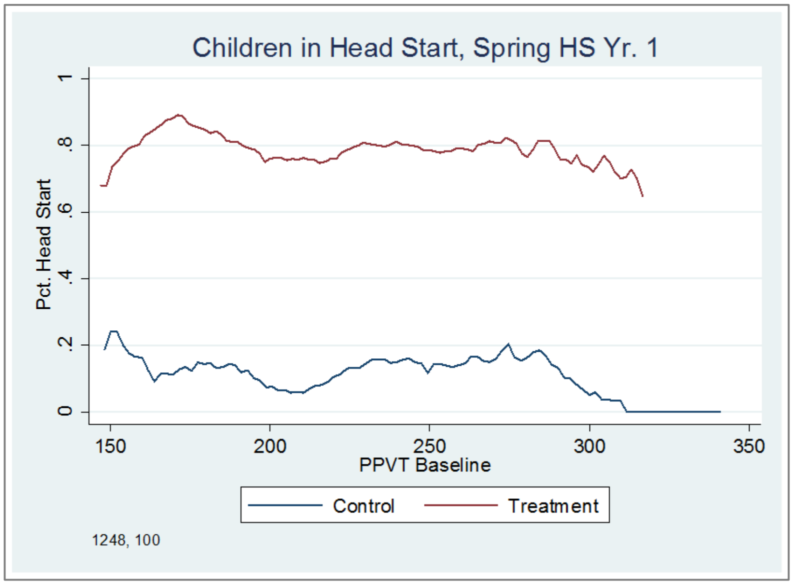


Figure 3B: Treatment-control difference in local linear estimates for 2003 PPVT on baseline score



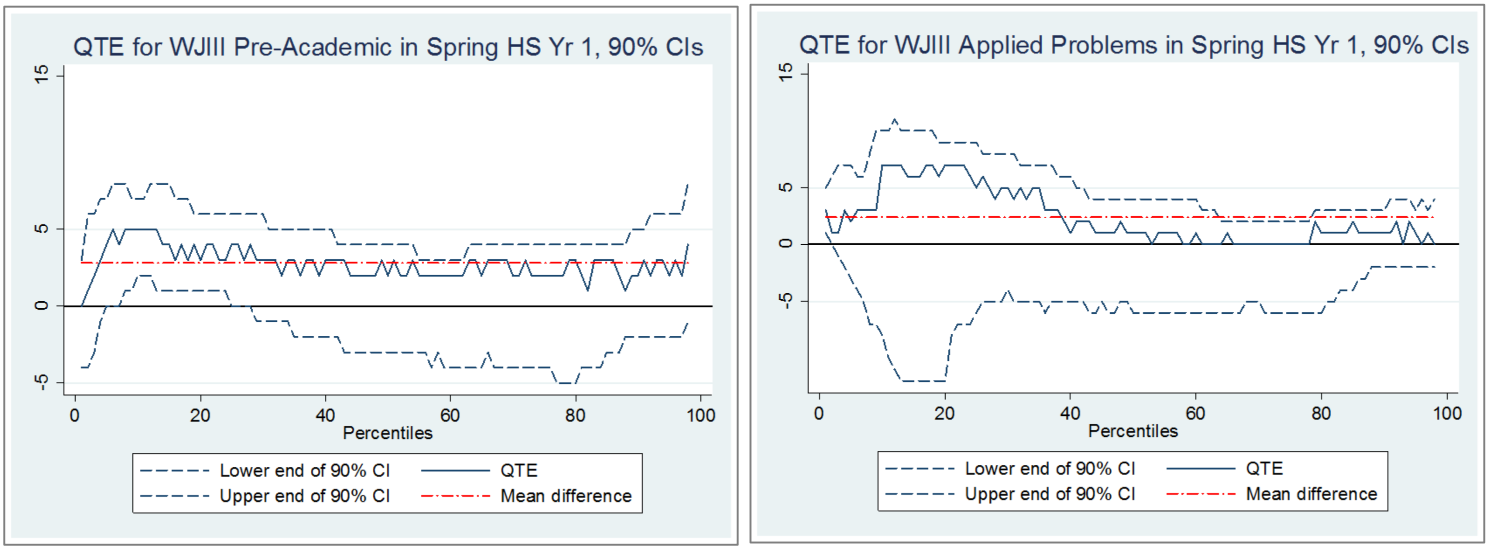
Notes: Calculations based on 3-year old cohort of HSIS using inverse propensity score weights. Top figure provides local linear regression of 2003 PPVT on baseline PPVT separately for treatment and control group. Bottom figure plots the treatment-control difference of the means in the above graph, by baseline score, along with 90% confidence interval.

