Quantifying Variation in Head Start Effects on Young Children's Cognitive and Socio-Emotional Skills Using Data from the National Head Start Impact Study

Howard S. Bloom
MDRC

Christina Weiland
University of Michigan

March 2015

Acknowledgments

This paper was funded under cooperative agreement #90YR0049/02 with the Agency for Children and Families (ACF), U.S. Department of Health and Human Services, and supported in part by a Grant from the Spencer Foundation. We thank members of the Center on Secondary Analysis of Variation in Impacts of Head Start (SAVI) — especially Pamela Morris, Hiro Yoshikawa, Lindsay Page, Avi Feller, Luke Miratrix and Larry Aber — for their helpful feedback on our work. We also thank Jennifer Brooks (formerly our ACF project officer) for her personal support of this work; Steve Raudenbush, Mike Weiss, and Kristin Porter for their collaboration in the development of the methodology that it uses; and the WT Grant Foundation for funding this methodological work. In addition, we thank Nonie Lesaux for sharing her expertise on dual language learners and the National Forum on Early Childhood Policy and Programs Meta-Analysis team for providing us with data from their meta-analysis of rigorous studies of the effects of early child care and education programs. Last, we thank Rebecca Coven for her invaluable fact-checking, copyediting, and manuscript production. However, all opinions expressed in the present paper and any errors that it might contain are solely the responsibility of the authors.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, The Harry and Jeanette Weinberg Foundation, Inc., The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

Abstract

This paper uses data from the Head Start Impact Study (HSIS), a nationally representative multi-site randomized trial, to quantify variation in effects of Head Start during 2002-2003 on children's cognitive and socio-emotional outcomes relative to the effects of other local alternatives, including parent care. We find that (1) treatment and control group differences in child care and educational settings varied substantially across Head Start centers (program sites); (2) Head Start exhibited a compensatory pattern of program effects that reduced disparities in cognitive outcomes among program-eligible children; (3) Head Start produced a striking pattern of subgroup effects that indicates it substantially compensated dual language learners and Spanish-speaking children with low pretest scores (two highly overlapping groups) for their limited prior exposure to English; and (4) Head Start centers ranged from much more effective to much less effective than their local alternatives, including parent care.

Contents

List of Exhibits

Table

Figure

Introduction

Created in 1965 as part of the U.S. War on Poverty, Head Start is now the largest federal program serving the developmental needs of young children from low-income families. The program began by providing eight-week summer sessions to 4-year-olds and expanded to nine-month or year-round half-day or full-day programming for over 900,000 predominantly 3- and 4-year-olds from all 50 U.S. states, Puerto Rico and the U.S. territories, and the District of Columbia (Office of Head Start, 2014a; 2014b). At a current cost of roughly $7 billion per year, Head Start is designed to serve the "whole child" through educational, health, nutritional, and social services.

The program is funded by a federal appropriation and administered by the Office of Head Start in the Administration for Children and Families of the U.S. Department of Health and Human Services. The Office of Head Start awards grants to public and private nonprofit organizations that operate local Head Start programs. The programs operated by these grantees or their delegate agencies must meet federal performance standards. There are currently about 1,800 Head Start grantees that provide program services through roughly 16,000 Head Start centers (Administration for Children and Families, 2014).[1] Program participants are mainly 3- and 4-year-olds, most of whom are served in classroom settings. About 2 percent of participants are served in their homes or through family-based child care (Office of Head Start, 2014a).

Head Start has been a perennial source of controversy and continues to be hotly debated, with periodic calls for its expansion or dissolution. For example, there currently are early education and child care proposals that could increase investment in the program (e.g., Strong Start for America's Children Act, 2013; The White House, 2013) and other proposals that could substantially alter its funding and oversight (e.g., Head Start Improvement Act, 2014).

Many studies have examined the average short-term, medium-term, and long-term effects of Head Start on child, youth, and young adult development (e.g., Abbott-Shim, Lambert, and McCarty, 2003; Currie and Thomas, 1995; Deming, 2009; Garces, Thomas, and Currie, 2002; Ludwig and Miller, 2007; Shager et al., 2013). To provide new insights about the program's effectiveness, the present study addresses a different question — namely, by how much do Head Start short-term effects on children's cognitive and socio-emotional skills vary across individuals, subgroups, and sites?

Prior researchers have suggested that Head Start effects vary substantially, especially across sites (e.g., Barnett, 2007; Currie, 2007; Ludwig and Phillips, 2007; Zaslow, 2008), but

[1]We determined the number of Head Start grantees and centers from the ACF (2014) Head Start location data set. All Head Start grantees and centers in the data set were included except for Early Head Start programs.

there is almost no empirical evidence that confirms or refutes this suggestion. Because understanding variation in effects is essential for managing a highly decentralized program like Head Start, the present paper attempts to help fill this knowledge gap.

Our analysis is based on data from the National Head Start Impact Study (HSIS),[2] a multisite randomized trial conducted in a nationally representative sample of Head Start centers that were "oversubscribed" (had more applicants than program slots) during program year 2002-2003 (Puma et al., 2010a).[3] This analysis produced the following main findings.

- *Cross-site variation in the HSIS treatment contrast:* Treatment and control group *differences* in child care and educational settings varied substantially across Head Start centers (program sites).
- *Individual variation in program effects:* Head Start exhibited a compensatory pattern of program effects that reduced disparities in key cognitive outcomes among program-eligible children.
- *Subgroup variation in program effects:* Head Start produced a striking pattern of subgroup effects, which indicates that it substantially compensated dual language learners and Spanish-speaking children with low pretest scores (two highly overlapping groups) for their limited prior exposure to English.
- *Cross-site variation in program effects:* Head Start centers ranged from much more effective to much less effective than their local alternatives, including parent care.

The sections that follow review prior knowledge about the effects of Head Start, describe the present research design and analysis, report the present findings, and conclude with a brief discussion.

[2]The central goal of the congressional mandate for the Head Start Impact Study (HSIS) was to determine "*... if overall, the Head Start programs have impacts consistent with their primary goal of ... increasing ... school readiness*" (Puma et al., 2010a, pp. 1-11). This mandate also called for the Head Start Impact Study to consider "*... possible sources of variation in impacts of Head Start programs*" (Puma et al., 2010a, pp. 1-11). The present paper addresses a question that is a precursor to the second component of the HSIS mandate, which is to *quantify* the amount of variation in Head Start effects that exists.

[3]To be included in the HSIS, grantees and centers had to have enough applicants to permit "creation of a control group without requiring Head Start slots to go unfilled" (Puma et al., 2010a, p. xviii). The study population is therefore the national population of oversubscribed Head Start centers in 2002-2003.

Prior Knowledge About Head Start Effects

There is a substantial body of empirical evidence about the average effects of Head Start, but much less evidence about variation in these effects.

Average Effects

Over Head Start's nearly 50-year history, three basic metrics have been used to assess the program's effects: (1) its average effects on the development of participating children relative to those of local alternative options, including parent care; (2) its average effects on child development relative to those of other specific preschool programs; and (3) its average effects on child development relative to its average costs.

In terms of the first metric, there is considerable evidence that Head Start increases school readiness beyond what it would be with only existing alternatives for its eligible population (e.g., Barnett, 1995; Currie, 2001; Deming, 2009; Ludwig and Phillips, 2008; Puma et al., 2010a; Yoshikawa et al., 2013), although the quality of this evidence varies. The evidence indicates that, on average, Head Start improved immediate post-program school readiness skills for children who participated before 1980 (Currie and Thomas, 1995), during the 1980s and 1990s (Abbott-Shim, Lambert, and McCarty, 2003; Deming, 2009), and during the 2000s (Puma et al., 2010a). For example, across the 27 most rigorous studies of Head Start effects between 1965 and 2007, a recent meta-analysis reports an average effect size of 0.27 standard deviation on immediate post-program cognitive outcomes (Shager et al., 2013). These findings are about the same for sample members who were in Head Start before 1974 (the year in which the first national program quality guidelines were implemented) and thereafter.

Studies of Head Start's medium-term effects (when children are between 5 and 18 years old) universally find that cognitive test scores of participants and their control group or comparison group counterparts converge over time. This pattern seems to hold regardless of what decade participants were in Head Start (Barnett, 1995; Cicirelli, 1969; Deming, 2009; Puma et al., 2012). However, there is some evidence of sustained medium-term effects on special education placements and grade retention (Barnett, 1995; Currie and Thomas, 1995).

The small number of studies that have examined longer-term Head Start effects by following participants and nonparticipants into adulthood provide consistent evidence of beneficial effects on outcomes such as high school completion, college enrollment, physical health, mortality, criminal behavior, and "idleness" (Deming, 2009; Garces, Thomas, and Currie, 2002; Johnson, 2013; Ludwig and Miller, 2007).[4] However, given the time that must elapse before

[4]Deming (2009) defines "idleness" as an individual not being in school and reporting zero wages in 2004, when his sample members were at least 19 years old.

these longer-term outcomes can be measured, existing studies of them represent only persons who were in Head Start during the 1980s or earlier.

With respect to the second metric — Head Start's effectiveness relative to that of other specific preschool options — existing evidence is mixed and incomplete. On the one hand, Head Start appears to be less beneficial (in the short and the long run) than small intensive model programs, such as the early Perry and Carolina Abecedarian preschools (Barnett, 1995). Head Start's short-term cognitive and socio-emotional effects are also smaller than those of some well-known current, large-scale, publicly funded prekindergarten programs, such as those in Boston (Weiland and Yoshikawa, 2013), Tulsa (Gormley et al., 2005; Gormley, Phillips, and Gayer, 2008; Gormley et al., 2011), and New Jersey (Wong et al., 2008). On the other hand, Head Start's estimated effects on children's language and mathematics skills are more favorable than those of two states in the recent evaluation of five state prekindergarten programs conducted by Wong and colleagues (2008). Head Start's short-term effects on vocabulary and mathematics are similar to those of the Tennessee prekindergarten program (Lipsey et al., 2013), and a recent study in Tulsa found no difference for two out of three cognitive outcomes between children who attended two years of Head Start and those who attended one year of Head Start plus a year of the city's publicly funded prekindergarten program (Jenkins et al., 2014). Another study in Tulsa found that while attending either Head Start or public prekindergarten for one year had positive effects on two measures of children's literacy skills, effects were much more pronounced for public prekindergarten participants (Gormley et al., 2010). Effects were positive and similar between programs for children's early mathematics skills. Moreover, descriptive studies find similar levels of emotional support and instructional quality in Head Start and state prekindergarten programs (Mashburn et al., 2008; Office of Head Start, 2013).

However, comparisons between Head Start and other types of early child education programs are difficult to interpret. One reason for this is that no study has randomly assigned children to Head Start versus other specific preschool programs, which is the only way to fully account for potential differences between the types of families who normally would choose these options. Thus observed differences between the effects of Head Start and those of other programs might reflect preexisting differences between their participants. For example, several studies find that children who enroll in Head Start are more disadvantaged than those who do not enroll (e.g., Currie and Thomas, 1995; Lee et al., 1990; Feller et al., 2014). In addition, much of the evidence on the effectiveness of public prekindergarten programs comes from studies that used a regression discontinuity design. These studies estimated a localized treatment effect — effects for those who just made a birthday cutoff versus who just missed it for program admission in a given year — while Head Start studies have estimated average effects across the full sample of participants. Further, regression discontinuity studies of prekindergarten have differed from standard regression discontinuity studies in ways that make their treatment effect estimates not directly comparable with those from experimental studies (Lipsey et al., 2014).

In terms of Head Start's third assessment metric — its benefit-cost ratio — there has been no comprehensive accounting of the program's economic benefits and costs. However, there have been at least three "back of the envelope" approximations (Currie, 2001; Deming, 2009; Ludwig and Phillips, 2008), all of which suggest that Head Start would pass a benefit-cost test. For example, because Head Start is much less expensive than intensive model programs like Perry and Carolina Abecedarian, the smaller effects observed for Head Start might be as cost effective as the larger effects observed for these other programs. Along these lines, Deming (2009) concludes that in the 1980s, Head Start generated roughly 80 percent of the long-term benefits of Perry and Abecedarian for 60 percent of their costs. At the current time, however, the jury is out on this issue.

Variation in Effects

While knowledge about average program effects is essential for guiding program policy, knowledge about variation in program effects is equally valuable for guiding program management. For example, evidence about how program effects vary across participant subgroups can help to target a program, and knowledge about how program effects vary across sites can be used to learn from sites that are especially effective how to improve other sites.

There are several reasons to expect the effects of Head Start to vary. First, children with different characteristics might respond to the program differently. For example, Kelchen et al. (2012) hypothesize that early education might affect girls and boys differently because of their differences in specific skills or differences in their susceptibility to environmental influences — although the authors found limited empirical evidence of such differentiation. In addition, developmental trajectories for girls with respect to socio-emotional skills appear to be steeper than those for boys (e.g., Card et al., 2008; Matthews, Ponitz, and Morrison, 2009), which might cause Head Start to produce different results for girls and boys. In addition, there is some evidence that boys are more sensitive than girls to environmental stressors (e.g., Kraemer, 2000).[5] If this is true, and if Head Start helps young children cope with these stressors, boys and girls might respond differently to the program. Similarly, existing theory and empirical research suggest that differential program effects might exist for subgroups defined by race/ethnicity (Magnuson and Waldfogel, 2005), home language (Barnett et al., 2007; Magnuson, Lahaie, and Waldfogel, 2006), preexisting skills (Bitler, Hoynes, and Domina, 2014), and special needs (Phillips and Meloy, 2012). Thus some subgroups might benefit more than others from Head Start and sites with higher concentrations of the former might benefit more than sites with higher concentrations of the latter.

[5]These findings are ambiguous, however, because through gender socialization, girls more than boys might be encouraged to inhibit displays of anger and aggression in response to environmental stressors (Zaslow and Hayes, 1986).

The effectiveness of Head Start might also vary because of differences in program dosage and quality. For example, some Head Start programs are full-day and others are half-day; and within these parameters, children's attendance can vary. There is some evidence that children with higher-than-average program dosage and sites with higher concentrations of such children experience larger-than-average program effects. During early childhood, there is consistent evidence of positive associations between higher dosage of center-based care and children's cognitive skills, and there is inconsistent evidence of adverse associations between higher dosage of center-based care and children's socio-emotional skills (Loeb et al., 2007; Magnuson, Ruhm, and Waldfogel, 2007; NICHD Early Child Care Research Network, 2002; Votruba-Drzal, Coley, and Chase-Lansdale, 2004). Furthermore, two preschool studies found effects that were larger for full-day programs than for part-day programs (Robin, Frede, and Barnett, 2006; Walters, 2014). Across the many existing studies of kindergarten, full-day programs have been found to be as effective as or more effective than half-day programs and never less effective (see Lee et al., 2006, for a review; Gibbs, 2014).

It is far more difficult to assess the extent to which variation in Head Start effects is driven by differences in individual child attendance. This is because attendance is likely to be endogenous to other longer-term outcomes and poor attendance is linked to multiple risk factors (Epstein and Sheldon, 2002). The one study that carefully attempted to account for endogenous child attendance patterns (Arbour et al., 2014) found that positive effects of the preschool quality improvement program being examined on children's literacy skills were only experienced by students with high attendance rates.

Existing evidence about the likely magnitude of variation in Head Start effectiveness due to variation in program quality is mixed. Some evidence suggests that high-quality early care and education programs produce effects on children's cognitive and socio-emotional skills that are demonstrably larger than those produced by lower-quality programs (Barnett, 1995; Yoshikawa et al., 2013). This would imply that if there is considerable variation in the quality of local Head Start programs, there should be considerable variation in their effectiveness. However, other studies find that the cross-site standard deviation of measured Head Start quality is relatively small (e.g., Moiduddin et al., 2012; Puma et al., 2010a),[6] which suggests little corresponding variation in program effects. Further, there are relatively weak relationships between observational measures of quality in preschool programs and gains in enrolled children's cognitive and socio-emotional skills (Burchinal, Kainz, and Kai, 2011; Mashburn et al., 2008;

[6]Resnick and Zill (1999) used data from the 1997 national Head Start observational study (the Family and Children Experiences Survey, FACES) to decompose variation in a measure of Head Start quality based on the Early Childhood Environmental Rating Scale (ECERS) into variation across Head Start classrooms, centers, and program/grantees. They found that roughly one-third of the total variation exists at each of these three levels.

Weiland et al., 2013). Nonetheless, experimental studies of interventions that improved measures of classroom emotional support and/or instructional quality by 0.40 standard deviation or more report positive effects — some of which are quite large — on children's language, literacy, executive function, and other socio-emotional skills (Bierman et al., 2008; Raver et al., 2008; Raver et al., 2009; Raver et al., 2011). Thus seemingly modest cross-site variation (modest cross-site standard deviations) in observed Head Start quality might reflect meaningful variation in program effectiveness.

There also could be variation in estimates of Head Start effectiveness due to differences in local alternative child care options, which represent the "counterfactual setting" against which Head Start is compared. For example, there are currently far more alternative preschool options for low-income families than there were when Head Start began (Shager et al., 2013), and many of these options have been found to increase children's cognitive skills (Leak et al., 2012). Thus the bar that Head Start must exceed in order for it to be judged effective has been rising over time and might well vary substantially across localities in which Head Start operates.

