Experimental evidence on distributional effects of Head Start

Marianne Bitler, UC Irvine,

Hilary Hoynes, UC Davis,

& Thurston Domina, UC Irvine.

Thanks to NICHD for financial support (1PO1HD065704-01A1).
Motivation (I)

- Wide body of evidence suggests human capital investments in early life have high returns.

- State of the Union Address: Goal of universal pre-K

- The most important public program currently is Head Start.
Motivation (II)

• Head Start’s goal is to improve outcomes for low-income pre-school children and their parents.

• Existing evidence shows some short-term cognitive impacts which fade out yet there is emerging evidence of other long-term effects.

• Several theories suggest alternative but offsetting effects on different parts of the distribution.
• QTE and subgroup specific estimates of HS impacts on cognitive and behavioral test scores.

• Means suggest little effect of a HS slot offer.

• QTE suggest positive effects at the bottom of the distribution on PPVT (early language) and the Woodcock-Johnson III pre-academic skills, fading out over time. No non-cognitive effects.

• Local linear regression and other subgroup results are consistent with QTE. Exploring effects by center.
Outline

• What is Head Start?

• Relevant literature

• Data: Head Start Impact Study

• Results: QTE, local linear regressions, subgroups, centers

• Conclusion/next steps
What is Head Start? (I)

- Head Start is a program to promote school readiness for low-income children aged 3–5, mostly for 3- and 4-year olds (Early Head Start for younger kids)

- Started as part of War on Poverty in 1965

- Mostly federally funded but locally administered (around 1400 local grantees), with some federal standards, leads to lots of variation in programs and quality.
• Recent rules require low quality centers to recompete for $.

• Served around 908,000 children in FY 2007
What is Head Start? (II)

- Eligibility: Most participants must be below poverty or on AFDC/TANF/SSI or homeless; also a requirement that 10% of slots be for the disabled

- Services: Education and cognitive development, health care, nutrition, & social services; for children and families

- Providers: Non-profits/governments/churches/school systems; 20% of funding required to come from elsewhere (e.g., First-5 San Francisco, DOE, in-kind donations); parents play a large role
<table>
<thead>
<tr>
<th>Program</th>
<th>Federal $2010, billions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head Start (no Early HS)</td>
<td>6.70</td>
</tr>
<tr>
<td>Early Head Start</td>
<td>1.00</td>
</tr>
<tr>
<td>Child care subsidies</td>
<td>5.70</td>
</tr>
<tr>
<td>Child Care Food Program</td>
<td>1.40</td>
</tr>
<tr>
<td>Child Care Tax Credits</td>
<td>2.20</td>
</tr>
<tr>
<td>DoD child care</td>
<td>0.75</td>
</tr>
<tr>
<td>Title 1 preschool</td>
<td>0.50</td>
</tr>
<tr>
<td>Preschool special education</td>
<td>0.57</td>
</tr>
<tr>
<td>Infant/toddler disability interventions</td>
<td>0.63</td>
</tr>
<tr>
<td>Home visits</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Source: Haskins and Barnett (2010).
Head Start is an important preschool program for low-income children

<table>
<thead>
<tr>
<th></th>
<th>All Kids</th>
<th>Income in Bottom 20%</th>
<th>Income in Top 20%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3-year olds</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head Start</td>
<td>8</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Special Education</td>
<td>4</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Private</td>
<td>36</td>
<td>22</td>
<td>68</td>
</tr>
<tr>
<td>Other Public</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Any Preschool</td>
<td>51</td>
<td>45</td>
<td>82</td>
</tr>
<tr>
<td><strong>4-year olds</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head Start</td>
<td>13</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Special Education</td>
<td>6</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Private</td>
<td>42</td>
<td>22</td>
<td>75</td>
</tr>
<tr>
<td>Other Public</td>
<td>13</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Any Preschool</td>
<td>74</td>
<td>64</td>
<td>90</td>
</tr>
</tbody>
</table>

Source: Haskins and Barnett (2010), from 2005 NHES.
Other pre-schools growing in importance

- HS is a large share of Federal early childhood $.

- State pre-K programs are growing in importance, especially for 4-year olds. 2011: 28% of 4-year olds in state pre-K, 11% in Head Start (NIEER).