Figure 4: Local linear regression of spring 2003 Head Start participation on baseline scores, by treatment and control



Notes: Calculations based on 3-year old cohort of HSIS using inverse propensity score weights. Head Start participation is from the parent interview in spring 2003 (end of Head Start year).

Figure 5: QTE for Woodcock Johnson III tests, spring 2003



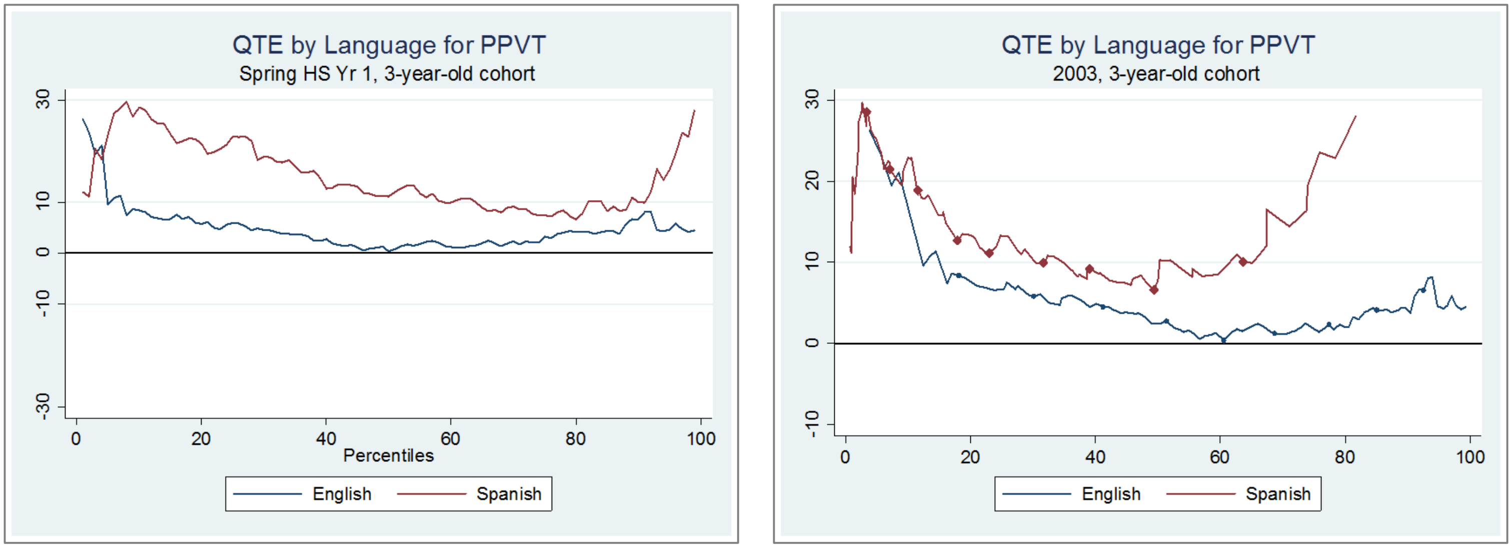
(A) Pre-Academic Skills

(B) Applied Problems

Notes: Calculations based on 3-year old cohort of HSIS using inverse propensity score weights. 90% confidence intervals obtained by bootstrapping Head Start center.