Along these lines, a recent study found that Head Start's effects on receptive vocabulary for children who would have stayed at home had they not attended the program are larger than for children who would have enrolled in alternative preschool programs (Feller et al., 2014). Two other such studies found that Head Start attendance was associated with improved cognitive and socio-emotional skills when compared to parent care (Zhai, Brooks-Gunn, and Waldfogel, 2011; 2014). When compared with other center-based programs, one of the two studies found positive short-term effects of Head Start attendance on socio-emotional outcomes (Zhai, Brooks-Gunn, and Waldfogel, 2011), while the other found no such benefits (Zhai, Brooks-Gunn, and Waldfogel, 2014). Likewise, based on their meta-analysis of 27 rigorous Head Start studies, Shager and colleagues note that: "We found that evaluation studies in which the control group actively sought alternative ECE services produced a smaller average effect size (0.08) than studies with passive control groups (0.31)" (Shager et al., 2013, p. 88).

Existing *direct* evidence about variation in Head Start effects focuses mainly on differences in effects across child subgroups defined by such factors as race/ethnicity, gender, special needs, and home language (e.g., Currie and Thomas, 1995; Deming, 2009; Puma et al., 2010a; Yoshikawa et al., 2013). No consistent subgroup pattern emerges from these findings, however. Interestingly, a recent paper by Bitler, Hoynes, and Domina (2014) that applies quantile regression analysis to data for the 3-year-old HSIS cohort finds that Head Start has a compensatory pattern of effects on some cognitive outcomes (it raises the bottom of the test score distribution by more than it raises the rest of the distribution). In addition, a recent paper by Walters (2014) uses HSIS data to demonstrate that Head Start effects on a composite measure of numerous cognitive outcomes and on a composite measure of numerous socio-emotional outcomes vary substantially across centers.

Even in the face of this limited information, many Head Start observers expect there to be substantial variation in the program's effects. These expectations are often motivated by perceptions that the quality of local Head Start programs varies substantially and that this variation in quality *must* produce variation in program effects. For example: Ludwig and Phillips (2007, p. 11) stated that "variation in quality and context matter for the delivery and impacts of early childhood programs." Zaslow (2008, p. 6) noted that the Head Start Impact Study "is a study of the impacts of a program as it was broadly implemented in a wide range of circumstances. This is not an evaluation of a small, tightly controlled demonstration program with uniform high quality." Barnett (2007, p. 675) noted that "the average estimated effects in these two Head Start studies conceal a great deal of heterogeneity in effects," and Currie (2007, p. 682) stated that "Head Start is run at a local level, so there is variation in quality."

To help fill this knowledge gap, the present paper provides direct empirical evidence about the magnitude of variation in Head Start effects across individuals, subgroups of individuals, and program sites (Head Start centers) on four key measures of young children's cognitive development (their receptive vocabulary, letter-word recognition, oral comprehension, and early numeracy skills) and on two key measures of young children's socio-emotional development (externalizing and self-regulation).

Methods and Measures

This section describes our samples, analysis period, outcome measures, estimation methods, and baseline sample balance. Appendix A describes how our analysis differs from that for the HSIS Final Report (Puma et al., 2010a).

Samples

The HSIS randomized 4,667 eligible 3- and 4-year-old first-time Head Start applicants from a national sample of 378 oversubscribed Head Start centers (Puma et al., 2010a). These children were randomly assigned to Head Start (the study's treatment group) or to a control group whose members could not enroll in the Head Start center for which they were randomized. The HSIS restricted-use file, which is the basis for the present analysis, omits sample members from Puerto Rico, resulting in a sample of 4,440 children from 351 Head Start centers.

For each of the six HSIS outcome measures that we examine, sample members with missing data for that outcome were omitted from its analysis. Also for each outcome, a few typically very small Head Start centers were dropped, because after omitting sample members with missing outcome data, these centers either had no treatment group members or no control group members. Hence they could not provide experimental estimates of Head Start effects. Last, a few typically very small Head Start centers were dropped because they had zero compliance

with random assignment (the proportion of their control group members who enrolled in Head Start equaled the proportion of their treatment group members who did so). These centers provide no information about Head Start effects.

This process produced six analysis samples (one for each outcome) that pool the HSIS 3-year-old and 4-year-old cohorts. These samples contain between 3,465 and 3,529 children from between 295 and 297 Head Start centers.[7] Each center represents a randomized trial for between 2 and 75 children. The arithmetic mean center sample size is 11.9 and its standard deviation is 8.9; the harmonic mean sample size is 7.3.[8]

Analysis Period

HSIS baseline data were collected in the fall of 2002, which was the beginning of sample members' "Head Start Year."[9] The study's first wave of follow-up data was collected in the spring of 2003, which was near the end of sample members' Head Start year. Subsequent waves of data were collected in the springs of sample members' kindergarten year, first-grade year, and third-grade year (Puma et al., 2012) and, for 3-year-olds only, the spring of their second preschool year.

The present analysis focuses on sample members' Head Start year (2002-2003) because thereafter the "treatment contrast" for 3-year-old cohort members differs substantially from that for 4-year-old cohort members. This difference exists because control group members in the 3-year-old cohort were made eligible for Head Start after 2002-2003 and many of them entered the program at this time. In contrast, most 4-year-old cohort members became eligible at this time for kindergarten and thus had no reason to enroll in Head Start. Thus during the second HSIS follow-up year, 65 percent of treatment group members and 49 percent of control group members in the 3-year-old cohort enrolled in Head Start, whereas only 6 percent and 7 percent of treatment and control group members in the 4-year-old cohort enrolled in Head Start, prekindergarten, or any other type of center-based care.[10]

---

[7]The number of children and Head Start centers in the analysis sample for each outcome is listed at the bottom of Table 9.

[8]These sample sizes are for the receptive vocabulary outcome. Sample sizes for other outcomes are very similar.

[9]HSIS baseline data were collected *after* sample members were randomized and in their preschool settings. Thus pretest scores for Head Start participants could have been influenced by program participation. Therefore controlling for pretest scores to improve the precision of program effect estimates could cause these estimates to understate true program effects. Our sensitivity analyses (discussed later) indicate that this potential bias is negligible.

[10]In the study's second year, parents of 4-year-old cohort members were asked whether their child was in "Head Start, pre-kindergarten, or any other type of center-based child care" (U.S. Department of Health and Human Services, 2004, p. 5). They were not asked specifically about enrolling in Head Start.

Because the timing of baseline and follow-up data collection varied across HSIS sites and data collection methods, we estimate that the "Head Start year" for sample members — which is the focus of the present analysis — ranges from three to seven months, with an average of five months.[11] Our estimates of Head Start effects reflect this range of exposure to Head Start and its local alternatives.

Outcome Measures

The present analysis examines variation in Head Start effects on the following four cognitive outcome measures and two socio-emotional outcome measures, the details of which are described in Appendix B.

Cognitive measures

- *Receptive vocabulary* measured by the Peabody Picture Vocabulary Test-III (PPVT; Dunn and Dunn, 1997)
- *Early reading* measured by the Woodcock-Johnson Letter-Word Identification subscale (WJ-LW; Woodcock, McGrew, and Mather, 2001)
- *Oral comprehension* measured by the Woodcock-Johnson Oral Comprehension subscale (WJ-OC; Woodcock, McGrew, and Mather, 2001)
- *Early numeracy* measured by the Woodcock-Johnson Applied Problems subscale (WJ-AP;Woodcock, McGrew, and Mather, 2001)

Socio-emotional measures

- *Externalizing behavior problems* measured by the Child Behavior Checklist (Achenbach, Edelbrock, and Howell, 1987)
- *Self-regulation skills* measured by the Leiter-R Assessor Report (Roid and Miller, 1997)

[11]Child assessments indicate the month in which each pretest was administered and the day and month in which each post-test was administered. To approximate the time interval between these two assessments, we assume that pretests were administered on the fifteenth day of each month. The resulting mean interval is 156 days (five months) and its standard deviation is 31 days (one month). Assuming approximate normality, this implies an interval that ranges from roughly three to seven months for 90 percent of sample members. Parent reports indicate the month in which pretests and post-tests were administered. Assuming that both were administered on the fifteenth day of each month implies a mean interval of 162 days and a standard deviation of 35 days. This also suggests a three- to seven-month interval for 90 percent of sample members.

Estimation

The present analysis examines variation in (1) the effects of random assignment to Head Start, which is an average effect of "intent to treat," or ITT; and (2) the effects of participation in Head Start, which is a local average treatment effect, or LATE (Angrist, Imbens, and Rubin, 1996).[12] Estimation for the first analysis is described in the present section; estimation for the second analysis, which is more complex, is described in Appendix C.[13]

The following two-level random-coefficients model was used to estimate cross-site variation in effects of Head Start assignment.

Level One: Individual Children

$$Y_{ij} = \alpha_j + \beta_j \cdot T_{ij} + \sum_{k=1}^{K} \pi_k \cdot X_{kij} + e_{ij} \quad (1)$$

Level Two: Head Start Centers

$$\alpha_j = \alpha_j \quad (2)$$

$$\beta_j = \beta_0 + r_j \quad (3)$$

where:

Y_{ij} = the value of the outcome measure for child *i* from Head Start center *j*,

T_{ij} = one if child *i* from Head Start center *j* was randomly assigned to the program and zero otherwise,

X_{kij} = baseline characteristic *k* for child *i* from Head Start center *j*,

α_j = the mean control group outcome for Head Start center *j* (which is fixed for each center),

β_j = the mean effect of random assignment to Head Start for Head Start center *j* (which varies randomly across centers),

β_0 = the cross-site grand mean effect of random assignment to Head Start (the mean of the site mean effects),

[12]Some authors refer to effects of program participation as effects of "treatment on the treated" (TOT). However given the instrumental variables method that is typically used to estimate these effects for randomized trials (and is used for the present analysis), these findings do not represent average effects of treatment on the treated when there are both "no-shows" in the treatment group (treatment group members who do not receive the treatment) *and* "cross-overs" in the control group (control group members who do receive the treatment). Instead they represent local average treatment effects — i.e., average program effects on compliers.

[13]Estimation of cross-site variation in the effects of Head Start participation (LATE) is based on an extension of the approach presented by Raudenbush, Reardon, and Nomi (2012).

e_{ij} = a random error that varies independently across individuals with a mean of zero and variances σ_T^2 and σ_C^2 for treatment group members and control group members, respectively, and

r_j = a random error that varies independently across sites with a mean of zero and a variance of τ_{ITT}^2.

The preceding model specifies a separate *fixed* intercept (α_j) for each Head Start center to represent its mean counterfactual untreated outcome. This eliminates a bias that could occur due to variation across Head Start centers in the proportion of sample members randomized to the program or treatment ($\bar{T}$).[14] The model also specifies *random variation* across Head Start centers in the mean effect of assignment to the program (β_j). The cross-site grand mean of this parameter (β_0) and its cross-site standard deviation (τ_{ITT}) are a central focus of the present analysis.[15] In addition, the model specifies *separate* individual-level residual outcome variances (σ_T^2 and σ_C^2) for the treatment group and control group, to account for the possibility that individual variation in Head Start effects changes the individual outcome variance. These individual-level outcome variances are policy-relevant parameters and play an important role in the present analysis.

Using a random-coefficients model like Equations 1-3 to study cross-site variation in program effects makes it possible to estimate variation in *true* program effects instead of merely reporting variation in program-effect *estimates*. This reflects the fact that a random-coefficients model can account for cross-site variation in program effect estimates that is due to variation in random estimation error. If instead one summarized the distribution of site-specific program effect estimates from a conventional model of site-specific fixed coefficients, the variance of these estimates (which reflects both cross-site variation in true program effects and cross-site variation in random estimation error) could be many times larger than the variance of true program effects.

In all analyses, our chosen minimum threshold for statistical significance was a p-value of 0.10. We chose this level because it matches the threshold used in the original HSIS (Puma et al., 2010a; 2010b).

[14]This bias can arise if the proportion of sample members randomized to treatment ($\bar{T}$) is correlated across sites with mean untreated counterfactual outcomes (α_j).

[15]We test the statistical significance of estimates of β_0 using the corresponding z statistic reported by conventional software for the multilevel model represented by Equations 1-3; we test the statistical significance of estimates of τ_{ITT}^2 (and thus τ_{ITT}) using the conventional Q statistic for random-effects meta-analyses (Hedges and Olkin, 1985).

Sample Baseline Balance

To assess the effect of missing data on the baseline balance of our analysis samples (and thus the internal validity of our estimates of Head Start effects), Table 1 compares treatment and control group means for all baseline covariates.[16] These findings, which are for our largest analysis sample (that for early reading) and were replicated for our other five analysis samples (but not reported here), indicate that observed treatment and control group baseline differences range from small to negligible. Thus sample attrition does not appear to threaten the internal validity of our findings. However, one can never be certain that unobserved treatment and control group differences do not exist.

Findings

This section presents our findings with respect to (1) cross-site variation in the HSIS treatment contrast, (2) individual variation in the effects of Head Start assignment, (3) subgroup variation in the effects of Head Start assignment, and (4) cross-site variation in the effects of Head Start assignment and participation.

Cross-Site Variation in the HSIS Treatment Contrast

Because estimates of Head Start effects are, by definition, *relative* to the effects of alternative child care and education settings experienced by sample members who were not in the program in 2002-2003, it is first useful to characterize the difference between these two counterfactual settings, which we refer to as the HSIS "treatment contrast." To do so, Table 2 presents estimates of the cross-site grand mean and cross-site standard deviation (with Head Start centers as sites) for five features of the HSIS treatment contrast. These features represent the treatment and control group *difference* in the percentage of sample members who (1) were enrolled in Head Start, (2) were enrolled in any center care (including Head Start), (3) had a teacher with a bachelor's degree (with a code of "no" for no teacher *or* for teacher with no BA), and (4) were in high-quality non-relative-based care (were in non-relative-based care *and* this care had a score of 5 or greater on the Early Childhood Environmental Rating [revised] or ECERS-R).[17] The table also presents estimates of the treatment-control difference in mean weekly hours of center care (including zeros for no center care). These results were obtained by estimating Equa-

[16]Findings in Table 1 were obtained by estimating Equations 1-3 and specifying each baseline characteristic as the dependent variable of Equation 1. We used the same baseline characteristics as the HSIS.

[17]As in the original HSIS, the present analysis conflates type of care with teachers' education and with quality of care. This is because children in parent care did not have teachers and because quality of care information is not available for children who were in parent care. Friedman-Krauss, Connors, and Morris (2014) use multinomial logistic regression to attempt to disentangle type of care from quality of care, but this approach is not compatible with the present analysis of cross-site impact variation.

tions 1-3 with each treatment-contrast feature as the dependent variable for Equation 1. Thus our HSIS treatment-contrast analysis parallels our analysis of Head Start assignment effects on child outcomes.

The cross-site grand mean of each feature of the treatment contrast represents its overall average magnitude. The larger this average is for a given feature, the larger the average effect of assignment to Head Start in 2002-2003 on child outcomes is likely to be if the feature influences child outcomes. The cross-site standard deviation of each treatment-contrast feature quantifies its cross-site variation. The greater this variation is, the more the effects of assignment to Head Start on child outcomes are likely to vary across sites, if the feature influences child outcomes.

With respect to the grand mean HSIS treatment contrast, note that (1) 86.6 percent of treatment group members versus 16.6 percent of control group members enrolled in Head Start during the present analysis period, for a difference of 70.0 percentage points; (2) 90.6 percent of treatment group members versus 49.3 percent of control group members were in some type of center care during this period, for a difference of 41.3 percentage points; and (3) 69.8 percent of treatment group members versus 27.0 percent of control group members were in high-quality, nonrelative care during this period, for a difference of 42.8 percentage points. In addition, treatment group members experienced 24.1 hours per week of center care versus 13.3 hours per week for control group members — a difference of 10.9 hours per week (numbers may not appear to sum correctly due to rounding). Thus all four of these treatment contrast features had a large treatment-control group difference. In contrast, only 33.5 percent of treatment group members versus 20.5 percent of control group members had a teacher with a bachelor's degree, for a difference of 13.1 percentage points.

What is most relevant for the present analysis, however, is the cross-site variation in these treatment-contrast features, which the findings in Table 2 indicate is *substantial*. For example, the estimated cross-site standard deviations for the four percentage measures range from 21.4 to 29.6 percentage points. To place these findings in perspective, note that if a given treatment-contrast feature were approximately normally distributed across Head Start centers, roughly 90 percent of centers would have a value that lies roughly within $\pm$ 41 percentage points of the grand mean. This is a very wide range.[18] In addition, the cross-site standard deviation of the treatment-control group difference in average weekly hours of center care is 4.6 hours per week. This is equivalent to an effect size of 0.27, which is substantial. Findings in Ta-

[18]Because of the bounds that exist for percentage measures, it is not possible for the cross-site distribution of the treatment-control group difference in the percentage of sample members who enrolled in Head Start to be normally distributed (given its estimated grand mean and cross-site standard deviation). Furthermore, given the present inability to identify the *shape* of the cross-site distribution of Head Start treatment-contrast features or program effects (discussed later), our use of a normal approximation is merely illustrative.

ble 2 thus provide strong reasons to expect substantial variation in Head Start effects during 2002-2003.

Individual Variation in Head Start Assignment Effects

Table 3 provides insights into the magnitude and nature of individual variation in Head Start effects during 2002-2003 by reporting the effect of assignment to Head Start on the individual-level residual variance for each cognitive and socio-emotional outcome that we examined.[19] These findings were obtained by estimating Equation 1 for each outcome and reporting the resulting estimates of residual variances for treatment group members and control group members (σ_T^2 and σ_C^2).