- Relative costs: $7582/child in HS compared to $4847 for Pre-K and $12,442 for K-12.

- Concern with SES and racial gaps: Is Head Start effective?
Existing evidence on Head Start (I)

- Family fixed effects design: Positive effects of HS on test scores but some fade out (Garces, Currie, & Thomas, 2002; Currie & Thomas, 2000; Currie & Thomas, 1995)

- Family fixed effects design: Test score effects fade out yet some long-term improvements in young adult outcomes; more HS graduation, fewer learning disabilities, less likely to be in poor health (Deming, 2009)
Existing evidence on Head Start (II)

- Administrative program data: Higher cognitive scores with higher spending (Currie & Neidell, 2007)

- RD design on county: Reduction in mortality; some educational improvements (Ludwig and Miller, 2007)

- Sibling fixed effects along with HS roll out: Preschool has largest effects for those with the most early human capital, less fadeout there (Aizer and Cunha, 2012)
Existing evidence on Head Start (III)

- HSIS experiment: Improvements in parental involvement (Gelber & Isen, 2011)

- HSIS Experiment: Center effects (Bloom & Weiland, 2013)

- RD design on income at age 4: Long-term positive effects on obesity, health, depression, & incarceration (Carneiro & Ginja, 2012)

- Structural model: Head Start expansion would have positive effects (Griffen, 2011)
Other related evidence (Abecedarian, Perry Pre-School, class-size reductions)

- Other pre-school interventions: Heckman, Moon, Pinto, Saveltev, & Yavitz (2010); Anderson (2008); Schweinart, Montie, Xiang, Barnett, Belfield, & Nores (2005); Gormley, Phillips, & Gayer (2008)

- Class size reductions: Dynarski, Hyman, & Schanzenbach (2011); Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011)

- Common theme: Fade-out in test score effects yet some long-term effects on important outcomes
Our contribution

• Experimental evidence on Head Start effects for academic and socio-emotional outcomes, distributional focus with QTE

• Positive effects for PPVT and Woodcock Johnson III tests in pre-school period at the bottom of the distribution

• Augmented by local linear and other subgroup estimates within subgroups as well as center effects
Why look at distributions?

• With offsetting effects across the distribution, could miss important impacts with means

• Does the fade-out of pre-school interventions in the existing literature happen across the whole distribution?

• Given policy interest in differences in outcomes across groups, effects across the distribution are important for a full accounting (SWF).
Expected effects on the distribution: Possible compensatory role for Head Start

- Observational evidence suggests low-achieving students gain the most from early education (e.g., Magnuson, 2004).

- Head Start performance standards explicitly focus attention on basic skills.
Expected effects on the distribution: Skill begets skill predictions

- Academic skills are cumulative.

- Does this hold for relation of pre-school outcomes to elementary school scores?
Outline

• What is Head Start?

• Relevant literature

• Data: Head Start Impact Study
  
  • Results: QTE, local linear regressions, subgroups, centers

• Conclusion/next steps
Head Start Impact Study (I)

- Mandated by Congress, goal to assess effects on school readiness and parental outcomes.

- HSIS randomly assigned 3- and 4-year old children at oversubscribed programs to either an offer of placement in a HS program, or no offer.

- All students were first time applicants; treatment lasted one year.

- Samples: 2449 in 3-year old cohort, 1993 in 4-year old cohort.
Head Start Impact Study (II)


• 3-year olds free to attend HS at age 4.

• Non-compliance (19% of the treatment group did not participate in HS) and cross-over (non-trivial share—12%—of control group in other HS centers—not center of application. We address by looking at compliers.
Head Start Impact Study (III)

- We predominately look at intent to treat effects.

- The control counterfactual is whatever alternative center-based or informal care arrangements the children would have experienced otherwise.