Figure 6: QTE for PPVT scores in spring 2003, for English and Spanish speakers

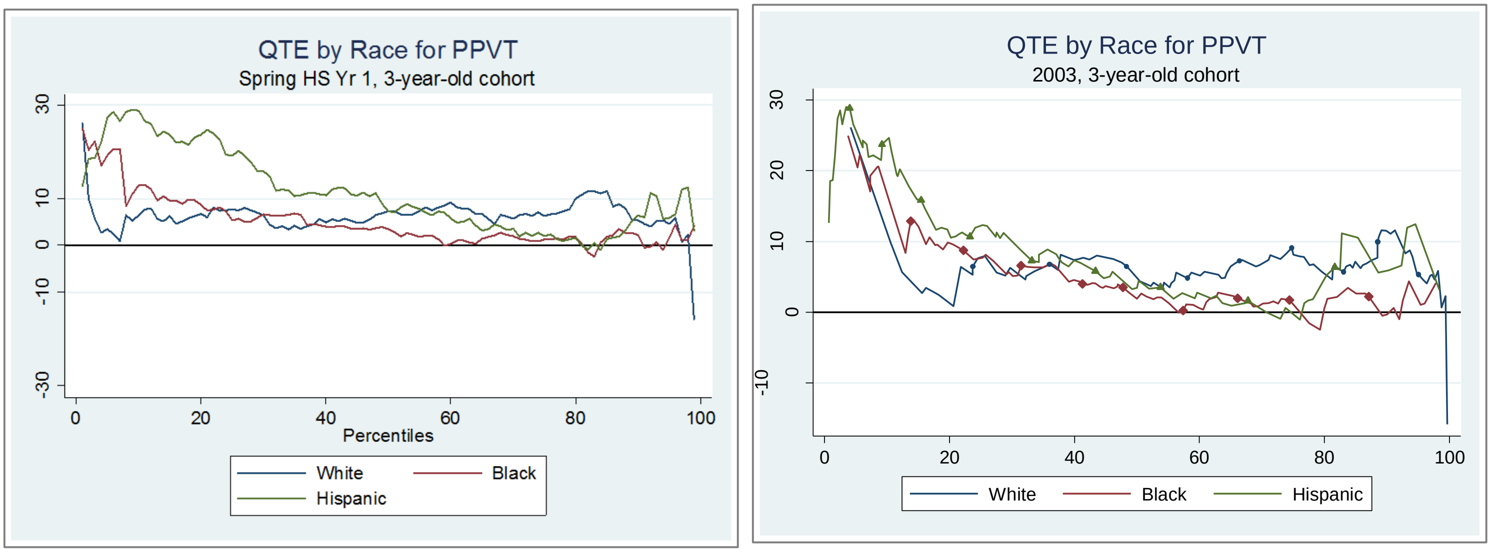
(A) Conditional QTE (B) "Translated" QTE



Notes: QTE using 3-year old cohort of HSIS and inverse propensity score weights. Figure (A) presents conditional QTE separately by language. Figure (B) translates the QTE so they are graphed on the same absolute scale, the full sample's control group percentiles. 90% confidence intervals obtained by bootstrapping Head Start center.

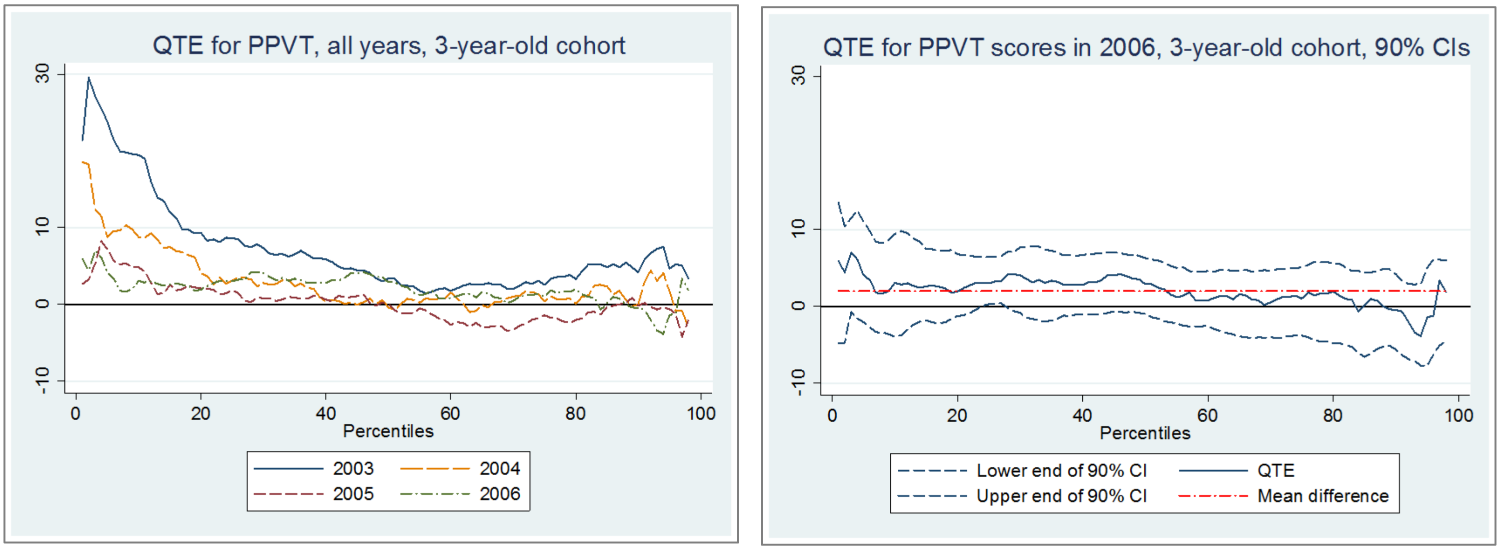
Figure 7: QTE for PPVT scores in spring 2003, by race/ethnicity (non-Hispanic black and white and Hispanic)

(A) Conditional QTE (B) Translated QTE



Notes: QTE using 3-year old cohort of HSIS and inverse propensity score weights. Figure (A) presents conditional QTE separately by race. Figure (B) translates the QTE so they are graphed on the same absolute scale, the full sample's control group percentiles. 90% confidence intervals obtained by bootstrapping Head Start center.