For receptive vocabulary, oral comprehension, and early numeracy, the residual variance for treatment group members is appreciably smaller than that for control group members, with a difference ranging from 13.2 percent to 22.5 percent (estimates of which are statistically significant). Because sample members were randomly assigned to the HSIS treatment group or control group, these differences are internally valid estimates of the effect of Head Start assignment on the heterogeneity of individual outcomes. For early reading and the two socioeconomic outcomes, the estimated variance effect is small in magnitude (ranging from a 1.4 percent to 5.0 percent reduction) and not statistically significant.

Given Head Start's remedial focus, a plausible explanation for the substantial observed variance reductions for the three cognitive outcomes is that Head Start increased them by more for children who, without the program, would have performed below average (perhaps well below average), thereby *compressing* the overall outcome distribution.[20] This explanation is consistent with Bitler, Hoynes, and Domina's (2014) quantile regression findings for the 3-year-old HSIS cohort, which indicate that Head Start assignment increased the mean of the bottom third of individual test scores by appreciably more than it increased the mean for the rest of the distribution. These quantile regression results were most pronounced for the two outcomes with the most pronounced program-induced variance reductions in Table 3 — receptive vocabulary and early numeracy. Furthermore, the variance reductions in Table 3 are consistent with Head Start effect estimates presented below for pretest performance-based subgroups.

[19] Any difference between the individual-level variances of outcomes for treatment group members and control group members must be caused by individual variation in Head Start effects (see Bloom et al., under review).

[20] Because one cannot identify the distribution of individual program effects without fairly strong assumptions, there is a virtual infinitude of alternative explanations for the variance reductions in Table 3 (Bloom et al., under review).

Thus during sample members' "Head Start year," it appears that the program produced compensatory effects on some important cognitive outcomes.

Subgroup Variation in Head Start Assignment Effects

Table 4 reports estimates of the grand mean effects of assignment to Head Start in 2002-2003 for selected HSIS subgroups. These findings were produced by estimating Equations 1-3 for each subgroup.[21] Estimated effects are reported as standardized mean difference effect sizes (effect sizes for short), which are measured as a multiple of the full sample, control group standard deviation for each outcome. The statistical significance level for each estimated effect in the table is denoted by stars and within each subgroup pair, effect-size estimates that differ statistically significantly at the 0.10 level are shaded in gray.

The first row in the table reports effect sizes for "low pretest performers" and the second row reports effect sizes for "other children" (all other sample members). Low pretest performers (treatment group members and control group members) are defined as sample members whose PPVT pretest scores were in the range spanned by the lower third of control group PPVT pretest scores, with cutoffs determined separately for members of the 3-year-old and 4-year-old HSIS cohorts. The PPVT pretest was used to define performance-based subgroups for all outcomes, which has two major benefits. First, it produces consistent subgroup samples that vary only slightly across outcomes due to modest differences in their attrition. Second, because the PPVT pretest was given only in English (unlike the Woodcock-Johnson pretest, which was given in English or Spanish), the PPVT pretest provides a baseline performance scale that is the same for all sample members.[22]

Note that the estimated effect sizes for receptive vocabulary and early numeracy are much larger for low pretest performers than for other children and the differences between these subgroup estimates are statistically significant. Specifically, the effect size for receptive vocabulary is 0.20 for low pretest performers versus 0.09 for other children and the effect size for early numeracy is 0.20 for low pretest performers versus 0.06 for other children.[23] These findings demonstrate a compensatory pattern of individual Head Start effects, which confirms

[21]Subgroup findings were produced by estimating Equations 1-3 for each subgroup from data for subgroup lotteries that were complete (they had at least one treatment group member and one control group member) and did not have zero compliance (the Head Start enrollment rate for their treatment group did not equal that for their control group).

[22]All post-tests were given in English only.

[23]We report effect sizes without stating their units (standard deviations) in order to avoid confusion when reporting the cross-site standard deviation of effect sizes, which is a standard deviation of a parameter that is measured in units of another standard deviation.

our compensatory interpretation of the program-induced individual-level variance reductions in Table 3.[24]

In terms of cognitive outcomes, the other striking findings in Table 4 are for dual language learners, Spanish-speaking children, and Hispanic children. The estimated Head Start effect size for each of these subgroups is much larger than that for its complement, and this difference is statistically significant.

Dual language learners were identified at their pretest based on responses by their teachers (if they were in a classroom) or their parents (if they were at home) to three questions about the language they prefer to speak.[25] Spanish-speaking children were identified by a response to a baseline parent interview question about the language spoken most frequently to the child at home. Hispanic children were identified by a response to a baseline parent interview question about whether their child was of Spanish, Hispanic, or Latino origin.

Almost all dual language learners in the present sample (96.4 percent) are Spanish-speaking children, and the vast majority of Spanish-speaking children (86.9 percent) are dual language learners. Hence these two subgroups overlap almost completely.[26] Hispanic sample members also overlap with these two subgroups (70.2 percent of Hispanic sample members are designated as Spanish-speaking children and 66.7 percent are designated as dual language learners). Thus the three subgroups share the fact that they had limited prior exposure to English.

The other subgroup findings of note in Table 4 indicate that Head Start in 2002-2003 improved both socio-emotional outcomes for girls but neither for boys. Estimated effect sizes for externalizing and self-regulation are -0.15 and 0.10 for girls (both of which are statistically significant) versus 0.01 and -0.01 for boys (neither of which is statistically significant). Note that a reduction in externalizing and an increase in self-regulation both represent an improve-

[24]Analyses that were conducted by the authors and are available upon request indicate that there is a compensatory pattern of Head Start effects on receptive vocabulary and early numeracy for both the 3-year-old HSIS cohort and the 4-year-old HSIS cohort. Thus pooling the data for the two cohorts does not distort these findings. Additional analyses, which were conducted by the authors and are available upon request, indicate that the treatment and control group difference in the mean value of the pretest covariate for each Head Start effect estimate is small to negligible, and the difference between this difference for the two pretest performance subgroups is negligible and not statistically significant for five of our six outcome measures. Thus the subgroup differences in estimated Head Start effects that we find cannot be attributed to subgroup differences in the influence of the pretest on these estimates.

[25]A child was considered to be a dual language learner (and the Woodcock-Johnson pretest was administered to him in Spanish rather than English) if his teacher or parent answered "Spanish" to two or more of the following questions: (1) "What language does the child speak most often at home?" (2) "What language does the child speak most often at this care setting?" (3) "What language does it appear the child prefers to speak?" Children who spoke neither English nor Spanish were not given the pretests used for the present analysis.

[26]Percentages reported in this paragraph are for students in the analysis sample for the PPVT post-test.

ment in behavior. These findings echo those of a recent meta-analysis which found that girls experience socio-emotional benefits from early education programs that are slightly larger than those experienced by boys (Kelchen et al., 2012). This might reflect that fact that boys tend to lag behind girls in their socio-emotional development and have different developmental trajectories (Else-Quest et al., 2006; Matthews, Ponitz, and Morrison, 2009).

One further issue to consider about the subgroup findings in Table 4 is the extent to which they represent subgroup differences in the effects of actually participating in Head Start rather than subgroup differences in compliance with HSIS random assignment. To address this issue, we estimated the effects of random assignment to Head Start on enrollment in the program (compliance with HSIS randomization) separately for each subgroup.[27] These findings, which are reported in Appendix Table D.1, indicate that subgroup compliance differences are too small to explain the observed subgroup differences in Head Start assignment effects.

Exploring Hypotheses About the Cognitive Subgroup Findings

The fact that Head Start effect sizes for receptive vocabulary and early numeracy — the two outcomes with the most pronounced compensatory pattern of individual Head Start effects — are most pronounced for subgroups with limited exposure to English suggests that this compensation might be largely for limited prior exposure to English. For example, attending Head Start might increase a child's English vocabulary beyond what it would have been otherwise, which in turn might increase his scores on the PPVT post-test of receptive vocabulary. This increased receptive vocabulary in English might in turn promote better understanding of (and thus more correct answers to) questions on the Woodcock-Johnson post-test of early numeracy.

However, a very different developmental hypothesis might explain the relationship between the especially large Head Start effects for dual language learners or Spanish-speaking children (and by extension, Hispanic children) and the observed compensatory pattern of Head Start effects. This hypothesis is based on findings from the neuroscience and child development literatures, which suggest that second-language or bilingual learning has cognitive benefits. For example, differences in brain organization have been identified between bilingual and monolingual children, and these differences tend to predict better cognitive development for bilingual children, particularly in the domain of executive function (Bialystok, 2011; Carlson and Meltzoff, 2008). Differences in executive function could be consequential because these skills have been found to promote academic achievement in general and mathematics achievement in particular (e.g., Duncan et al., 2007; Geary et al., 2007). In addition, although research suggests that bilingual children have smaller vocabularies than monolingual children (Bialystok, 2011;

[27]To do so, we estimated Equations 1-3 with a binary indicator of Head Start enrollment as the dependent variable for Equation 1. This estimation was repeated for each of our six outcome samples.

Bialystok et al., 2010), researchers have found that *growth* rates for word reading and oral language skills of children from lower-income Spanish-speaking homes can surpass those of native-born children (Mancilla-Martinez and Lesaux, 2011). Thus stronger Head Start effects for dual language learners and Spanish-speaking children might reflect a bilingual developmental *advantage* for these subgroups.

We present this hypothesis, however, with several caveats. First, in contrast to our preschool-aged, disadvantaged Head Start sample, studies of bilingual children have typically included school-aged samples of primarily middle-class students. Further, while some work suggests that the bilingual advantage might begin in infancy (Bialystok, 2011), it is not clear how quickly benefits of second language learning manifest themselves. Recall that the time between baseline and follow-up assessments in the present study ranged from three to seven months, with an average of five months. Consequently, it is not clear that findings from previous studies of bilingual learning advantages are generalizable to the present sample, nor is it clear that the timeframe of the present study is adequate for such a bilingual advantage to accrue.

With these caveats in mind, looking across the preceding results we see that:

1. The observed Head Start reduction of the individual-level outcome variance for receptive vocabulary and early numeracy suggests a compensatory pattern of individual program effects on these outcomes.

2. The fact that observed Head Start effects on receptive vocabulary and early numeracy are much larger for low-pretest performers than for other sample members confirms the existence of a compensatory pattern of individual Head Start effects on these outcomes.

3. The fact that observed Head Start effects on receptive vocabulary and early numeracy are much larger for dual language learners or Spanish-speaking children than for other sample members suggests two hypotheses: (A) that Head Start compensation is largely a compensation for limited prior exposure to English and/or (B) that bilingual learners have a Head Start learning advantage.

The problem that arises when trying to interpret these findings is that pretest-based subgroups are confounded with language-based subgroups. For example, low pretest performers are far more likely than other sample members to be dual language learners (44.3 versus 12.6 percent, respectively). Conversely, dual language learners are far more likely than other sample members to be low pretest performers (65.0 versus 25.2 percent, respectively).

To help break this confound, it is useful to subdivide each set of subgroups by the other set and report estimates of average Head Start effects for the resulting sub-subgroups. Thus Ta-

ble 5 reports estimates of grand mean ITT effects for dual language learners and English-only sample members by pretest performance sub-subgroups. For this table, dual language learners who are low pretest performers are defined as sample members with a PPVT pretest score that falls within the lower third of PPVT pretest scores for dual language control group members (with separate pretest cutoffs for 3- and 4-year-old cohort members). English-only sample members who are low pretest performers are defined as sample members with a PPVT pretest score that falls within the lower third of PPVT pretest scores for English language control group members (with separate pretest cutoffs for 3- and 4-year-old cohort members). Findings that differ statistically significantly between sub-subgroups at the 0.10 level are shaded in gray.

If the compensatory pattern of Head Start effects were due entirely to compensation for limited prior English (hypothesis A), this pattern would exist for dual language learners but not for other sample members. Findings in Table 5 indicate that indeed this is indeed the case. Dual language learners (96.8 percent of whom are Spanish speakers) exhibit a striking compensatory pattern whereas English-only sample members exhibit no such pattern. This suggests that the compensatory pattern of Head Start effects represents compensation for limited prior English.[28]

To explore this issue further, Table 6 presents findings for sub-subgroups that are defined differently from their counterparts in Table 5. Low pretest performers are now defined as sample members with a PPVT pretest score that falls within the lower third of PPVT pretest scores for *all* control group members (with separate pretest cutoffs for 3- and 4-year-old cohort members). Findings in Table 6 indicate a striking "dual language advantage" for low pretest performers but not for other sample members. This pattern is the *opposite* of what one would expect from a bilingual Head Start learning advantage because, to the extent that any sample members are truly operating in a bilingual mode, it should be children who know enough English to do so.

Together the results in Tables 5 and 6 indicate (1) a compensatory pattern of Head Start effects for dual language learners but not for other children (which is consistent with Head Start compensation for limited prior English), and (2) a "dual language learning advantage" for low pretest performers but not for other children (which is inconsistent with a bilingual learning advantage). Furthermore, the observed dual language learning advantage for low pretest performers might actually reflect Head Start compensation for limited prior English because the average pretest score (in English only) for dual language learners among low pretest performers is substantially lower than that for their English-only counterparts. The magnitude of this difference

[28]To assess whether the large positive estimated effects reported in Table 5 for low-pretest dual language learners might be due to chance baseline imbalance, Appendix Table E.1 compares baseline characteristics for treatment and control group members in the sub-subgroup. These findings indicate no baseline imbalance problem. Appendix Tables E.2 through E.4 present similar findings for the three other sub-subgroups.

stated as an effect size is 0.31 for 3-year-old cohort members and 0.27 for 4-year-old cohort members.[29]

Consequently, it appears that the much larger than average Head Start effects observed for dual language learners and Spanish-speaking children is more likely to reflect compensation for limited prior exposure to English than it is to reflect a bilingual learning advantage. However, many other factors could be at play here, including differences in the Head Start centers and counterfactual care settings experienced by members of each subgroup. To explore some of these alternative explanations, Table 7 presents estimates of key features of the HSIS treatment contrast separately for dual language learners and English-only sample members and by pretest status. These findings suggest that differences in the HSIS treatment contrast between dual language learners and English-only learners by pretest status caused by (1) differences in their local Head Start programs, (2) differences in their local alternative programs, or (3) differences in their propensities to choose these options *do not explain* the striking subgroup differences that exist in their effects of Head Start.

For example, the ITT effect of Head Start on the percentage of children in nonrelative care that was rated to be of high quality (i.e., had an ECERS-R score of 5 or greater) was actually smaller for dual language learners with low pretest scores than for the rest of the sample. Overall, the magnitude of the differences in the treatment contrast across sub-subgroups are too small to explain the fact that Head Start effects on receptive vocabulary and early numeracy are much larger for dual language learners with low pretest scores than for the other sub-subgroups. (See Table 5.) In other words, these differences cannot be explained by larger compliance with Head Start, a greater shift into center care, a greater increase in hourly weeks of center care, or a greater shift into high-quality center care for dual language learners.

Yet another potential explanation for the preceding pattern of subgroup findings is that low-pretest dual language learners attend Head Start centers that might be more effective overall (relative to their local competing alternatives) than the centers attended by other sample members. To test this hypothesis, we reestimated grand mean program effects separately for low-pretest dual language learners and all other sample members for the subset of Head Start centers that had a randomized trial for *both subgroups* (44 centers for receptive vocabulary and 43 centers for early numeracy).[30] We further refined this analysis by giving the two subgroups from

[29]We report these differences by age cohort because we use cohort-specific pretest performance thresholds.

[30]For the 44 Head Start centers in the receptive vocabulary reanalysis there were 289 low-pretest dual language learners and 419 other children. For the 43 Head Start centers in the early numeracy reanalysis there were 286 low-pretest dual language learners and 399 other children. The threshold for a low pretest used for this analysis was the same for dual language learners and other sample members.

each center the same weight.[31] When we thereby constrained the two subgroup samples to represent the *same Head Start centers*, our estimates of ITT effects on receptive vocabulary and early numeracy were between 0.24 and 0.29 (and statistically significant) for low-pretest dual language learners and between 0.00 and 0.05 (and not statistically significant) for all other sample members. Thus unobserved differences in the overall effectiveness of Head Start centers attended by low-pretest dual language learners and all other sample members cannot explain the striking observed differences in their Head Start effects.

The important policy and developmental question underlying these results is whether stronger effects for low-pretest dual language learners represent meaningful learning for children or whether they are simply learning the language of the post-test (English). To examine this question, we estimated effects separately by language and pretest status for receptive vocabulary and mathematics outcomes from data for the *third HSIS follow-up wave*. By this point in time, the majority of children in the 4-year-old cohort were in first grade and the majority of children in the 3-year-old cohort were in kindergarten. Thus presumably all control group members now had extensive exposure to English. If stronger short-term effects for low-pretest dual language learners were due only to knowing the language of the test, we would expect the control group to fully catch up to the treatment group once control group members were also immersed in English-language schooling.

Results of this analysis are shown in Table 8, where we also repeat our estimates of the spring 2003 (end of Head Start) effects to facilitate a comparison of short-term and medium-term program effects. As can be seen, at the third follow-up wave, though not statistically significant, effects for receptive vocabulary and early mathematics are still largest for dual language learners with low pretest scores. In fact, dual language learners with low pretest scores are the only group to show any lasting boost from Head Start on mathematics. These results suggest that Head Start may have had an effect on the development of dual language learners with low pretest scores that had implications beyond their being taught the language of the HSIS assessments. Because 97.6 percent of low-pretest dual language learners are Spanish-speaking children and 94.6 percent of low-pretest Spanish-speaking children are dual language learners, the conclusions that we draw for each of these subgroups hold for the other.