- Importantly, the control group is not randomized to receive informal care.
Child care arrangements in Spring 2003

<table>
<thead>
<tr>
<th>Type of Care</th>
<th>Treatment Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head Start</td>
<td>0.78</td>
<td>0.14</td>
</tr>
<tr>
<td>Other center</td>
<td>0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>Non-center</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>Parent/rel.</td>
<td>0.10</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Spring 2003.
HSIS outcomes

- Cognitive: PPVT (vocabulary), Woodcock Johnson III (oral comprehension, phonetic awareness, pre-reading/reading, pre-writing, & math)

- Social/emotional: Adjustment Scales for Preschool Intervention; PIANTA closeness scales, various parent reports
HSIS cognitive outcomes

- Peabody Picture Vocabulary Test (PPVT): Measures receptive vocabulary (listening comprehension for the spoken word in standard English) and verbal ability

- Composite achievement measures from the Woodcock-Johnson III battery
Components of Woodcock Johnson III (I)

- Pre-Academic Skills: Pre-reading skills, identifying letters and words, developing math skills, and writing skills

- Basic Reading Skills: Sight vocabulary, phonics, structural analysis (K/1st grade)

- Math Reasoning: Mathematical problem solving, analysis, reasoning, and vocabulary (K/1st grade)
Components of Woodcock Johnson III (II)

- Academic Skills: Reading decoding, math calculation, and spelling of single-word responses (1st grade)

- Academic Applications: Application of academic skills to academic problems (1st grade)
HSIS social/emotional outcomes

• Adjustment Scales for Preschool Intervention (ASPI)
  – Designed to measure emotional and behavioral adjustment (teacher reports)
  – 144 items, collapsed to 5 behavioral dimensions and 3 situational dimensions
  – We combine into 3 standardized indices
  – When asked of teachers, only useful for K/first grade, younger kids don’t all have teachers
HSIS social/emotional outcomes: ASPI

- “Negative” behavior: Average of standardized scales for Aggressive, Oppositional, and Inattentive/Hyperactive

- “Shy/withdrawn” behavior: Average of standardized scales for Withdrawn-Low Energy and Socially Reticent

- “Situational” behavior: Average of standardized scales for Problems with Structured Learning, Problems with Peer Interaction, and Problems with Teacher Interaction
Other socio-emotional

- PIANTA Scale: Measures closeness, conflict, and total positive relationship; developed for child-teacher relationship, also asked of parents

- Some items like ASPI asked of parents.
HSIS: Design and weights

- Sample design was very complicated. There are baseline weights (sample design) as well as weights including non-response adjustments.

- The data include some other cognitive and behavioral/socio-emotional test scores. We have explored those which are defined for everyone and continuous.
Adjusting for observables and baseline test scores (I)

- If the baseline test assessments were conducted at the time of RA, then baseline test score differences in means (and quantiles) should be 0.

- Slight challenge, baseline tests administered in Fall of first year, at somewhat different times for T and C.

- Randomization done within centers, different shares in T/C across centers.
Adjusting for observables and baseline test scores (II)

- We adjust for differences in observables (including baseline tests) with inverse propensity score weights. Firpo (2007) shows this yields consistent estimates of the unconditional QTE.

- There is slight (but statistically insignificant) imbalance across the T and C groups in the QTE for the baseline test score (with/without imputed scores).

- This could be early effects of treatment, we also investigate this in robustness.
Adjusting for observables and baseline test scores (III)

• Estimate the probability of being in the treatment group as a function of demographics, dummies for Head Start center applied to, and deciles of baseline scores within assessment month group.

• Use a logistic regression, with baseline weights; then predict the probability of being in the T group \( \hat{p}_i \).

• Weight by the inverse p-score weight:
\[
D_i/\hat{p}_i + (1 - D_i)/(1 - \hat{p}_i).
\]
Adjusting for observables and baseline test scores (IV)

- Controls from Table 1 and center of application dummies.

- Deciles within assessment month group; imputed scores are one group.

- Accounts for differential timing of baseline test administration, with the treatment group having been tested earlier.
Adjusting for observables and baseline test scores (V)

- Few demographic observables are significant in logit as would expect given random assignment.

- Overlap of the p-scores is very complete, most far from 0/1.