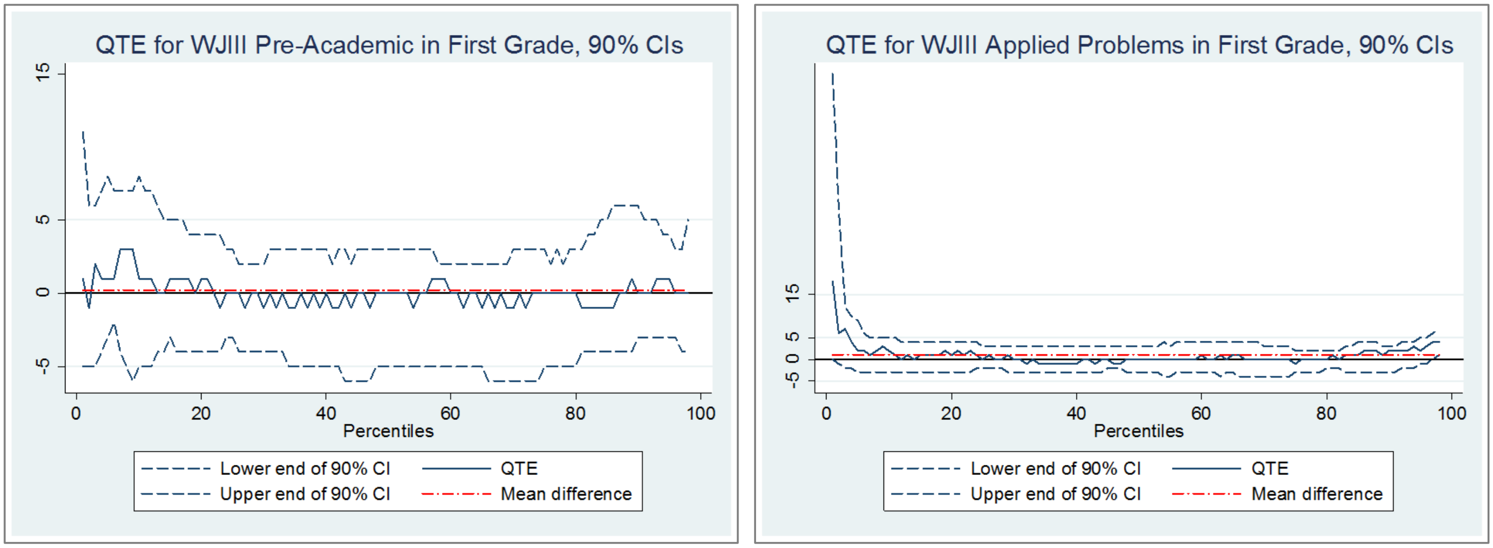
Figure 8: QTE for PPVT scores, Full sample


(A) All years

(B) Grade 1, spring 2006

Notes: QTE using 3-year old cohort of HSIS and inverse propensity score weights. Figure (A) presents QTE separately for each year. Figure (B) presented QTE for grade 1 as well as 90% confidence intervals obtained by bootstrapping Head Start center.

Figure 9: QTE for WJIII scores, Grade 1


(A) Pre-Academic

(B) Applied Problems

Notes: QTE using 3-year old cohort of HSIS and inverse propensity score weights. Figure (A) presents QTE for pre-academic skills for 2006 (grade 1) and Figure (B) presents the QTE for applied problems for 2006. Both contain 90% confidence intervals obtained by bootstrapping Head Start center.