[31]This weight was established for each center by (1) computing the error variance of the impact estimate for every subgroup from every center in the reanalysis, (2) setting the error variance for the two subgroups from each center equal to their mean value for that center, and (3) using a random effects meta-analysis (often referred to as "V-known estimation") to estimate the grand mean program effect for each subgroup. This model weights the impact estimate for each center inversely proportionally to the sum of its average estimation error (which is the same for the two subgroups from each center) plus the estimated true cross-block variance of program effects (which is the same for all subgroups from all centers). Doing so ensures that the cross-center distribution of the weight of the data in the reanalysis is the *same* for low-pretest dual language learners and all other sample members.

Cross-Site Variation in Head Start Assignment Effects and Participation Effects

Table 9 reports estimates of the cross-site grand mean and cross-site standard deviation of Head Start effect sizes in 2002-2003. The first panel in the table reports results for effects of Head Start assignment (ITT). These findings were produced by estimating Equations 1-3 for each outcome measure.[32] To test the sensitivity of these estimates to the fact that baseline pretests were administered in the fall of 2002 after Head Start participants had begun the program, Appendix Table F.1 compares the ITT results in Table 7, which were estimated using a pretest covariate, with their counterparts estimated without a baseline covariate. There are no substantial or systematic differences between the two sets of estimates.

The second panel in Table 9 reports results for effects of Head Start participation (LATE). These findings were produced using the procedure described in Appendix C.[33] The third panel reports a rough approximation of the likely percentage of Head Start centers with a negative ITT effect size for each outcome (i.e., the percentage of centers where the effect of Head Start assignment on that outcome is less than the effect of local alternatives, including parent care). The basis for this approximation and an important caveat are explained below. The final panel in the table reports the number of children and Head Start centers in the present analysis sample for each outcome.

Assignment Effects (ITT)

The estimated grand mean ITT effect size is positive and statistically significant for three of the four cognitive outcomes examined, with magnitudes ranging from 0.12 to 0.17. The corresponding estimate for the fourth cognitive outcome, oral comprehension, is near zero (0.01).[34] This latter finding might reflect limited syntactic complexity in the child-directed speech of Head Start teachers. This is consistent with prior findings that indicate that even though young children are rapidly developing syntactic complexity skills and this development is sensitive to teacher inputs, many preschool teachers use little complex syntax with their students (Huttenlocher et al., 2002; Justice et al., 2013).

The estimated grand mean effect size for socio-emotional outcomes is -0.05 (which is statistically significant) for externalizing and 0.02 (which is not statistically significant) for self-

[32]The estimated cross-site grand mean effect size for Head Start assignment is the estimated value of β_0 in Equation 3. The estimated cross-site standard deviation of Head Start effect sizes is the estimated value of τ_{ITT} for Equation 3.

[33]The estimated cross-site grand mean effect size for Head Start participation is the estimated value of δ_0 in appendix Equation C.6. The estimated cross-site standard deviation of Head Start effect sizes is the estimated value of τ_{LATE} for appendix Equation C.6.

[34]The preceding findings are consistent with those in the HSIS Final Report (Puma et al., 2010a).

regulation.[35] These modest program-induced improvements in socio-emotional skills should not be considered discouraging, because past research suggests that center-based care can sometimes produce *adverse* effects on these outcomes (e.g., Loeb et al., 2007; Magnuson, Ruhm, and Waldfogel, 2007; NICHD Early Child Care Research Network, 2003). On the other hand, when early care settings are of high quality, null or positive effects have been reported for these outcomes (Gormley et al., 2011; Loeb et al., 2004; Votruba-Drzal, Coley, and Chase-Lansdale, 2004). As shown in Table 2, the majority of HSIS treatment-group members were in high-quality centers (centers with an ECERS-R score of 5 or higher).

Of greater interest for the present analysis, however, is that the estimated cross-site standard deviation of Head Start ITT effect sizes is substantial and statistically significant for five of the six outcomes examined, with magnitudes ranging from 0.12 to 0.25. The smallest estimate (for early numeracy) is 0.07 and is not statistically significant. Thus for all but one outcome, the present findings provide strong evidence of substantial cross-site variation in the effects of assignment to Head Start relative to the effects of counterfactual care settings experienced by control group members.

Figure 1 graphically illustrates this variation for the five outcome measures with statistically significant estimates of cross-site variation in Head Start effects. This was done using histograms of "adjusted empirical Bayes estimates" (described in Appendix H) of ITT effect sizes for each Head Start center. Although this approach has limitations, which are discussed below, it is a useful way to illustrate a rough approximation to a cross-site distribution of "true" program effects.

Consider the distribution of PPVT-ITT effect sizes. The estimated cross-site grand mean for this outcome is 0.14, and its estimated cross-site standard deviation is 0.12. If effect sizes for the outcome are approximately normally distributed across Head Start centers, then about 90 percent of centers would have a PPVT-ITT effect size between -0.06 and 0.34 and about 12 percent would have a negative PPVT-ITT effect size (i.e., the Head Start center would be less effective than its local alternatives at promoting receptive vocabulary). The PPVT-ITT distribution in Figure 1 illustrates this pattern. Its peak (mean, median, and mode) is approximately 0.15, and only about 9 percent of the distribution lies below zero (according to findings reported in the bottom row of Table 9). Thus even though the data indicate substantial cross-site variation in PPVT-ITT effect sizes, they suggest that the overwhelming majority of these effect sizes are positive and that most of the small number of negative effect sizes are modest in magnitude.

[35]Findings for these outcomes were not included in the HSIS Final Report (Puma et al., 2010a).

The distribution of ITT effect sizes for oral comprehension represents a very different situation, with a near-zero grand mean (0.01) and a large cross-site standard deviation (0.12). Consequently, 43.2 percent of its adjusted empirical Bayes estimates are negative and 56.8 percent are positive. If true oral comprehension effect sizes are approximately normally distributed across Head Start centers, then about 90 percent of them would lie between negative 0.19 and positive 0.21. Consequently, the near-zero mean effect size for this outcome masks a wide range of positive and negative program effects.

A similar result was obtained for the two socio-emotional outcomes, given their near-zero grand mean effect sizes (-0.05 and 0.02) and large cross-site effect size standard deviations (0.16 and 0.22). Thus it appears that Head Start centers range from substantially more effective to substantially less effective than their local alternatives, including parent care, at improving socio-emotional outcomes.

However, even though one can be fairly confident about present estimates of the cross-site grand mean and standard deviation of Head Start effect sizes, one cannot be as confident about the *shape* of the cross-site distribution of these effect sizes.[36] This is because the cross-site distribution of effect size estimates (be they OLS estimates, empirical Bayes estimates, or adjusted empirical Bayes estimates) is a *combination* of two distributions: (1) the cross-site distribution of "true" Head Start effect sizes and (2) the cross-site distribution of site-level estimation error due to within-site randomization. In many (if not most) cases, the cross-site distribution of site-level estimation error for OLS estimates of ITT effects will be approximately normal. This is because these estimates represent some form of a treatment-control group difference in means, the error for which is approximately normally distributed when either (1) the distribution of individual outcomes is approximately normal or (2) site samples are large enough for the Central Limit Theorem to overcome departures from normally distributed individual outcomes.[37] However, the cross-site distribution of true program effects can have almost any shape. Because the cross-site distribution of site-level estimation error is confounded with the cross-site distribution of true program effects, it is not possible to distinguish the shapes of the two distributions without further assumptions and/or more complex analytic methods.

Nonetheless, the present findings clearly indicate that there is substantial cross-site variation in ITT effects of Head Start on important child outcomes. In addition, these findings

[36]The authors thank Professor Henry May of the University of Delaware for bringing this issue to our attention.

[37]The further the distribution of individual outcomes departs from normality, the larger site samples must be in order for the Central Limit Theorem to produce a normal distribution of site-level estimation error. However, even for dichotomous individual outcomes, the distributions of which are highly nonnormal, site-level estimation error can be approximately normally distributed for surprisingly small samples. This fact can be easily demonstrated through simulations.

strongly suggest that for at least the three outcome measures with near-zero grand mean effect-size estimates (oral comprehension, externalizing, and self-regulation), Head Start centers range from substantially more effective to substantially less effective than their local alternatives.

Participation Effects (LATE)

The second panel in Table 9 reports estimates of cross-site grand mean effects of participating in Head Start (δ_0 in appendix Equation C.6) and cross-site standard deviations of effects of participating in Head Start (τ_{LATE} for appendix Equation C.6) in 2002-2003.[38] Note that for five of the six outcomes considered, the estimated grand mean effect of Head Start participation is larger than its counterparts for Head Start assignment. This reflects the extent to which non-compliance with random assignment (no-shows among treatment group members and cross-overs among control group members) "diluted" the HSIS treatment contrast and thereby reduced its ITT effects.

Findings in the second panel of Table 9 also indicate that there is substantial cross-site variation in Head Start participation effects. The estimated cross-site standard deviation of Head Start participation effects is large (ranging from 0.14 to 0.28) and statistically significant for the five outcomes with corresponding results for Head Start assignment. These findings make it possible to assess the extent to which cross-site variation in Head Start assignment effects (ITT) is due to cross-site variation in compliance with random assignment versus cross-site variation in the actual effects of participating in Head Start. A finding of little cross-site variation in Head Start participation effects would imply that most of the variation in Head Start assignment effects is due to variation in compliance. However, the present findings indicate that variation in the effectiveness of Head Start centers relative to their local alternatives is the primary source of cross-site variation in effects of Head Start assignment.[39]

To conclude the present analysis, Table 10 links key findings about cross-site variation in Head Start effects to key findings about subgroup variation in Head Start effects by addressing the question: "To what extent is cross-site variation in program effects *predicted* by cross-site variation in the representation of low-pretest dual language learners?" To do so, Table 10 reports results obtained by replacing Equation 3 in our two-level model of Head Start ITT effects with the following level-two *predictive* model.

---

[38]As noted earlier, estimates of Head Start participation effects (LATE) were obtained using an extension of the approach presented by Raudenbush, Reardon, and Nomi (2012).

[39]The ratio of LATE to ITT grand mean effect estimates varies across outcomes, which is possible for a multi-instrument estimation strategy like the present one. However, this ratio would be constant across outcomes if a single instrument were used. As a sensitivity test, the authors reestimated the LATE findings in Table 7 using a single instrument and obtained results that are similar to those reported here. Appendix Table C.1 reports these alternative estimates.

$$\beta_j = \beta_0^* + \pi \cdot LPDLL_j + r_j^* \quad (4)$$

where:

β_j = the mean Head Start ITT effect-size for center *j*,

$LPDLL_j$ = the percentage of sample members from center *j* who were low-pretest dual language learners,

r_j^* = a random error that varies independently and identically across centers with a mean of zero and a variance of τ_{ITT}^{2*}.

The first row in Table 10 reports, for each of our six outcomes, the estimated intercept ($\hat{\beta}_0^*$) for Equation 4. This represents the grand mean Head Start ITT effect size for centers with *zero* low-pretest dual language learners. The second row in the table reports, for each outcome, the estimated slope ($\hat{\pi}$) for Equation 4. This represents the rate of change in the mean Head Start ITT effect size *per percentage point increase* in low-pretest dual language learners.

We should expect the representation of low-pretest dual language learners to predict Head Start effects only for receptive vocabulary, because it is our only outcome with both cross-site impact variation *and* a pronounced differential effect for low-pretest dual language learners.[40] As can be seen, this is in fact the case because receptive vocabulary has an estimated slope that is highly statistically significant (p = 0.004), whereas none of the other outcomes do so (their p-values are 0.990, 0.129, 0.910, 0.470, and 0.119, respectively).

The estimated intercept for receptive vocabulary indicates that its grand mean ITT effect size for centers with *no* low-pretest dual language learners is 0.10. The estimated slope for this outcome indicates that the grand mean ITT effect size increases by 0.003 per percentage-point increase in low-pretest dual language learners. This implies that (1) the grand mean ITT effect size is 0.149 for centers with the mean percentage of low-pretest dual language learners (16.2 percent) and (2) the grand mean ITT effect size is 0.40 for centers with 100 percent low-pretest dual language learners.[41] These findings are consistent with the grand mean effect size of 0.14 reported in Table 9 and the pronounced differential subsubgroup effect sizes reported in Tables 5 and 6.

[40]Our only other outcome measure with a pronounced differential effect for low-pretest dual language learners — early numeracy — has no discernible cross-variation in Head Start effects (see Table 9).

[41]The cross-site distribution of the percentage of treatment group members who were low-pretest dual language learners ranges from zero percent for 211 Head Start centers to 100 percent for 6 Head Start centers, with a wide range of values in between. Thus our interpretation of the estimated parameters for Equation 4 does not extrapolate beyond the distribution of the present data.

By comparing our estimate of the "unconditional" cross-site variance of Head Start effect sizes ($\tau^2_{ITT} = 0.01346$) for receptive vocabulary (obtained from Equations 1-3) with our estimate of the "conditional" cross-site variance of Head Start effect sizes ($\tau^{2*}_{ITT} = 0.01138$) for receptive vocabulary (obtained from Equations 1, 2, and 4) we estimate that about 15.5 *percent* of the cross-site effect-size variance for this outcome is "explained" by cross-site variation in the representation of low-pretest dual language learners.[42]

Because our differential subgroup findings suggest that Head Start compensated some participants for their limited prior exposure to English, the findings discussed above suggest that for one of our six outcome measures — receptive vocabulary — *a small portion of cross-site variation in Head Start effects might be explained by cross-site variation in Head Start compensation for limited prior English.* We do not purport to explain the remainder of the cross-site variation in Head Start effects for this outcome or any of the cross-site variation in Head Start effects for other outcomes.

Discussion

The preceding results confirm that, as hypothesized by much prior research (e.g., Barnett, 2007; Currie, 2007; Ludwig and Phillips, 2007; Zaslow, 2008), there is substantial variation in the effects of Head Start (at least during 2002-2003) on multiple measures of children's cognitive and socio-emotional school readiness skills. This variation is manifest across individual children, across policy-relevant subgroups of children, and across Head Start centers (program sites).

With respect to variation across individuals and subgroups, there appears to be a *compensatory* pattern of Head Start effects on cognitive outcomes, which is consistent with the program's mission of serving this country's most disadvantaged young children (Zigler and Styfco, 2010). The pattern was observed two different ways. First, Head Start reduced the individual-level residual outcome variance for early language, literacy, and mathematics skills. Hence it reduced disparities in these outcomes for program-eligible children. Second, Head Start effects on early language and mathematics skills were much larger for children with low baseline language skills (low pretest performers) than for other children, which explains how the program reduced the variances of these outcomes. This finding is consistent with the results of Bitler,

[42]Our results for the model represented by Equations 1-3 indicate that the "unconditional" cross-site variance of Head Start effect sizes for receptive vocabulary ($\hat{\tau}^2_{ITT}$) equals 0.01346. Our results for the model represented by Equations 1, 2, and 4 indicate that the "conditional" cross-site variance of Head Start effect sizes for receptive vocabulary ($\hat{\tau}^{2*}_{ITT}$) equals 0.01138. The fact that $\left(1 - \frac{\tau^{2*}_{ITT}}{\tau^2_{ITT}}\right) = \left(1 - \frac{0.01138}{0.01346}\right) = 0.155$ implies that 15.5 percent of the cross-site variation in Head Start effect sizes for receptive vocabulary is predicted by corresponding variation in the representation of low-pretest dual language learners.

Hoynes, and Domina (2014) who find that Head Start produced its largest cognitive effects on the low end of the cognitive outcome distribution. Thus it appears that several key Head Start cognitive effects are largest for participants with the "most room to grow."

However, these findings do not necessarily represent a case for targeting Head Start on program-eligible children with the weakest skills. This is because reducing disparities *within* the population of low-income disadvantaged children who are eligible for Head Start (the present finding) is not the same as reducing disparities *between* this disadvantaged population and other children (the goal of Head Start). Thus maximizing the effectiveness of Head Start for *all* eligible children is more in line with the program's goal than is maximizing Head Start's *average* effectiveness by targeting a particularly responsive eligible subpopulation.

Further, stronger effects for those with the "most room to grow" do not help to resolve debate regarding whether early education programs are more effective for children with lower baseline skills (Sameroff and Chandler, 1975) or higher baseline skills (Heckman, 2000), in part because both theories are mute regarding the match between the intervention and children's initial skills. Instead, we hypothesize, following Vygotskian theories of learning and child development (1978), that the program match for children matters more than do children's baseline skills. That is, Head Start, with its explicit remedial focus, may have offered lower-skilled but not higher-skilled children a "zone of proximal development" (Vygotsky, 1978), or support for obtainable learning goals, just beyond what they already knew. Empirically, repeating content that children already know — which may have been the case with the higher-skilled children in Head Start — has been found to be negatively associated with children's mathematics development in kindergarten (Engels, Claessens, and Finch, 2013).

In addition, we find that Head Start effects are much larger for dual language learners and Spanish-speaking children (two highly overlapping subgroups) than for other sample members, especially for receptive vocabulary and early numeracy (the two outcomes with the most pronounced compensatory patterns of program effects). Further analysis suggests that the much larger than average Head Start effects for dual language learners, Spanish-speaking children, and low pretest performers probably represents Head Start compensation for *their limited prior exposure to English*. This, in turn, markedly improves post-test performance, especially for receptive vocabulary and early numeracy. We examined several alternative explanations for stronger effects for low-pretest dual language learners and found little support that they were driven by differences in the treatment contrast, differences in local alternative programs, or differences in propensities to choose different care settings.