- Adjustments eliminate the average T-C differences in non-response.
<table>
<thead>
<tr>
<th>Demographics</th>
<th>3-year olds</th>
<th>4-year olds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Teen mother</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>Caregiver ≤ 24</td>
<td>0.34</td>
<td>0.28</td>
</tr>
<tr>
<td>Caregiver 25–29</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>Caregiver ≥ 30</td>
<td>0.32</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Mother</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS DO</td>
<td>0.35</td>
<td>0.43</td>
</tr>
<tr>
<td>HS grad./GED</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Single</td>
<td>0.40</td>
<td>0.32</td>
</tr>
<tr>
<td>Divorced</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>Recent immig.</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td><strong>Child</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.33</td>
<td>0.18</td>
</tr>
<tr>
<td>White</td>
<td>0.34</td>
<td>0.31</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.32</td>
<td>0.51</td>
</tr>
<tr>
<td>Special needs</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>English at home</td>
<td>0.73</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>3-year olds</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>Mean C</td>
<td>Unadj. T-C</td>
</tr>
<tr>
<td>Teen mother</td>
<td>0.17</td>
<td>-0.04*</td>
</tr>
<tr>
<td>Caregiver ≤ 24</td>
<td>0.34</td>
<td>-0.05**</td>
</tr>
<tr>
<td>Caregiver 25–29</td>
<td>0.34</td>
<td>-0.02</td>
</tr>
<tr>
<td>Caregiver ≥ 30</td>
<td>0.32</td>
<td>0.07**</td>
</tr>
<tr>
<td><strong>Mother</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS DO</td>
<td>0.35</td>
<td>-0.03</td>
</tr>
<tr>
<td>HS grad./GED</td>
<td>0.33</td>
<td>0.03</td>
</tr>
<tr>
<td>Single</td>
<td>0.40</td>
<td>0.02</td>
</tr>
<tr>
<td>Divorced</td>
<td>0.15</td>
<td>-0.01</td>
</tr>
<tr>
<td>Recent immig.</td>
<td>0.15</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Child</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.33</td>
<td>0.01</td>
</tr>
<tr>
<td>White</td>
<td>0.34</td>
<td>0.02</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.32</td>
<td>0.02</td>
</tr>
<tr>
<td>Special needs</td>
<td>0.10</td>
<td>0.03**</td>
</tr>
<tr>
<td>Eng. at home</td>
<td>0.73</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Mean impacts in baseline scores and non-response

- With baseline weights (for sampling design only), there are large and significant differences across T and C in non-response, they disappear when we use the inverse p-score weights.

- Mean impacts for 2002 (baseline score) are insignificant with any of these weights.
## Differences in PPVT non-response, 3-year olds

<table>
<thead>
<tr>
<th></th>
<th>Inv. p-score wgts.</th>
<th>Control Mean [SD]</th>
<th>Mean T-C (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No baseline score/imputed</td>
<td></td>
<td>0.222 (0.422)</td>
<td>0.031 (0.022)</td>
</tr>
<tr>
<td>No preschool Y 1 score</td>
<td></td>
<td>0.190 (0.392)</td>
<td>-0.009 (0.023)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.179 (0.488)</td>
<td>0.009 (0.022)</td>
</tr>
<tr>
<td>No Kindergarten score</td>
<td></td>
<td>0.233 (0.423)</td>
<td>0.005 (0.022)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.255 (0.436)</td>
<td>0.001 (0.022)</td>
</tr>
</tbody>
</table>
### Mean impacts for PPVT for 3-year olds

<table>
<thead>
<tr>
<th></th>
<th>Control Mean [SD]</th>
<th>Mean T-C (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>231 [38]</td>
<td>-0.00 (1.84)</td>
</tr>
<tr>
<td><strong>Preschool Y 1</strong></td>
<td>251 [38]</td>
<td>7.20*** (1.63)</td>
</tr>
<tr>
<td>(Spring 2003)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Preschool Y 2</strong></td>
<td>298 [40]</td>
<td>2.89 (1.81)</td>
</tr>
<tr>
<td>(Spring 2004)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Kindergarten</strong></td>
<td>339 [29]</td>
<td>0.21 (1.29)</td>
</tr>
<tr>
<td>(Spring 2005)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1st grade</strong></td>
<td>358 [30]</td>
<td>2.00 (1.42)</td>
</tr>
<tr>
<td>(Spring 2006)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Mean impacts for PPVT for 4-year olds

<table>
<thead>
<tr>
<th></th>
<th>Control Mean</th>
<th>T-C (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>271 [38]</td>
<td>0.42 (1.80)</td>
</tr>
<tr>
<td>Preschool Y 2 (Spring 2003)</td>
<td>291 [39]</td>
<td>1.08 (1.84)</td>
</tr>
<tr>
<td>Kindergarten (Spring 2004)</td>
<td>334 [43]</td>
<td>-3.50* (2.05)</td>
</tr>
<tr>
<td>1st grade (Spring 2005)</td>
<td>360 [32]</td>
<td>-0.33 (1.71)</td>
</tr>
</tbody>
</table>
Methodology: Quantile treatment effects (I)

- Instead of estimating differences in means, we estimate differences across the distributions.