Figure 10A: Translated QTE for spring 2006, by language

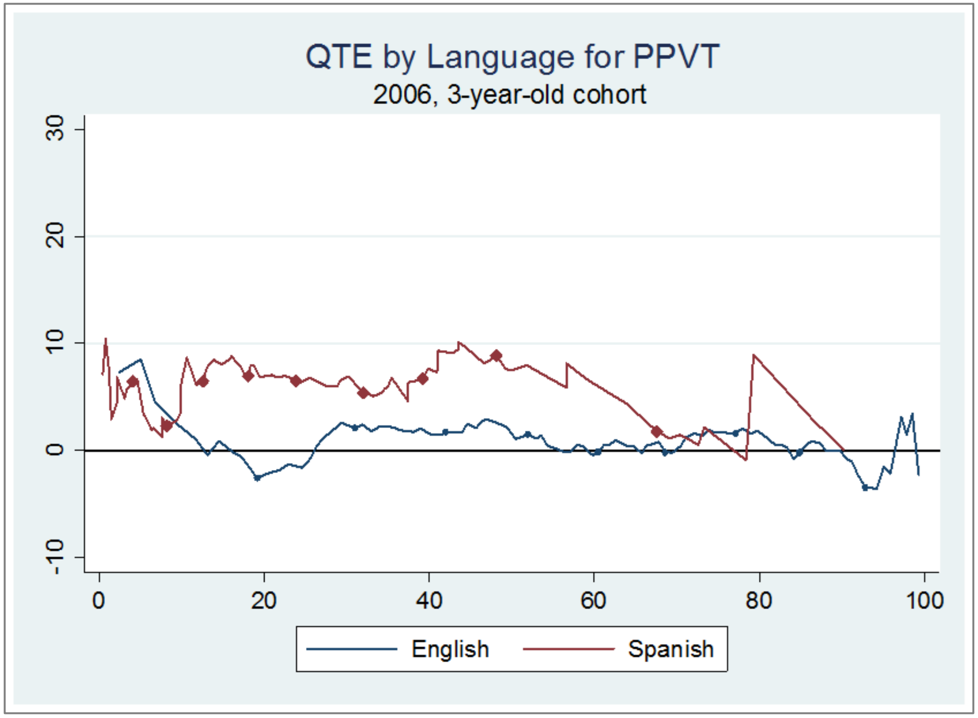
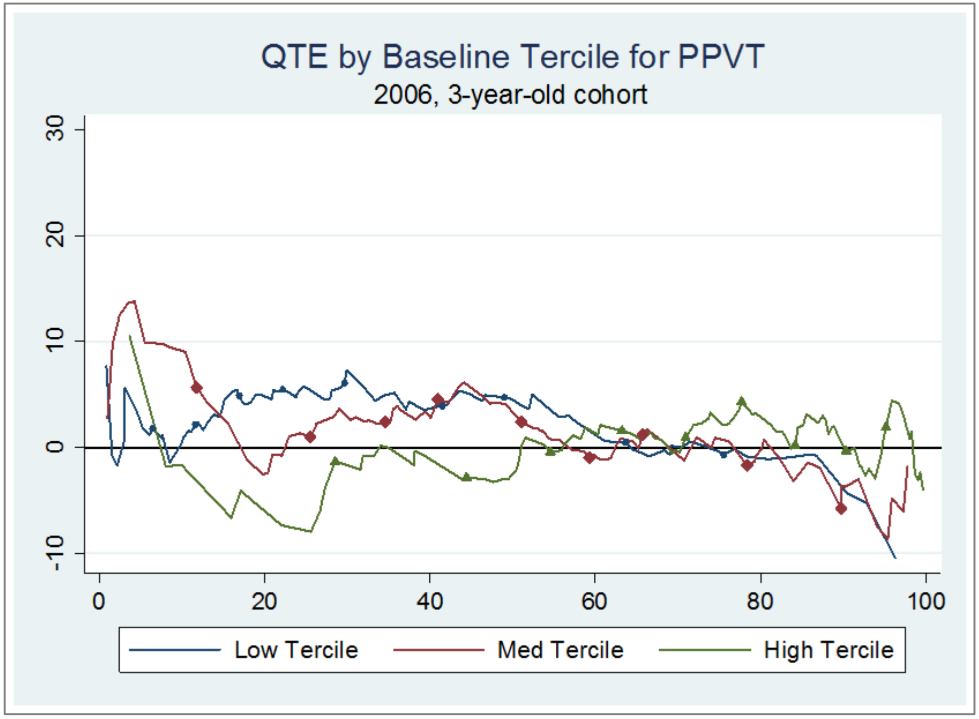


Figure 10B: Translated QTE for spring 2006, by terciles of baseline score



Notes: QTE using 3-year old cohort of HSIS and inverse propensity score weights. Both figures are "translated" QTE; each subgroup is graphed on the same absolute scale (x axis), the full sample's control group percentiles. 90% confidence intervals obtained by bootstrapping Head Start center.