We took steps to examine whether this apparent early improvement in English for a subset of Head Start participants represents the beginning for them of a long-lasting improvement in cognitive outcomes or simply an early improvement in their ability to take tests in Eng-

lish. To do so, we examined effects of Head Start by language and pretest status at the third wave of HSIS data collection, when sample children were nearing the completion of kindergarten or first grade and thus control group members also had considerable exposure to English-language instruction. We found that positive treatment effects for dual language learners with low pretest scores persisted even after dual language learners in the control group learned English, which suggests that Head Start's positive effects on this sub-subgroup were more profound than simply improving its members' ability to take English-language tests. However, estimates of these effects were no longer statistically significant at the third follow-up wave. Nonetheless, this evidence provides support for efforts to enroll more dual language learners (especially those with very limited English skills) in Head Start or programs like it. This step could be particularly important given that nationally, dual language learners (particularly those from Spanish-speaking households, like most HSIS dual language learners) are proportionally underenrolled in preschool programs, and their enrollment rates have declined even further in recent years (Fuller and Kim, 2011; Weiland and Yoshikawa, 2013).

The substantial *cross-site variation* that we observe for Head Start effects in 2002-2003 on five of the six outcomes we examined verifies what has been hypothesized for decades: that this large-scale, nationally funded, locally implemented program (with 1,800 grantees and 16,000 centers at present) produces results that vary widely relative to those of competing local alternatives. This is the case both for outcomes with substantial grand mean effects and for outcomes with near-zero grand mean effects. Such variation can be produced by differences in the effectiveness of local Head Start programs, differences in the effectiveness of local program alternatives (including parent care), or both.

For example, although, on average, Head Start in 2002-2003 was more effective than its local alternatives at reducing externalizing behaviors, this average masks the finding that many Head Start centers were less effective than their local alternatives at improving this outcome. The latter finding is consistent with prior research that suggests that center-based care can sometimes have adverse effects on externalizing behavior (Loeb et al., 2007; Magnuson, Ruhm, and Waldfogel, 2007; NICHD Early Child Care Research Network, 2002), and that these adverse effects are most likely to occur in low-quality center-based care (Vandell et al., 2010; Votruba-Drzal, Coley, and Chase-Lansdale, 2004). The finding is also consistent with research demonstrating that relative Head Start effects depend on the nature of the counterfactual care setting with which the program is compared (for example, center-based care or parent care, Feller et al., 2014). In addition, the finding is consistent with our findings that the HSIS treatment contrast varies substantially across Head Start centers in terms of program dosage, teacher education, and classroom quality.

The one outcome measure with a positive grand mean Head Start effect size and little or no observable cross-site variation is early numeracy. We hypothesize that this finding reflects

little cross-site variation in the depth and breadth of early mathematics taught by Head Start, other preschools, and children's parents during 2002-2003. For example, the literature suggests that preschool teachers tend to (1) feel less comfortable teaching math than teaching language or literacy (Ginsburg, Lee, and Boyd, 2008), (2) spend less time teaching math than teaching other topics (Early et al., 2010), and (3) limit their math instruction to simple skills such as counting and recognizing shapes or numerals (Clements and Sarama, 2007). This is the case even though research has shown that effective preschool mathematics curricula allocate more time to mathematics and cover a richer and deeper set of mathematical skills than is typical for preschool classrooms (Clements et al., 2013; Ginsburg, Lee, and Boyd, 2008; Starkey, Klein, and Wakeley, 2004). However, because these curricula were not widely available in 2002-2003, there was little room for variation in preschool mathematics curricula at that time. Furthermore, at home, without specific training interventions, low-income parents tend to provide little support for the development of young children's mathematical skills (Starkey and Klein, 2000).

Taken together, our cross-site findings are relevant to policy concerns that the high quality and large effects of small, model preschool programs studied in the past cannot be maintained when such programs are taken to a large scale (Burke and Sheffield, 2013). For example, recent data suggest that while Head Start and state prekindergarten programs provide good classroom emotional support, their average classroom instructional support is not adequate (Office of Head Start, 2013; Mashburn et al., 2008). Nonetheless, our findings suggest that although the average effectiveness of Head Start programs could be improved (as evidenced by the encouraging results for the most effective Head Start centers in the HSIS sample), the Head Start program in 2002-2003 outperformed its local alternatives on average overall and at the majority of Head Start centers in terms of effects on children's early language, early literacy, early numeracy, externalizing, and self-regulatory skills.[43]

In closing, it is important to note some important limitations of our findings. First, it is beyond the scope of the present analysis to determine the extent to which the substantial cross-site variation in Head Start effects that we observe represents variation in *overall program effectiveness* rather than variation in *outcome-specific program effectiveness*. Strong cross-site associations among Head Start effects on different outcomes would suggest that some Head Start centers are relatively stronger overall than others. Weak cross-site associations among Head Start effects on different outcomes would suggest that some Head Start centers are relatively stronger than others with respect to specific cognitive or socio-emotional outcomes.

[43]As noted earlier, this result assumes that true Head Start effects on these outcomes are distributed approximately normally across sites.

To properly document these associations, however, requires complex estimation methods that account for random error (lack of reliability) in site-specific estimates of Head Start effects. This is necessary in order to estimate cross-site associations among true program effects, not just cross-site variation in program effect estimates.[44] Because such methods have not yet been developed and tested on data like those for the present analysis, the generality versus outcome-specific nature of cross-site variation in Head Start effects remains an interesting question for future research.

Second, Head Start and its alternatives have changed a great deal during the past decade. Hence findings for 2002-2003 might not generalize fully to the current program. For example, most Head Start centers in 2002-2003 lacked the supports and investments that recent studies show are critical for achieving large program effects — especially focused, domain-specific curriculum and coaching for teachers (Yoshikawa et al., 2013). These are features that are currently being introduced to some local Head Start programs. In addition, most Head Start teachers in 2002-2003 did not have a bachelor's degree, which is required for the demonstrably successful preschool programs in Boston and Tulsa, required for the federal Preschool for All Plan, and currently emphasized by the national Head Start program (Improving Head Start for School Readiness Act, 2007). Furthermore, since 2002-2003, there has been a major expansion of state prekindergarten programs and widespread merging of funding for Head Start and these programs. Thus many children who receive funding from these different sources attend the same preschools and are in the same classrooms.

Third, it is not currently possible to rigorously compare the observed cross-site variation in Head Start effects with that for other related programs. This is because the present methodology has not yet been used widely on data for multisite trials. Just as Cohen (1988) and then Hill and colleagues (2008) established empirical benchmarks for interpreting average program effect sizes, researchers need comparable future studies to establish yardsticks for cross-site variation in effect sizes.

Meanwhile it is useful to compare our estimates of variation in effect sizes *across Head Start centers* with the effect-size variation that exists *across past studies* of early child care and education programs. Appendix G describes how we used a random-effects meta-analysis to

[44] A simple correlation between site-specific OLS estimates of Head Start assignment effects on two outcome measures will understate (attenuate) the corresponding cross-site correlation between true Head Start assignment effects. This attenuation is due to random estimation error (lack of reliability) in the site-specific estimates. If one tried to account for this lack of reliability by correlating site-specific empirical Bayes "shrinkage" estimates, this would impart a positive bias to the resulting estimated cross-site correlation because shrinkage affects estimates of site-specific effects for different outcomes similarly. Bayesian modeling might produce better estimates, but these methods have not yet (to our knowledge) been tested on data like those for the present analysis.

produce two estimates of the latter from data provided by the National Forum on Early Childhood Policy and Programs Meta-Analysis Project.[45] One of our estimates was based on information about effect sizes on cognitive outcomes and their standard errors for 31 treatment-control group contrasts from 24 studies of Head Start programs that operated between 1963 and 2002. The other estimate was based on corresponding information for 74 treatment-control group contrasts from 55 studies of early child care and education programs (including Head Start) that operated between 1961 and 2009. Because these studies estimated the effects of *participating* in a specific early education program relative to a counterfactual mix of child care and education alternatives, including parent care,[46] their findings are arguably most comparable to our estimates of the effects of Head Start *participation* (LATE).

The estimated standard deviation of "true" effect sizes on cognitive outcomes for the 31 treatment-control group contrasts from 24 Head Start studies is 0.27 (p-value = 0.099). The corresponding finding for the 74 treatment-control group contrasts from 55 studies of early child care and education programs (including Head Start) is 0.28 (p-value < 0.001). These findings represent different types of programs, programs that were operated at different times and in different environments, programs that served different populations, studies that focused on different outcome measures, and studies that used different methodologies. Hence they should reflect a great deal of variation in findings.

Estimates in Table 9 of the cross-site standard deviation of Head Start participation effect sizes for three of our six outcome measures — early reading, oral comprehension, and self-regulation — range from 0.20 to 0.28, which is close to the preceding meta-analytic findings. Estimates for the two other outcomes with statistically significant cross-site variation (receptive vocabulary and externalizing) are 0.14 and 0.15, which is substantial but not as large as those for our meta-analytic findings. Thus it appears that in 2002-2003 there was a great deal of variation across Head Start centers in program effects on cognitive and socio-emotional outcomes.

In conclusion, we would like to add two further thoughts. First, we hope that in addition to providing valuable information for the national Head Start program and its local operations, the present paper will serve as a model for detecting, quantifying, reporting, and interpreting variation in program effects using data from multisite trials. Second, we note that although, with one exception, identifying predictors of cross-site variation in Head Start effects is beyond the scope of the present paper, other researchers are taking up this charge (e.g., Friedman-Krauss,

[45]See Duncan and Magnuson (2013), Grindal et al. (2013), Leak et al. (2012), Schindler et al. (2013), and Shager et al. (2013) for analytic papers produced from this meta-analytic database.

[46]The present meta-analysis does not include studies that compared alternative versions of Head Start with each other or compared Head Start with some other specific program.

Connors, Morris, et al., 2014; Feller et al., 2014; Peck and Bell, 2014; Walters, 2014). In addition, there is considerable interest in similar research for other educational and social programs.

To advance this larger research agenda it will be necessary to develop realistic but practical conceptual frameworks and program theories. Likewise, it will be essential for the next generation of multisite trials to collect high-quality data on the core elements of these conceptual frameworks and to be designed in ways that facilitate the study of variation in program effects. Furthermore, it will be necessary to continue to develop new analytic methods that can make the most of this information. Together, these new statistical methods, conceptual frameworks, and data collection systems can produce the accumulation of knowledge that is needed by policymakers, practitioners, and researchers to move beyond average impacts and to improve future programs.

Exhibits

Table 1

Baseline balance of the analysis sample

Baseline characteristic	Mean value of the baseline characteristic				
	Treatment group	Control group	Difference		P-value of difference
Pretest results					
Receptive vocabulary (PPVT)	249.1	252.1	-3.0	**	0.037
Early reading (WJ-LW)	300.9	300.2	0.7		0.469
Oral comprehension (WJ-OC)[1]	N/A	N/A	N/A		N/A
Early numeracy (WJ-AP)	377.1	377.2	-0.1		0.924
Externalizing (parent reports)	1.7	1.7	0.0	**	0.028
Self-regulation (assessor reports)	3.1	3.1	0.0		0.807
Child characteristics					
Male (%)	49.2	50.3	-1.0		0.557
Black (%)	31.1	30.3	0.9		0.287
Hispanic (%)	36.6	36.6	0.0		0.980
English is home language (%)	70.9	71.1	-0.2		0.869
Family characteristics					
Mother's age (years)	29.2	28.9	0.3		0.265
Mother has less than HS education (%)	36.4	40.0	-3.6	**	0.031
Mother has HS education (%)	34.1	31.5	2.6		0.123
Mother is married (%)	44.4	45.8	-1.4		0.409
Mother was previously married (%)	16.2	15.3	0.9		0.483
Mother is a teenager (%)	16.1	17.4	-1.3		0.337
Mother is a recent immigrant (%)	17.7	18.7	-1.0		0.367
Child lives with both biological parents (%)	49.7	49.7	0.1		0.974
Assessment characteristics					
Child age at spring testing	4.0	4.0	0.0		0.980
Spring child assessment date[2]	32.6	33.8	-1.2	***	0.000
Child was tested in English (%)	75.0	75.7	-0.7		0.496
Spring parent interview date[2]	33.5	33.9	-0.4	***	0.000

Notes: Sample includes children in complete randomized blocks with nonzero compliance and with nonmissing WJ-LW outcome data. The largest possible N is 3,529 (nonmissing WJ-LW outcome data and in a complete, nonzero compliance randomized block). N = 3,529 for the following baseline characteristics: male, child age at spring testing, child was tested in English, spring parent interview date. Sample sizes for other baseline characteristics were as follows: receptive vocabulary (3,097), early reading (3,076), early numeracy (2,304), externalizing (3,217), self-regulation (3,156), black (3,513), Hispanic (3,513), English is home language (3,501), mother's age (3,519), mother has less than HS education (3,401), mother has HS education (3,401), mother is married (3,403), mother was previously married (3,403), mother is a teenager (3,219), mother is a recent immigrant (3,488), and child lives with both biological parents (3,431).

[1]Pretest data were not collected for this outcome.

[2]In weeks since September 1, 2002.

Table 2

HSIS treatment contrast for the present sample

	Treatment group grand mean	Control group grand mean	Difference	P-value of difference	Cross-site standard deviation of difference	P-value of cross-site standard deviation
Percentage in Head Start	86.6	16.6	70.0***	<0.0001	22.3***	<0.0001
Percentage in any center care	90.6	49.3	41.3***	<0.0001	21.4***	<0.0001
Average weekly hours in center care[1]	24.1	13.3	10.9***	<0.0001	4.6***	<0.0001
Percentage with teacher who has a BA	33.5	20.5	13.1***	<0.0001	29.6***	<0.0001
Percentage in nonrelative care with an ECERS-R score of 5 or greater	69.8	27.0	42.8***	<0.0001	28.4***	<0.0001

Notes: Samples include children in complete randomized blocks with nonzero compliance and nonmissing WJ-LW outcome data. Estimation models used as covariates: nonresidualized pretest scores, standard HSIS covariates, a binary indicator for age cohorts, and fixed intercepts for Head Start centers. For all percentage outcomes, the cross-site standard deviation is expressed in percentage points. The standard deviation for hours is expressed in hours.

*** = p<0.01

[1]This variable was set equal to zero for sample members who were not in center care.

Table 3

Individual-level residual variances for treatment and control group members

Outcome measure	Treatment group		Control group		Difference		Percentage difference
Cognitive outcomes							
Receptive vocabulary (PPVT)	526	***	679	***	-153	***	22.5
Early reading (WJ-LW)	430	***	436	***	-6		1.4
Oral comprehension (WJ-OC)	112	***	129	***	-17	***	13.2
Early numeracy (WJ-AP)	455	***	563	***	-108	***	19.2
Socio-emotional outcomes							
Externalizing (parent reports)	0.102	***	0.106	***	-0.004		3.8
Self-regulation (assessor reports)	0.383	***	0.403	***	-0.020		5.0

Note: Samples include children in complete randomized blocks with nonzero compliance and nonmissing outcome data. Estimation models used as covariates: nonresidualized pretest scores, standard HSIS covariates, a binary indicator for age cohorts, and fixed intercepts for Head Start centers.

*** = $p<0.01$

Table 4

Grand mean ITT effect sizes, by subgroup

	Cognitive outcomes									*Socio-emotional outcomes*				
	Receptive vocabulary (PPVT)		Early reading (WJ-LW)		Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)		Externalizing (parent reports)		Self-regulation (assessor reports)				
Full-sample grand mean	0.15	***	0.17	***	0.01	0.12	***	-0.05	*	0.02				
Pretest performance														
Low pretest performers	0.20	***	0.16	**	0.03	0.20	***	-0.10		0.00				
Other children	0.09	***	0.18	***	-0.02	0.06	*	-0.09	**	0.02				
Dual language status														
Dual language learner	0.26	***	0.23	***	-0.01	0.30	***	-0.09		0.02				
English only	0.10	***	0.15	***	0.02	0.06	**	-0.05		0.00				
Home language														
Spanish	0.27	***	0.20	***	-0.02	0.28	***	-0.06		0.01				
Other	0.10	***	0.17	***	0.03	0.04		-0.06		0.00				
Special needs														
Yes	0.24	**	0.12		0.07	0.09		-0.06		0.00				
No	0.14	***	0.19	***	0.01	0.12	***	-0.07	**	0.05				
Age cohort														
Age 3	0.15	***	0.20	***	0.02	0.12	***	-0.11	**	0.01				
Age 4	0.11	***	0.16	***	-0.02	0.11	***	-0.01		0.05				
Gender														
Male	0.17	***	0.17	***	0.03	0.09	**	0.01		-0.01				
Female	0.19	***	0.22	***	0.00	0.14	***	-0.15	***	0.10	**			

(continued)

Table 4 (continued)

Race/ethnicity											
Black	0.04		0.19	***	-0.02	0.03		0.02		0.04	
Hispanic	0.24	***	0.18	***	0.03	0.21	***	-0.08	*	0.01	
White/other	0.12	***	0.16	***	0.03	0.04		-0.05		0.01	

Notes: Within each subgroup, models were fit: using children with available outcome data in nonzero compliance, complete randomized blocks; including the standard HSIS covariates; using fixed intercepts for Head Start centers; using the appropriate nonresidualized pretest; using data from both age cohorts; and including a control for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect for each outcome in its original units by the control-group standard deviation for that outcome. Statistically significant impact differences between subgroups for a given outcome are indicated with shading ($p<0.10$). Statistical significance of differences in subgroup impacts was determined by a t-test of the interaction between the subgroup characteristic and the treatment variable. For race/ethnicity, the overall statistical significance of differences among subgroups was determined by an omnibus test of the joint statistical significance of interactions between the coefficients for treatment interacted with multiple subgroup indicators. For children with nonmissing outcome data, missing data were imputed once, except for the relevant subgroup characteristic, which was not imputed.