- For treatment $d$, the $q$th quantile of distribution $F_d$, $y_{qd}$, is $y_{qd} \equiv \inf_y \{ y : F_d(y_{qd}) \geq q \}$.

- Then the $qth$ quantile treatment effect (QTE), $\Delta y(q) = y_1(q) - y_0(q)$, where $y_j(q)$ is the $qth$ quantile of the $jth$ distribution.

- With random assignment, non-parametric estimator of effect of $T$ on the ex-post distribution.
• This is simply the difference between the two distributions at the $q$th quantile.
Methodology: Quantile treatment effects (II)

- This is the same as the horizontal difference between the control and treatment group CDFs, or vertical distance between inverse CDFs at the \( q \)th percentile.

- Results are summarized as graphs of the vertical differences in the inverse CDFs (horizontal differences in CDF) at each percentile index.
Inference and CIs for QTE

- For inference (using bootstrapping), we use inverse propensity score weights and incorporate the baseline weights in the p-score estimation.

- We reestimate the inverse p-score weights within the bootstrap replicates.

- CIs are 90% pointwise (may shrink if estimate jointly).
What do QTE estimate? (I)

- The QTE tell us how the distribution of the outcome changes when assigning an offer of a Head Start slot randomly.

- It does not tell us anything about the distribution of individual treatment effects without further assumptions.

- Useful if we think there are offsetting effects across the distribution.

- Ceiling effects/floor effects do not affect quantiles in the middle.
Balance, no-shows, and crossovers

- With inverse propensity score weighting, QTE on baseline scores are balanced across the groups.

- The no-shows/crossovers are evenly spread across the baseline score distribution. (First stage is difference in these.)
Local Linear Regressions: Head Start Participation

3–year–old cohort, 2003
What can we learn about the distributional effects from observables?

• In addition to the full sample QTE, we can estimate QTE for subgroups (conditional QTE).

• Subgroups include pre-RA test scores, race/ethnicity, gender, language (English/not), disadvantaged, special needs, and various characteristics of the center of application.

• Also explore whether the subgroups are systematically located at specific points in the baseline test score distribution.
• Local linear regressions of current outcomes on baseline test score (targeting). Way of looking at subgroups.
Balance in baseline distributions (I)

- We started with baseline PPVT score.

- Pretests are important in education, so this is a control in many specifications. It was imputed for nearly all non-respondents in Fall 2002.

- Means for non-response and baseline score are balanced, need to check the distributions.
Balance in baseline distributions (II)

- With the baseline weights, there was some lack of balance, insignificant at the 10% level but troubling.

- Looks much better after using inverse p-score weights.

- Next, I’ll show QTE of baseline score for 3-year olds and 4-year olds, using the inverse p-score weights. (All the other figures also use the inverse p-score weights.)

- Balance is better. Deviations from 0 are small away from the very ends of the distribution.
QTE for PPVT scores in 2002, 3–year–old cohort, 90% CIs
QTE for PPVT scores in 2002, 4−year−old cohort, 90% CIs
Outline

• What is Head Start?

• Relevant literature

• Data: Head Start Impact Study

• Results: QTE, local linear regressions, sub-groups

• Conclusion/next steps
Results: QTE for 3-year olds

- Evidence of positive effects at the bottom, significant at the 10% level, for 2003 (through most of first 40 percentiles) and 2004 (through about quantile 20).

- Negative significant effects for very few quantiles in 2005, no positive significant effects.

- Positive significant effects for 2006 for quantiles 26 and 27, no negative significant effects.
QTE for PPVT scores in 2003, 3–year–old cohort, 90% CIs
QTE for PPVT scores in 2005, 3–year–old cohort, 90% CIs
QTE for PPVT scores in 2006, 3–year–old cohort, 90% CIs
QTE for PPVT, all years, 3–year–old cohort
Results: QTE for 4-year olds

- Evidence of positive effects at the bottom, significant at the 10% level, for 2003 (through around quantile 20).

- For 2004, negative and significant effects for quantiles 40–75.

- Few negative and significant estimates for low quantiles (near 25) in 2005.
QTE for PPVT scores in 2003, 4–year–old cohort, 90% CIs
QTE for PPVT scores in 2004, 4–year–old cohort, 90% CIs

Lower end of 90% CI
Upper end of 90% CI
Mean difference

Percentiles

QTE
QTE for PPVT scores in 2005, 4–year–old cohort, 90% CIs
QTE for PPVT, all years, 4–year–old cohort

Percentiles

2003 2004
2005

QTE for PPVT, all years, 4–year–old cohort
Other outcomes: Cognitive tests

- Woodcock Johnson III tests

- There are many different components, including early language, reading, and math.

- Some positive effects at the bottom for 4-year olds for pre-academic skills, for both 3- and 4-year olds in math reasoning (in K and 1st grade); nothing for reading.
WJ III Pre-academic skills, 3-year old cohort
Mean WJ3 Pre–Academic Impacts, 3–Year–Old Cohort

Treatment – Control (Using Inverse Propensity Score Weighting)

- ▲ Significant at 95% level
- △ Not Significant
QTE for WJ3 Pre–Academic scores in 2003, 3–year–old cohort, 90% CIs

Percentiles
Lower end of 90% CI QTE
Upper end of 90% CI Mean difference

QTE for WJ3 Pre–Academic scores in 2003, 3–year–old cohort, 90% CIs

Percentiles
Lower end of 90% CI QTE
Upper end of 90% CI Mean difference
QTE for WJ3 Pre-Academic scores in 2004, 3-year-old cohort, 90% CIs
QTE for WJ3 Pre–Academic scores in 2005, 3–year–old cohort, 90% CIs
QTE for WJ3 Pre–Academic scores in 2006, 3–year–old cohort, 90% CIs
WJ III Pre-academic skills, 4-year old cohort
Mean WJ3 Pre-Academic Impacts, 4-Year-Old Cohort

Treatment – Control (Using Inverse Propensity Score Weighting)

- Significant at 95% level
- Not Significant

<table>
<thead>
<tr>
<th>Year</th>
<th>Difference in WJ3 Pre-Academic score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>+1</td>
</tr>
<tr>
<td>2004</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>+1</td>
</tr>
<tr>
<td>2006</td>
<td></td>
</tr>
</tbody>
</table>
QTE for WJ3 Pre-Academic scores in 2003, 4-year-old cohort, 90% CIs
QTE for WJ3 Pre–Academic scores in 2004, 4–year–old cohort, 90% CIs
QTE for WJ3 Pre–Academic scores in 2005, 4–year–old cohort, 90% CIs

Percentiles
Lower end of 90% CI QTE
Upper end of 90% CI Mean difference

QTE for WJ3 Pre–Academic scores in 2005, 4–year–old cohort, 90% CIs
WJ III Basic Reading (only for K/1st grade)
Other WJ III outcomes: Not much

- WJ III Math Reasoning (only for K/1st grade)
- WJ III Academic Skills (only for 1st grade)
- WJ III Academic Applications (only for 1st grade)
Other outcomes: Social/behavioral

- ASPI test composites: Negative behavior, shy/withdrawn, and situational behavior poor

- PIANTA tests: Closeness

- Not much going on with parents or teachers
ASPI negative (only for K/1st grade)
Local linear regression

- In the education literature, heavy reliance on models that control for previous test scores (means and distributional impacts).