* = p<0.10, ** = p<0.05, *** = p<0.01

Table 5

Grand mean ITT effect sizes for dual language learners and other sample members, by pretest performance subgroup

Subgroup	Estimated ITT effect size: Receptive vocabulary (PPVT)		Estimated ITT effect size: Early numeracy (WJ-AP)	
Dual language learners[1]				
Low pretest performers	0.44	***	0.38	**
Other sample members	0.17	***	0.11	
English-only sample members[1]				
Low pretest performers	0.13	**	0.03	
Other sample members	0.09	***	0.07	**

Notes: Within each subgroup, models were fit: using children with available outcome data in nonzero compliance and complete randomized blocks; including the standard HSIS covariates; using fixed intercepts for centers; using the appropriate nonresidualized pretest; using data from both cohorts; and including a binary indicator for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect on each outcome in its original units by the control group standard deviation for that outcome.

* = p<0.10, ** = p<0.05, *** = p<0.01

[1]Dual language learners who are low pretest performers have a PPVT pretest score that falls within the lower third of PPVT pretest scores for dual language control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members. English-only sample members who are low pretest performers have a PPVT pretest score that falls within the lower third of PPVT pretest scores for English-only control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members. Findings that differ statistically significantly between sub-subgroups at the 0.10 level are shaded in gray.

Table 6

Grand mean ITT effect sizes for low pretest performers and other sample members, by language subgroup

Subgroup	Estimated ITT effect size: Receptive vocabulary (PPVT)		Estimated ITT effect size: Early numeracy (WJ-AP)	
Low pretest performers[1]				
Dual language learners	0.30	***	0.30	***
English-only sample members	0.05		0.05	
Other sample members				
Dual language learners	0.05		-0.09	
English-only sample members	0.09	***	0.06	*

Notes: Within each relevant subgroup, models were fit: using children with available outcome data in nonzero compliance and complete randomized blocks; including the standard HSIS covariates; using fixed intercepts for centers; using the appropriate nonresidualized pretest; using data from both cohorts; and including a binary indicator for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect on each outcome in its original units by the control group standard deviation for that outcome.

* = $p<0.10$, ** = $p<0.05$, *** = $p<0.01$

[1]Low pretest performers have a PPVT pretest score that falls within the lower third of PPVT pretest scores for *all* control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members. Findings that differ statistically significantly between sub-subgroups at the 0.10 level are shaded in gray.

Table 7

Grand mean ITT effects on features of the HSIS treatment contrast for dual language learners and English-only learners, by pretest performance subgroup

Subgroup	Estimated ITT effect: Percentage in Head Start	Percentage in any center care	Average weekly hours in center care[1]	Percentage with a teacher who has a BA	Percentage in nonrelative care with an ECERS-R of 5 or greater	Percentage in parent care
Dual language learners						
Low pretest performers	84.7***	53.5***	11.8***	11.1	44.9***	-40.8***
Other sample members	79.1***	44.0***	12.5***	0.62	50.9***	-35.3***
English-only sample members						
Low pretest performers	80.7***	49.2***	13.9***	12.5**	45.5***	-30.6***
Other sample members	74.1***	45.6***	11.3***	17.6***	47.2***	-32.3***

Notes: Within each subgroup, samples include children in complete randomized blocks with nonzero compliance and nonmissing WJ-LW outcome data. Estimation models used as covariates: nonresidualized pretest scores, standard HSIS covariates, a binary indicator for age cohorts, and fixed intercepts for Head Start centers. Statistically significant impact differences between subgroups for a given outcome are shaded in gray ($p<0.10$). Statistical significance of differences in subgroup impacts was determined by a t-test of the interaction between the subgroup characteristic and the treatment variable. For children with nonmissing outcome data and nonmissing data on their age cohort, missing baseline data were imputed once.

*** = p<0.01

[1]This variable was set equal to zero for sample members who were not in center care.

Table 8

Receptive vocabulary and mathematics grand mean ITT effect sizes for all sample members and by dual language learner and pretest performance status, spring 2003 and spring 2005

	Estimated ITT effect size							
	End of HS year (spring 03)				End of K/1st (spring 05)			
Subgroup	Receptive vocabulary (PPVT)		Early numeracy (WJ-AP)		Receptive vocabulary (PPVT)		Early numeracy (WJ-AP)	
All sample members	0.14	***	0.12	***	0.04		-0.00	
Dual language status								
Dual-language learner	0.26	***	0.30	***	0.05		0.12	
English only	0.10	***	0.06	**	0.02		-0.05	
Dual language learners[1]								
Low pretest performers	0.44	***	0.38	**	0.13		0.13	
Other sample members	0.17	***	0.11		-0.09		0.01	
English-only sample members[1]								
Low pretest performers	0.13	**	0.03		0.09		-0.01	
Other sample members	0.09	***	0.07	**	0.06	*	-0.02	

Notes: Within each subgroup, models were fit: using children with available outcome data in nonzero compliance and complete randomized blocks; including the standard HSIS covariates; using fixed intercepts for centers; using the appropriate nonresidualized pretest; using data from both cohorts; and including a binary indicator for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect on each outcome in its original units by the control group standard deviation for that outcome.

* = $p<0.10$, ** = $p<0.05$, *** = $p<0.01$

[1]Dual language learners who are low pretest performers have a PPVT pretest score that falls within the lower third of PPVT pretest scores for dual language control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members. English-only sample members who are low pretest performers have a PPVT pretest score that falls within the lower third of PPVT pretest scores for English-only control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members. Findings that differ statistically significantly between sub-subgroups at the 0.10 level are shaded in gray.

Table 9

Cross-site grand means and standard deviations for Head Start effect sizes (ITT and LATE)

	Cognitive outcomes				Socio-emotional outcomes	
	Receptive vocabulary (PPVT)	Early reading (WJ-LW)	Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)	Externalizing (parent reports)	Self-regulation (assessor reports)
Effects of assignment to Head Start (ITT)						
Grand mean	0.14*** (<0.001)	0.17*** (<0.001)	0.01 (0.625)	0.12*** (<0.001)	-0.05* (0.088)	0.02 (0.568)
Standard deviation	0.12** (0.030)	0.25*** (<0.001)	0.12* (0.097)	0.07 (0.230)	0.16*** (0.009)	0.22** (0.014)
Effects of participation in Head Start (LATE)						
Grand mean	0.17*** (<0.001)	0.25*** (<0.001)	0.03 (0.354)	0.15*** (<0.001)	-0.07* (0.052)	0.02 (0.572)
Standard deviation	0.15** (0.004)	0.26** (0.002)	0.20* (0.057)	0.00 (0.560)	0.14** (0.019)	0.28** (0.009)
Percentage of HS centers that are less effective than their local alternatives	9	23	43	--	36	45
N children	3,523	3,529	3,465	3,491	3,524	3,486
N centers	297	297	296	296	297	295

Notes: Within each relevant subgroup, models were fit: using children with available outcome data in nonzero compliance and complete randomized blocks; including the standard HSIS covariates; using fixed intercepts for centers; using the appropriate nonresidualized pretest; using data from both cohorts; and including a binary indicator for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect on each outcome in its original units by the control group standard deviation for that outcome. P-values are in parentheses below each parameter estimate.

* = p<0.10, ** = p<0.05, *** = p<0.01

Table 10

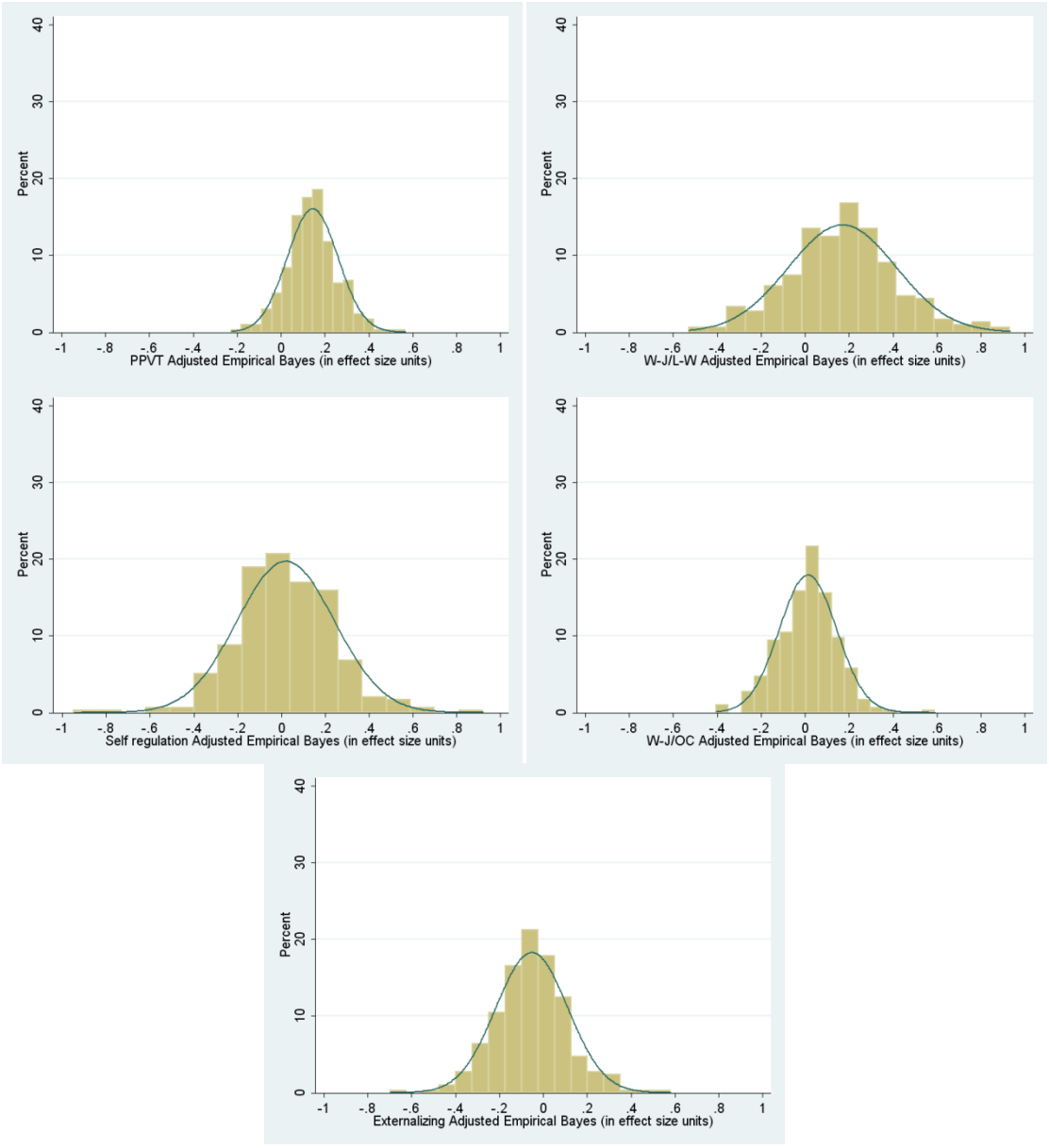
Predicting mean Head Start ITT effect sizes with the percentage of sample members who are low-pretest dual language learners

	Cognitive outcomes				Socio-emotional outcomes	
	Receptive vocabulary (PPVT)	Early reading (WJ-LW)	Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)	Externalizing (parent reports)	Self-regulation (assessor reports)
Intercept (β_0^*)	0.101*** (0.000)	0.165*** (<0.001)	0.076** (0.019)	0.011 (0.725)	-0.070* (0.070)	-0.024 (0.553)
Slope (π)	0.003*** (0.004)	0.000 (0.990)	0.002 (0.129)	0.000 (0.910)	0.001 (0.470)	0.002 (0.119)

Note: Findings in this table were obtained for each outcome by estimating the two-level random-coefficients model represented by Equations 1, 2, and 4.

Figure 1

Inferred cross-site distributions of Head Start ITT effect sizes by outcome measure



Appendix A

Departures from the HSIS Analysis

The present analysis differs in several ways from that for the HSIS Final Report (Puma et al., 2010a). In this analysis, we pool data for the HSIS 3- and 4-year-old cohorts in order to maximize statistical power. In contrast, the HSIS reported findings separately for the two age cohorts. We believe that pooling data for the two age cohorts is justified for several reasons. First, 3- and 4-year-olds were randomized together in a single block per Head Start center, not separately in two blocks per center. Second, many members of the two age cohorts experienced Head Start together; 43 percent of Head Start classrooms in the HSIS served both sample cohorts.[1] Third, the two age cohorts partially overlap across sites because they were defined in terms of the month that determines a child's eligibility for kindergarten, which varies locally. Fourth, during the study's first follow-up year (the focus of the present analysis) the two age cohorts have similar rates of compliance with randomization (71 percent and 67 percent). Last, the two age cohorts experienced similar Head Start effects on five of the six outcomes examined.[2]

A second departure from the HSIS Final Report is that the present analysis does not use the HSIS sampling weights that were developed to extrapolate the study's findings to the 2002-2003 national population of oversubscribed Head Start centers (Puma et al., 2010b, Chapter 2). Not using these weights made it possible to avoid the ambiguity that exists about how they were created, which was greatly complicated by the need for the weights to account for many different facets of the HSIS sampling process. Not using these weights also made it possible to avoid the added complexity that would result from their use when computing statistical tests for analyses of variation in Head Start effects. Fortunately, these weights have little effect on HSIS point estimates of average program effects and only increase their standard errors (Bloom and Weiland, 2014).[3]

The present analysis also differs from that for the HSIS Final Report by not "residualizing" pretests when using them as covariates to improve statistical precision. Instead we use actual pretest values. The HSIS residualized pretests by computing them as a deviation from the control group mean for control group members and as a deviation from the treatment group mean for treatment group members. This was done to ensure that the pretest would not influence estimates of Head Start effects, which was a concern because pretests were administered after the beginning of the Head Start year.[4] But if a covariate cannot influence an estimate it

[1]This 43 percent figure probably understates the percentage of sample members who were in multiage classrooms, because children of varying age who were not in the present sample shared classrooms with its sample members.

[2]Table 4 presents these findings with those for other subgroups of children.

[3]This increase in standard errors reflects the well-known phenomenon that weighting an estimate to extrapolate a finding beyond a sample reduces precision.

[4]According to the HSIS Final Report (Puma et al, 2010a, pp. 2-57), "The 'residualization' procedure … removes any systematic differences between treatment and control group levels in the fall measures [the pretest], including those potentially due to Head Start's impact."

cannot influence its precision, and this is not properly accounted for by the regression models used by the HSIS to estimate Head Start effects. This did not materially affect HSIS results, however (Bloom and Weiland, 2014). In recognition of this issue, the recent HSIS third-grade follow-up report uses nonresidualized pretests (Puma et al., 2012).

Two other differences between the present analysis and the HSIS Final Report are worth noting. First, the present estimation model, which focuses on cross-site variation in Head Start effects, has a random-coefficients specification, whereas the HSIS Final Report model, which focused on national average program effects, has a fixed-coefficient specification. Second, the HSIS Final Report used hot-decking to impute missing baseline data, whereas the present analysis uses a single replicate of a multiple imputation model for this purpose. Neither of these differences changes the basic story reflected by the findings obtained.

Appendix B

Further Detail About Our Outcome Measures

This appendix presents details about the four cognitive and two socio-emotional outcome measures used for the present analysis.

Cognitive Measures

During the Head Start year, HSIS sample members were assessed on a battery of 14 cognitive outcome measures. These assessments were typically administered individually to sample members in their primary care setting by a specially trained assessor (with the exception of the Emergent Literacy Scale, which was a parent report). For parsimony, we study four of the 14 measures. Six measures were eliminated because they are nonstandardized, had limited psychometric evidence of validity, or had scoring problems which led the HSIS team to exclude them from its reports.[1] From the remaining eight measures, we chose the following four, which are most commonly used in early childhood research and tap domains of child development that have been shown to predict later outcomes.[2]

1. *Receptive vocabulary* measured by the Peabody Picture Vocabulary Test-III (PPVT; Dunn and Dunn, 1997)

2. *Early reading* measured by the Woodcock-Johnson Letter-Word Identification subscale (WJ-LW; Woodcock, McGrew, and Mather, 2001)

3. *Oral comprehension* measured by the Woodcock-Johnson Oral Comprehension subscale (WJ-OC; Woodcock, McGrew, and Mather, 2001)

4. *Early numeracy* measured by the Woodcock-Johnson Applied Problems subscale (WJ-AP; Woodcock, McGrew, and Mather, 2001)

The PPVT measures children's receptive vocabulary skills. It is a nationally normed measure that has been used widely for diverse samples of young children (e.g., Weiland and Yoshikawa, 2013; Wong et al., 2008). The measure has excellent split-half and test-retest reliability plus strong qualitative and quantitative indicators of validity (Dunn and Dunn, 1997). The PPVT requires children to choose (verbally or nonverbally) which of four pictures best

[1]The following tests used in the HSIS have no or limited published reliability information: Color Identification, Counting Bears, the Emergent Literacy Scale, and Letter Naming (see pp. 2-25 to 2-30 of the HSIS Final Report [Puma et al., 2010a] for more details). Two additional tests — the Preschool Comprehensive Test of Phonological and Print Processing (CTOPPP) Print Awareness Subtest and the Story and Print Concepts test — were not included in analysis by the original HSIS team due to problems in scoring and interpreting their results (see pp. 3-5 and 3-7 of the HSIS Technical Report [Puma et al., 2010b] for more details).