- We do so here, using Fan local linear regression. We must condition on baseline test scores however. Bandwidth for PPVT is 11.45 and we use the Epanechnikov kernel.
Fan Local Linear Regressions Difference for PPVT
3–year–old cohort, 2003
Fan Local Linear Regressions Difference for PPVT

3–year–old cohort, 2006

0 100 200 300
Observations/regression, control
−30 −15 0 15 30
Difference
150 200 250 300
2002 Baseline PPVT Score
Difference
Observations/regression, control
3–year–old cohort, 2006
Fan Local Linear Regressions Difference for PPVT

−30
−15
0
15
30

−15
0
15
30

150 200 250 300
2002 Baseline PPVT Score

Difference

Observations/regression, control

Difference

Differences
Fan Local Linear Regressions Difference for PPVT
4–year–old cohort, 2003

2002 Baseline PPVT Score
Difference
Observations/regression, control

Observations/regression, control

Difference
Fan Local Linear Regressions Difference for PPVT

4–year–old cohort, 2005

Difference

2002 Baseline PPVT Score

Observations/regression, control
Local linear regressions

• PPVT scores: Show some positive effects for those with low baseline scores for both the 3-year old and 4-year old cohorts in 2003, nothing in first grade

• What is different about this from the QTE?

• What, if anything, can we learn from subgroup QTE and not here and vice verse?
Subgroup QTE and joint distribution of baseline and this year test scores (I)

- First, we’ll examine where the various subgroups are in the overall full sample outcome distributions. (We’ll start with subgroups based on baseline test scores, given the reliance on baseline test scores in the education literature.)

- If previous test scores and current scores were perfectly correlated, then those in the bottom third of the baseline score distribution would be the bottom third of the later distributions.
Subgroup QTE and joint distribution of baseline and this year test scores (II)

- When this and last period scores are highly correlated, then QTE and local linear regressions lead to similar implications.

- If the scores are not highly correlated, then the implications can differ.

- Local linear regressions give information useful for targeting; for example, average impacts for those at different points in the pre-RA test score distribution. They cannot by themselves tell you what
share gains from these different impacts by pre-RA score or where they end up.
Subgroup QTE and joint distribution of baseline and this year test scores (III)

- QTE on the full sample are useful for assessing the effects on the full distribution. For example, if the concern is with achievement gaps at the bottom, QTE reveal this, while it might not show up in the local linear regressions (if pre-RA score is not predictive of the later test scores).
Baseline Tercile Composition based on PPVT in 2003, 3-year-olds
Control Group

Percentile
Low Tercile
Med Tercile
High Tercile

Baseline Tercile Composition based on PPVT in 2003, 3-year-olds
Control Group
Baseline Tercile Composition based on PPVT in 2006, 3-year-olds
Control Group
Subgroups across the distribution

- In 2003, there is clearly a systematic relationship between test score at baseline and outcomes in the first year. It’s stronger for the 4-year old cohort.
Subgroup mean impacts and QTE

- Next, we'll show subgroup means and subgroup QTE.
- Big mean differences by subgroups for 3-year olds, less for 4-year olds.
- Some differences in conditional on being in subgroup QTE as well.
Mean PPVT Impacts By Tercile, 3-Year-Old Cohort

Treatment – Control (Using Inverse Propensity Score Weighting)
QTE by Baseline Tercile for PPVT
2003, 3-year-old cohort
QTE by Baseline Tercile for PPVT
2006, 3–year–old cohort
Other cuts by subgroup

• Other observables are common proxies for disadvantage, yet cannot be used flexibly as can test scores (e.g., race).

• But we can calculate subgroup means and QTE and look at the shares.
Mean impacts and QTE by race/ethnicity
Mean PPVT Impacts By Race, 3-Year-Old Cohort

Treatment – Control (Using Inverse Propensity Score Weighting)
Race Composition based on PPVT in 2003, 3-year-olds
Control Group
QTE by Race for PPVT
2003, 3–year–old cohort

Legend:
- **White**
- **Black**
- **Hispanic**
Mean PPVT Impacts By Race, 4–Year–Old Cohort
Treatment – Control (Using Inverse Propensity Score Weighting)
QTE by Race for PPVT
2003, 4–year–old cohort

Control
White Black
Hispanic
Shares and QTE by gender
Gender Composition based on PPVT in 2003, 3-year-olds
Control Group

![Graph showing gender composition based on PPVT in 2003 for 3-year-olds. The graph compares the male and female percentiles.]
QTE by Gender for PPVT

2003, 3–year–old cohort

Control
Male Female

112
Gender Composition based on PPVT in 2003, 4-year-olds

Control Group
Shares and QTE by language
Language Composition based on PPVT in 2003, 4-year-olds
Control Group
QTE by Language for PPVT
2003, 3-year-old cohort

Control

English

Spanish

116
QTE by Language for PPVT
2003, 4–year–old cohort

Control
English
Spanish
Shares and QTE by other subgroups

- We also looked at subgroups based on children’s being disadvantaged or special needs, and based on many characteristics of the center of random assignment (teacher qualifications, full day/part day, at capacity, & director’s education). These were not very informative.
Other outcomes

- School quality as an outcome; share proficient, share free and reduced price lunch, race/ethnicity

- No evidence that treatment group children are going to better schools.
Center effects

• First crack, fixed effects for treatment by center.