[2]Three of these outcomes were measured as a pretest in the fall of 2002 *and* as a post-test in the spring of 2003. The fourth outcome (Woodcock-Johnson Oral Comprehension) was measured only as a post-test in the spring of 2003.

represents a spoken stimulus word. The HSIS administered a short version of the PPVT that was adapted using Item Response Theory (IRT). The present analysis employs this measure.[3]

Sample members' scores on Woodcock-Johnson subscales were obtained from the Woodcock-Johnson Test of Achievement III. These subscales are nationally normed and widely used, although the Oral Comprehension subtest is used less frequently than the other subtests.[4] The Letter-Word Identification subscale measures children's early reading skills and reflects their ability to identify and pronounce isolated written letters and words. Its test-retest reliability is 0.96 (Woodcock, McGrew, and Mather, 2001). The Oral Comprehension subscale represents children's oral comprehension skills, such as their ability to understand a short spoken passage and provide missing words based on syntactic and semantic clues. Its test-retest reliability is 0.82 (Woodcock, McGrew, and Mather, 2001). The Applied Problems subscale represents children's early numeracy and mathematics skills based on their ability to perform simple calculations and solve simple arithmetic problems. Its test-retest reliability is 0.90 (Woodcock, McGrew, and Mather, 2001).[5]

To reduce the time needed to administer the Woodcock-Johnson subscales, the HSIS used a three-item stop rule, instead of the six-item stop rule recommended by the test's developers (Puma et al., 2010b). This modification might have reduced average scores for sample members but should not have differentially affected treatment and control group scores.

Socio-Emotional Measures

The present analysis uses two outcomes from this domain.

- *Externalizing behavior problems* measured by the Child Behavior Checklist (Achenbach, Edelbrock, and Howell, 1987)
- *Self regulation skills* measured by the Leiter-R Assessor Report (Roid and Miller, 1997)

[3]The HSIS used three-parameter IRT models to score children's test results. PPVT-IRT scores have an advantage over non-IRT PPVT scores because the former correct for guessing. Details about the shortened PPVT and its IRT scoring are available from the Head Start Impact Study Technical Report (Puma et al., 2010b; pp. 3-16 to 3-19).

[4] For examples, see Gormley et al., 2005; Lipsey et al., 2013; Peisner-Feinberg et al., 2001; Weiland and Yoshikawa, 2013; Wong et al., 2008; and Woodcock, McGrew, and Mather, 2001.

[5]This subtest does not measure geometric and spatial capacities, and some researchers have raised concerns about its comprehensiveness, appropriateness, and sensitivity for use with young children (Clements, Sarama and Liu, 2008).

A composite measure of externalizing problems that reflects children's aggressive and hyperactive behavior was constructed based on parent responses to seven items on the Child Behavior Checklist.[6] This instrument is used widely to assess early childhood social-emotional functioning (Duncan et al., 2007; Raver et al., 2009). Our composite measure (alpha = 0.71) reverse-codes each item so that higher scores represent more severe problems. We focus on externalizing problems, rather than some other socio-emotional outcomes in the HSIS, for several reasons. First, other parent-reported socio-emotional measures collected for the HSIS have somewhat lower internal consistency.[7] Second, externalizing is arguably the most substantively important socio-emotional construct assessed for the HSIS. Externalizing behaviors are relatively stable during childhood (Campbell, 1995); they are associated with underachievement in adolescence (Hinshaw, 1992); and when occurring in early childhood, they are considered a major risk factor for juvenile delinquency, adult crime, and violence (Liu, 2004; Moffitt, 1993). In addition, some preschool interventions have been shown to reduce externalizing behaviors, so they are potentially malleable (Raver et al., 2009; Schindler et al., 2013).

The present analysis also uses the only measure administered by the HSIS to assess self-regulation skills — the Leiter-R Assessor Report. This report was completed by assessors after they tested children's cognitive skills. The self-regulation measure is an average of assessor ratings of children's task persistence, attention span, body movement, and attention to direction (alpha = 0.82). Self-regulation skills are an important developmental domain, and empirical studies have demonstrated that these skills are sensitive to preschool experiences (Morris et al., 2013; Raver et al., 2011; Weiland and Yoshikawa, 2013).

Data Coverage Rates

Data coverage rates for post-tests (Appendix Table B.1) are very high for treatment group members (ranging from 86.6 percent to 87.8 percent) and moderately high for control group members (ranging from 74.0 percent to 78.3 percent). The 8.3 to 12.8 percentage-point difference in these coverage rates for treatment and control group members probably reflects the fact that it was more difficult to locate control group members, who were widely dispersed, than it was to locate treatment group members, who were concentrated at Head Start centers.

Data coverage rates for four of the five pretests that were administered are high for treatment group members — ranging from 88.5 percent to 92.3 percent — and somewhat lower for control group members — ranging from 81.8 percent to 87.4 percent (Appendix Table

[6]For each sample member with data for at least five of these seven items, we averaged item scores to create a composite measure. For other sample members we coded the measure as missing.

[7]For example, measures of internal consistency (alphas) for social competencies and approaches to learning were 0.60 and 0.63, respectively.

B.2).[8] Data coverage rates for all other baseline covariates (Appendix Table B.2) range from 87 percent to 100 percent for treatment and control group members. Missing values for baseline covariates were imputed from a single replicate of a multiple imputation model using all baseline and follow-up data, including treatment assignment status.[9]

[8]Pretest data were not obtained for oral comprehension, and pretest data coverage rates for early numeracy were 65.6 percent for treatment group members and 60.6 percent for control group members.

[9]A single replicate instead of multiple replicates was used to impute missing covariate values in order to minimize complexity. We believe that this decision is justified because (1) the presence or total absence of covariates had only a modest effect on our results, (2) there was little missing data for covariates, and (3) sensitivity tests of alternative methods for imputing missing covariate values (including the use of a binary missing-data indicator without imputation) indicate that our results are robust.

Appendix Table B.1

Data coverage rates for each outcome measure

Outcome measure	Percentage with data for the measure: Treatment group	Control group	Difference	P-value of difference
Follow-up outcome measure (post-test)				
Receptive vocabulary (PPVT)	87.6	76.0	11.6***	<0.001
Early reading (WJ-LW)	87.8	76.1	11.7***	<0.001
Oral comprehension (WJ-OC)	87.4	75.5	11.9***	<0.001
Early numeracy (WJ-AP)	86.8	74.0	12.8***	<0.001
Externalizing (parent report)	86.6	78.3	8.3***	<0.001
Self-regulation (assessor report)	87.3	75.8	11.6***	<0.001

Note: Sample includes all children who were in complete lotteries with nonzero compliance. N = 4,315 children in 318 centers.

Appendix Table B.2

Data coverage rates for each baseline covariate

Covariate	Mean value of covariate: Treatment group	Control group	Difference		P-value of difference
Pretest					
Receptive vocabulary (PPVT)	89.5	81.8	7.8	***	<0.001
Early reading (WJ-LW)	88.5	82.0	6.5	***	<0.001
Oral comprehension (WJ-OC)[1]	N/A	N/A	N/A		N/A
Early numeracy (WJ-AP)	65.6	60.6	5.0	***	0.000
Externalizing (parent report)	92.3	87.4	4.9	***	<0.001
Self-regulation (assessor report)	90.9	84.1	6.8	***	<0.001
Child characteristic					
Male (%)	100.0	100.0	0.0		--
Black (%)	99.7	99.3	0.4	*	0.099
Hispanic (%)	99.7	99.3	0.4	*	0.099
English is home language (%)	99.4	99.2	0.2		0.449
Family characteristic					
Mother's age (years)	99.8	99.6	0.2		0.350
Mother has less than HS education (%)	96.5	95.0	1.5	**	0.019
Mother has HS education (%)	96.5	95.0	1.5	**	0.019
Mother is married (%)	96.5	95.3	1.2	*	0.074
Mother was previously married (%)	96.5	95.3	1.2	*	0.074
Mother is a teenager (%)	92.4	87.4	5.1	***	<0.001
Mother is a recent immigrant (%)	98.9	98.6	0.2		0.556
Child lives with both biological parents (%)	97.2	96.2	1.0	*	0.071
Assessment characteristic					
Child age at spring testing	98.4	97.4	1.0		0.033
Spring child assessment date[2]	98.4	97.4	1.0		0.033
Child was tested in English (%)	98.4	97.4	1.0		0.033
Spring parent interview date[2]	100.0	100.0	0.0		--

Note: Sample includes all children who were in complete lotteries with nonzero compliance and who had data for any of the six outcomes. N = 3,785 children in 297 Head Start centers.

[1]Pretest data were not collected for this outcome.

[2]In weeks since September 1, 2002.

Appendix C

Estimating Head Start Participation Effects (LATE)

The following approach, which builds on that developed by Raudenbush, Reardon, and Nomi (2012), was used to estimate the grand mean effect of participating (enrolling) in Head Start and the variation in these effects across Head Start centers. This estimation was conducted in two steps. Step one applied two-stage least squares (2SLS) to a multiple instrumental variables model in order to estimate the mean effect of Head Start enrollment for children from each Head Start center. Step two employed a random-effects meta-analysis (referred to as V-known estimation in SAS or HLM) to estimate the cross-center grand mean effect of Head Start enrollment and its cross-center variance or standard deviation.[1]

Step One: Site-Specific Estimation

This section describes how we obtained site-specific point estimates of mean Head Start effects and their estimated standard errors.

Point Estimates

Equations C.1 and C.2 below represent the 2SLS instrumental variables model used to estimate the mean effect of Head Start enrollment for each Head Start center.[2] The first stage of this model estimates the effect of random assignment to Head Start (T) on whether an individual child enrolled in Head Start (E) during the present follow-up period. The second stage estimates the effect of Head Start enrollment on a follow-up outcome (Y).

First-stage child-level model

$$E_{ij} = \sum_{m=1}^{J} \theta_m \cdot C_{mj} + \sum_{m=1}^{J} \gamma_m \cdot T_{ij} \cdot C_{mj} + \sum_{k=1}^{K} \psi_k \cdot X_{kij} + e_{1ij} \qquad \text{(C.1)}$$

Second-stage child-level model

$$Y_{ij} = \sum_{m=1}^{J} \phi_m \cdot C_{mj} + \sum_{m=1}^{J} \delta_m \cdot \hat{E}_{ij} \cdot C_{mj} + \sum_{k=1}^{K} \eta_k \cdot X_{kij} + e_{2ij} \qquad \text{(C.2)}$$

where:

E_{ij} = one if child i from Head Start center j enrolled in Head Start during the present follow-up period and zero otherwise,

C_{mj} = one if Head Start center j is Head Start center m and zero if not,

[1] We test the statistical significance of estimates of τ^2_{LATE} (and thus τ_{LATE}) using the conventional Q statistic for random-effects meta-analyses (Hedges and Olkin, 1985).

[2] It was necessary to develop a two-stage least squares procedure that could accommodate the large number of interactions in our model and its two individual-level residual outcome variances.

T_{ij} = one if child *i* from Head Start center *j* was randomly assigned to the program and zero if not,

X_{kij} = baseline characteristic *k* for child *i* from Head Start center *j*,

$\hat{E}_{ij}$ = the predicted value of Head Start enrollment (based on the estimated parameters for Equation C.1) for child *i* from Head Start center *j*,

e_{1ij} = a random error that is independently distributed across individuals, with a mean of zero and a variance of σ_{1T}^2 for treatment group members and σ_{1C}^2 for control group members,

e_{2ij} = a random error that is independently distributed across individuals with a mean of zero and a variance of σ_{2T}^2 for treatment group members and σ_{2C}^2 for control group members.

The parameter γ_m in Equation C.1 is the mean effect of random assignment to Head Start on the probability of enrolling in Head Start for children from Head Start center *m* (referred to elsewhere as center *j*). Parameter δ_m in Equation C.2 is the mean effect of Head Start enrollment on the outcome for children from Head Start center m.

Two-stage least squares estimation was implemented by:

1. using OLS to estimate the parameters of Equation C.1,
2. using these parameter estimates to compute the probability of Head Start enrollment or "predicted enrollment" ($\hat{E}_{ij}$) for each sample member,
3. substituting these predicted enrollment values into Equation C.2 and estimating its parameters using OLS, and
4. estimating standard errors of the parameter estimates for Equation C.2 using an approach developed by Brachet (2007) and described below.

This process produced consistent estimates of the mean effect of Head Start enrollment for each Head Start center ($\hat{\delta}_j$) and its standard error ($\widehat{se}(\hat{\delta}_j)$), the square of which is its estimated error variance ($\hat{V}_j$).

Standard Errors

We obtained valid estimates of 2SLS standard errors for Equation C.2 using SAS PROC MIXED to estimate the following OLS regression, which uses the independent variables

in Equation C.2 with a different dependent variable. For a proof of the method, see Brachet (2007).[3]

$$RE_{ij} = \sum_{m=1}^{J} \phi_m^* \cdot C_{mj} + \sum_{m=1}^{J} \delta_m^* \cdot \hat{E}_{ij} \cdot C_{mj} + \sum_{k=1}^{K} \eta_k^* \cdot X_{kij} + e_{2ij}^* \qquad \text{(C.3)}$$

where:

RE_{ij} = the 2SLS residual (defined below) for child *i* from Head Start center *j*,

C_{mj} = one if Head Start center *j* is Head Start center *m* and zero if not,

$\hat{E}_{ij}$ = the predicted value of Head Start enrollment, based on the estimated parameters of Equation C.1, for child *i* from Head Start center *j*,

X_{kij} = baseline characteristic *k* for child *i* from Head Start center *j*,

e_{2ij}^* = a random error that is independently distributed across individuals with a mean of zero and separate variances for treatment group members and control group members.

The estimated standard errors for Equation C.3 are valid estimates of the standard errors for corresponding parameter estimates in Equation C.2.

Note that RE_i is the residual that would result from predicting each sample member's outcome using the parameter estimates from Equation C.1 with the actual values of the Head Start enrollment indicator for each sample member (E_{ij}) instead of its predicted values ($\hat{E}_{ij}$). In symbols:

$$RE_{ij} = Y_{ij} - \hat{Y}_{ij} \qquad \text{(C.4)}$$

where:

$$\hat{Y}_{ij} \equiv \sum_{m=1}^{J} \hat{\phi}_m \cdot C_{mj} + \sum_{m=1}^{J} \hat{\delta}_m \cdot E_{ij} \cdot C_{mj} + \sum_{k=1}^{K} \hat{\eta}_k \cdot X_{kij} \qquad \text{(C.5)}$$

and the estimated parameters used for Equation C.5 are those obtained by estimating Equation C.2.

[3]The procedure described in this appendix is based on work by Brachet (2007) that demonstrates how to use SAS to estimate 2SLS standard errors that account for the clustering of sample members. This approach is quite general, however, and can be used with or without accounting for clustering. We used PROC MIXED to implement the procedure because it can emulate OLS and estimate separate treatment and control group residual variances. If multiple residual variances are not necessary, any OLS routine will suffice.

Step Two: Cross-Site Estimation

The next step was to input the values of $\hat{\delta}_j$ and $\hat{V}_j$ to a random-effects meta-analysis and estimate the following model of cross-site variation in Head Start enrollment effects.

Center-level model

$$\delta_j = \delta_0 + w_j \quad \text{(C.6)}$$

where:

δ_j = the mean effect of Head Start enrollment on the outcome for children from Head Start center j,

δ_0 = the cross-site grand mean effect of Head Start enrollment on the outcome,

w_j = a random error that varies independently and identically across Head Start centers, with a mean of zero and a variance of τ^2_{LATE}.

The random-effects meta-analysis produces consistent estimates of our parameters of interest, δ_0 and τ^2_{LATE}.

Alternative Estimates Using a Single Instrument

As a sensitivity analysis, we reestimated the cross-site grand mean and standard deviation of the effects of Head Start participation using a single instrument (random assignment to the HSIS treatment group or control group) based on the approach presented as "Option B" by Raudenbush, Reardon, and Nomi (2012). Appendix Table C.1 reports these findings. Note that Option B does not provide p-values for estimates of the cross-site standard deviation of program effects.

Appendix Table C.1

Alternative estimates of cross-site grand means and standard deviations of Head Start participation effects (LATE) using a single instrument

	Cognitive outcomes				Socio-emotional outcomes	
	Receptive vocabulary (PPVT)	Early reading (WJ-LW)	Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)	Externalizing (parent reports)	Self-regulation (assessor reports)
Grand mean	0.21*** (<0.001)	0.24*** (<0.001)	0.02 (0.625)	0.16*** (<0.001)	-0.07* (0.088)	0.03 (0.568)
Standard deviation[1]	0.15 N/A	0.33 N/A	0.17 N/A	0.08 N/A	0.22 N/A	0.30 N/A
N children	3,523	3,529	3,465	3,491	3,524	3,486
N centers	297	297	296	296	297	295

Notes: Samples include all children from both HSIS age cohorts who had available outcome data and were part of a complete randomized block with nonzero compliance. Estimation models included: the standard HSIS covariates; fixed intercepts for Head Start centers; the appropriate nonresidualized pretest; and a binary indicator for age cohort. Effect sizes were calculated for each outcome by dividing the Head Start effect estimate in its original units by the control group standard deviation for the outcome. P-values are in parentheses below each parameter estimate.