• We lose a bunch of centers with 1 or 2 students.

• Test of whether center by treatment FE are significant in baseline period is 1.04 (p=0.29) for pooling 3 and 4s. Same test rejects in 2004.

• If separate kids by age, reject in 2003 for both, 2004 for 3 year olds.
First pass

- Within center treatment effects, using our inverse p-score weights.

- Estimate with full set of center dummies and full set of center by treatment interactions.

- Plotted, open bars are share of cell which had mean effects significantly different from zero.

- Caveat: Many small centers (noisy), many drop out.
Difference in PPVT Score 2003 3–year–olds
Difference in PPVT Score 2003 4–year–olds

Mean Impacts by RA Center

Treatment − Control (By Center)

All p−val<.1

Difference in PPVT Score 2003 4−year−olds

Mean Impacts by RA Center

−200 −100 0 100

Treatment − Control (By Center)

All p−val<.1

Difference in PPVT Score 2003 4−year−olds

124
Do center effects persist?

- Plot 2003 T-C vs. 2002 T-C for each age group.

- Caveat: Many small centers in here.
2002, 2003 4-year-olds
Do center effects persist?

- Plot 2004 T-C vs. 2003 T-C for each age group.

- Caveat: Many small centers in here.
2003, 2004 3-year-olds
More to do with centers

- Net out some individual characteristics
- Get center by treatment effects
- Compare to histogram of QTE
- Look at effect of center characteristics (better measures than typically, can use all those who applied to the center)
• Calculate SD of treatment effects by center (fixed and random), adjust for sampling error.

• Compare to random sorting.
Interpretation

- One possibility to explain fadeout is that while PPVT at 3–4 is a good measure of something which also shows up in later life outcomes, it has independent information from early elementary outcomes.

- To explore this, we took NLSYC-79 data on children who could have attended Head Start in 86–90, and look at their outcomes at ages 20–24. We predict outcomes as a function of PPVT at 3–5 and Math PIAT scores at 9–11. If the PPVT has independent predictive power, perhaps the HSIS effects may impact these children as young adults.
NLSYC-79

- We look at two samples; all children and children predicted to be at high risk of Head Start participation.

- Participation is predicted as a function of mother’s AFQT, mother’s education, mother’s age, family income while 3–5, father’s presence, race and a dummy for being first born.

- Outcomes include whether by 18 the child has repeated a grade or been diagnosed with a disability,
whether by 17-23 the child has been out of the labor force, committed a crime, reported poor health, completed some schooling, and for girls, whether they have had a teen pregnancy.
NLSYC findings

- Math PIAT at 9-11 predicts many good outcomes.

- PPVT at 3 helps predict less grade repetition, fewer learning disabled diagnoses and a lower probability of being idle.

- PPVT effects are smaller, but effects range from -0.5 percent to 1.5 percent.

- Observational data, so speculative.
ECLS-K

- Another hypotheses is that teachers focus attention on the bottom, and “catch up” the worse off non-HS kids.

- We look at this using the ECLS-K. Not much evidence for this.
Conclusions (I)

• We have comprehensively examined the effects of an offer of Head Start on pre-school and later outcomes using distributional approaches.
  
  – Evidence of some cognitive gains concentrated at the bottom of the distribution, which fadeout at school entry.
  
  – No important impacts on socio/emotional outcomes.
  
  – Big effects for non-English speakers, Hispanics; few other differences.
- No important role for center characteristics so far.
Conclusions (II)

• Qualitative conclusions using local linear regressions on previous scores and QTE of contemporaneous scores are similar.

• First start at looking at effects by center and what might drive that.
Future work

- More about centers.

- Bounding based on assumptions about attrition

- Quantile IV for compliers (but we expect it to just scale up what we've found).