Estimation follows Option B in Raudenbush, Reardon, and Nomi (2012).

* = p<0.10; ** = p<0.05; *** = p<0.01

[1]No p-value for the cross-site standard deviation is currently available for this approach.

Appendix D

Subgroup Estimates of the Effect of Head Start Assignment on Head Start Enrollment

Appendix Table D.1

Grand mean ITT effect on Head Start enrollment, by subgroup

	Cognitive outcomes				Socio-emotional outcomes	
	Receptive vocabulary (PPVT)	Early reading (WJ-LW)	Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)	Externalizing	Self-regulation
Grand mean	70.0***	70.1***	69.5***	70.2***	72.9***	70.8***
Pretest performance						
Low pretest performers	79.9***	79.3***	78.1***	79.6***	82.4***	80.8***
Other children	74.9***	75.2***	74.8***	75.1***	76.7***	75.2***
DLL status						
Dual language learner	77.2***	77.0***	75.4***	77.0***	78.9***	78.3***
English only	69.6***	69.8***	69.3***	69.6***	71.4***	70.2***
Home language						
Spanish	75.1***	75.0***	73.2***	75.0***	77.0***	76.1***
Other	70.5***	70.5***	70.4***	70.7***	73.0***	71.1***
Special needs						
Yes	71.3***	72.9***	70.9***	75.9***	80.0***	71.2***
No	72.2***	72.3***	71.5***	72.1***	74.4***	73.1***
Age cohort						
Age 3	73.1***	73.3***	72.6***	73.4***	75.3***	73.7***
Age 4	73.7***	73.7***	73.4***	73.9***	75.6***	73.9***

(continued)

Appendix Table D.1 (continued)

Gender						
Male	70.6***	70.5***	69.8***	70.4***	73.9***	71.3***
Female	75.9***	75.9***	75.4***	76.0***	77.8***	76.0***
Race/ethnicity						
Black	66.9***	67.0***	66.7***	66.1***	69.7***	68.3***
Hispanic	72.2***	72.2***	70.4***	72.0***	74.9***	72.6***
White/other	76.7***	76.5***	76.8***	76.5***	76.8***	76.5***

Note: Within each relevant subgroup, models were fit: using children with available outcome data in nonzero compliance, complete randomized blocks; including the standard HSIS covariates; using fixed intercepts for centers; using the appropriate nonresidualized pretest; using data from both cohorts; and including a control for age cohort. Statistically significant impact differences between subgroups for a given outcome are shaded in gray ($p<0.10$). Statistical significance of differences in subgroup impacts was determined via a t-test of the interaction between the subgroup characteristic and the treatment variable. For race/ethnicity, statistical significance of differences between subgroups was determined via an omnibus test. For children with nonmissing outcome data, missing data were imputed once, except for the relevant subgroup characteristic, which was not imputed.

* = p<0.10; ** = p<0.05; *** = p<0.01

Appendix E

Baseline Balance Tests for Key Subgroups

Appendix Table E.1

Baseline balance of the analysis sample: DLL low pretest sample

Baseline characteristic	Mean value of the baseline characteristic			
	Treatment group	Control group	Difference	P-value of difference
Pretest results				
Receptive vocabulary (PPVT)	192.9	195.5	-2.6	0.512
Early reading (WJ-LW)	285.6	285.5	0.1	0.981
Oral comprehension (WJ-OC)[1]	N/A	N/A	N/A	N/A
Early numeracy (WJ-AP)[1]	N/A	N/A	N/A	N/A
Externalizing (parent reports)	1.9	1.8	0.0	0.652
Self-regulation (assessor reports)	3.1	3.2	0.0	0.893
Child characteristics				
Male (%)	51.1	44.1	7.0	0.383
Black (%)	1.1	0.0	1.1	0.381
Hispanic (%)	98.9	99.5	-0.6	0.646
English is home language (%)	1.1	0.0	1.0	0.407
Family characteristics				
Mother's age (years)	30.0	28.4	1.6	0.111
Mother has less than HS education (%)	73.9	83.3	-9.4	0.158
Mother has HS education (%)	12.0	11.8	0.2	0.970
Mother is married (%)	0.6	0.7	-0.1	0.479
Mother was previously married (%)	8.8	11.8	-3.0	0.533
Mother is a teenager (%)	10.9	17.6	-6.8	0.221
Mother is a recent immigrant (%)	54.3	65.7	-11.4	0.149
Child lives with both biological parents (%)	75.0	74.8	0.2	0.978
Assessment characteristics				
Child age at spring testing	4.0	4.1	-0.1	0.496
Spring child assessment date[2]	32.4	33.1	-0.7*	0.087
Child was tested in English (%)	0.0	0.0	0.0	--
Spring parent interview date[2]	33.2	32.6	0.6	0.189

Notes: Sample includes children in complete randomized blocks with nonzero compliance and with nonmissing WJ-LW outcome data. Dual language learners who are low pretest performers have a PPVT pretest score that falls within the lower third of PPVT pretest scores for dual language control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members. DLL = dual language learner.

[1]Pretest data were not collected for this outcome.

[2]In weeks since September 1, 2002.

Appendix Table E.2

Baseline balance of the analysis sample: DLL other sample members

Baseline characteristic	Mean value of the baseline characteristic: Treatment group	Control group	Difference	P-value of difference
Pretest results				
Receptive vocabulary (PPVT)	242.6	243.1	-0.6	0.858
Early reading (WJ-LW)	297.3	299.5	-2.3	0.425
Oral comprehension (WJ-OC)[1]	N/A	N/A	N/A	N/A
Early numeracy (WJ-AP)[1]	N/A	N/A	N/A	N/A
Externalizing (parent reports)	1.8	1.8	0.0	0.987
Self-regulation (assessor reports)	3.4	3.4	0.1	0.475
Child characteristics				
Male (%)	49.0	45.7	3.4	0.556
Black (%)	0.0	0.0	0.0	--
Hispanic (%)	98.6	97.8	0.7	0.629
English is home language (%)	3.8	1.6	2.2	0.259
Family characteristics				
Mother's age (years)	30.2	30.1	0.0	0.970
Mother has less than HS education (%)	61.2	66.7	-5.5	0.319
Mother has HS education (%)	23.8	20.5	3.3	0.492
Mother is married (%)	0.7	0.7	0.0	0.623
Mother was previously married (%)	12.3	9.1	3.1	0.388
Mother is a teenager (%)	9.3	5.8	3.4	0.272
Mother is a recent immigrant (%)	52.9	59.3	-6.4	0.261
Child lives with both biological parents (%)	74.8	74.3	0.5	0.920
Assessment characteristics				
Child age at spring testing	4.3	4.3	0.0	0.660
Spring child assessment date[2]	32.4	33.7	-1.3***	0.001
Child was tested in English (%)	0.0	0.0	0.0	--
Spring parent interview date[2]	33.6	33.8	-0.3	0.504

Notes: Sample includes children in complete randomized blocks with nonzero compliance and with nonmissing WJ-LW outcome data. Dual language learners who are not low pretest performers have a PPVT pretest score that falls within the upper two-thirds of PPVT pretest scores for dual language control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members.

[1]Pretest data were not collected for this outcome.

[2]In weeks since September 1, 2002.

Appendix Table E.3

Baseline balance of the analysis sample: English-only low pretest sample

Baseline characteristic	Mean value of the baseline characteristic: Treatment group	Control group	Difference		P-value of difference
Pretest results					
Receptive vocabulary (PPVT)	221.1	218.6	2.5		0.316
Early reading (WJ-LW)	292.4	291.1	1.3		0.533
Oral comprehension (WJ-OC)[1]	N/A	N/A	N/A		N/A
Early numeracy (WJ-AP)	366.1	360.7	5.4	**	0.033
Externalizing (parent reports)	1.7	1.8	-0.1	*	0.052
Self-regulation (assessor reports)	2.8	2.8	0.0		0.992
Child characteristics					
Male (%)	51.4	48.8	2.7		0.537
Black (%)	60.4	57.1	3.3		0.436
Hispanic (%)	11.8	16.6	-4.7		0.111
English is home language (%)	94.8	97.0	-2.2		0.203
Family characteristics					
Mother's age (years)	28.3	28.3	-0.1		0.922
Mother has less than HS education (%)	27.9	40.4	-12.5	***	0.002
Mother has HS education (%)	44.2	36.3	8.0	*	0.064
Mother is married (%)	0.3	0.3	0.0		0.898
Mother was previously married (%)	14.0	19.3	-5.3	*	0.097
Mother is a teenager (%)	21.1	23.8	-2.8		0.442
Mother is a recent immigrant (%)	2.9	2.6	0.3		0.831
Child lives with both biological parents (%)	35.6	33.1	2.5		0.552
Assessment characteristics					
Child age at spring testing	3.8	3.8	0.0		0.695
Spring child assessment date[2]	32.2	33.0	-0.8	***	0.002
Child was tested in English (%)	99.7	99.5	0.2		0.717
Spring parent interview date[2]	33.4	33.1	0.3		0.361

Notes: Sample includes children in complete randomized blocks with nonzero compliance and with nonmissing WJ-LW outcome data. English-only sample members who are low pretest performers have a PPVT pretest score that falls within the lower third of PPVT pretest scores for English-only control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members.

[1]Pretest data were not collected for this outcome.

[2]In weeks since September 1, 2002.

Appendix Table E.4

Baseline balance of the analysis sample: English-only other sample members

Baseline characteristic	Mean value of the baseline characteristic				
	Treatment group	Control group	Difference		P-value of difference
Pretest results					
Receptive vocabulary (PPVT)	276.3	276.8	-0.5		0.800
Early reading (WJ-LW)	309.0	305.4	3.6	***	0.012
Oral comprehension (WJ-OC)[1]	N/A	N/A	N/A		N/A
Early numeracy (WJ-AP)	384.1	382.6	1.5		0.346
Externalizing (parent reports)	1.6	1.7	0.0	*	0.062
Self-regulation (assessor reports)	3.2	3.2	0.0		0.252
Child characteristics					
Male (%)	47.0	49.0	-2.0		0.459
Black (%)	37.2	32.2	4.9	*	0.061
Hispanic (%)	13.9	15.9	-2.0		0.293
English is home language (%)	97.3	96.4	1.0		0.305
Family characteristics					
Mother's age (years)	28.8	28.4	0.4		0.384
Mother has less than HS education (%)	23.8	24.2	-0.4		0.869
Mother has HS education (%)	36.6	34.8	1.8		0.505
Mother is married (%)	0.4	0.4	0.0		0.122
Mother was previously married (%)	19.8	17.4	2.4		0.269
Mother is a teenager (%)	18.4	19.5	-1.1		0.605
Mother is a recent immigrant (%)	3.4	2.2	1.2		0.209
Child lives with both biological parents (%)	28.8	28.4	0.4		0.384
Assessment characteristics					
Child age at spring testing	41.5	43.1	-1.6		0.564
Spring child assessment date[2]	4.1	4.1	0.1		0.140
Child was tested in English (%)	32.2	33.1	-0.8	***	0.000
Spring parent interview date[2]	99.8	99.7	0.1		0.792

Notes: Sample includes children in complete randomized blocks with nonzero compliance and with nonmissing WJ-LW outcome data. English-only sample members who are not low pretest performers have a PPVT pretest score that falls within the upper two-thirds of PPVT pretest scores for English-only control group members, with separate pretest cutoffs for 3- and 4-year-old cohort members.

[1]Pretest data were not collected for this outcome.

[2]In weeks since September 1, 2002.

Appendix F

Cross-Site Grand Means and Standard Deviations for Head Start Effect Sizes Estimated With and Without a Pretest Covariate

Appendix Table F.1 below compares estimates of the cross-site grand mean and standard deviation of Head Start effect sizes estimated with a pretest covariate and their counterparts estimated without a pretest covariate. There are no substantial or systematic differences between the two sets of estimates.

Appendix Table F.1

Cross-site grand means and standard deviations for Head Start ITT effect sizes estimated with and without a pretest covariate

	Cognitive outcomes				Socio-emotional outcomes	
Parameter	Receptive vocabulary (PPVT)	Early reading (WJ-LW)	Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)	Externalizing (parent reports)	Self-regulation (assessor reports)
With pretest						
Grand mean	0.14*** (<0.001)	0.17*** (<0.001)	0.01 (0.625)	0.12*** (<0.001)	-0.05* (0.088)	0.02 (0.568)
Standard deviation	0.12** (0.030)	0.25*** (<0.001)	0.12* (0.097)	0.07 (0.230)	0.16*** (0.010)	0.22** (0.014)
Without pretest						
Grand mean	0.12*** (<0.001)	0.19*** (<0.001)	0.01 (0.723)	0.10*** (0.001)	-0.09*** (0.008)	0.01 (0.723)
Standard deviation	0.10* (0.080)	0.27*** (0.002)	0.14* (0.055)	0.04 (0.250)	0.12 (0.181)	0.21** (0.048)

Notes: Models were fit: using children with available outcome data in nonzero compliance and complete randomized blocks; including the standard HSIS covariates; using fixed intercepts for centers; using the appropriate nonresidualized pretest; using data from both cohorts; and including a binary indicator for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect on each outcome in its original units by the control group standard deviation for that outcome. P-values are in parentheses below each parameter estimate.

* = p<0.10, ** = p<0.05, *** = p<0.01

Appendix G

Using a Random-Effects Meta-Analysis to Estimate Variation in Program Effect Sizes Across Past Studies

The meta-analytic database that we used to estimate variation in program effect sizes across past studies was obtained from the National Forum on Early Childhood Policy and Programs Meta-Analysis Project. This database synthesizes over four decades of evaluations of programs for children from their prenatal period to age 5 (1962-2009).

Shager and colleagues (2013, p. 80) state that be included in this database, "studies must have had (a) a comparison group (either an observed control or alternative treatment group) and (b) at least 10 participants in each condition, with attrition of less than 50% in each condition. Evaluations could have been experimental or quasi-experimental, using one of the following methods: regression discontinuity, fixed effects (individual or family), residualized or other longitudinal change models, difference in difference, instrumental variables, propensity score matching, or interrupted time series. Quasi-experimental evaluations not using one of the former analytic strategies were also included if they had a comparison group *plus* pre- and posttest information on the outcome of interest or demonstrated adequate comparability of groups on baseline characteristics." For further details see Duncan and Magnuson (2013), Shager et al. (2013), and Schindler et al. (2013).

We obtained the database from the online appendix for Duncan and Magnuson (2013). Each row in the database represents a treatment-control group *contrast* from a given study, where a contrast is defined as a comparison between one group of children who received a given intervention and another group of children who received no other services that were a result of participating in the study (Shager et al., 2013). Treatment-control group contrasts are nested within studies. Because the Head Start Impact Study includes only children who are between 3 and 5 years old, we dropped 11 contrasts from 10 studies of intervention services that began before age 3 (out of 85 contrasts in 65 studies total). Doing so left us with 74 contrasts from 55 studies. Head Start studies are identified in the database by a 0/1 indicator. The original meta-analytic team also provided us with findings for an additional study (Boston prekindergarten; Weiland & Yoshikawa, 2013) that was not included in the online appendix for Duncan and Magnuson (2013).

Effect sizes in the database represent the average *effect of treatment on the treated* across all cognitive outcomes assessed for given contrast. Effect sizes were coded in two steps. First, for each cognitive outcome, the original meta-analysis team calculated effect sizes using Hedges's g, which adjusts standardized mean differences (Cohen's d) to account for small-sample bias. The effect sizes chosen for this purpose were those that were measured as close to the end of treatment as possible for each original study. This resulted in follow-up intervals ending as late as one year after program completion and as early as three-quarters of the way through a program (Duncan and Magnuson, 2013). For the online appendix, the original meta-analysis team calculated a contrast-level effect size (which is what we analyze) by taking a

simple mean of all cognitive effect size estimates for each contrast (Duncan and Magnuson, 2013).

The database also reports the "inverse of squared standard errors of the average estimates, which is calculated by a Bayesian shrinkage model to take sampling variation of the within-study estimates into account" (online appendix, Duncan and Magnuson, 2013, p. 4). Weights were truncated from above at 100 in order "to avoid sensitivity to extremely large variance weights" (online appendix, Duncan and Magnuson, 2013, p. 4).

To estimate cross-study variation in program effect sizes using effect-size information from the database, we performed a random-effects meta-analysis using V-known estimation in SAS (Hedges and Olkin, 1985). We tested the statistical significance of our estimates of cross-study variation using a conventional Q statistic, which is analogous to step 2 in the LATE estimation method for the present study.

Appendix H

A Constrained Empirical Bayes Method for Estimating Site-Specific Mean ITT Program Effects to Reflect the Estimated Cross-Site Variance of True Program Effects

Empirical Bayes estimators (often referred to as "shrinkage estimators") have the smallest mean squared error for predicting a specific parameter value, such as the mean ITT program effect for a site (Lindley and Smith, 1972). However, they are biased toward the grand mean and thereby "overshrink" their OLS counterparts toward the grand mean (Raudenbush and Bryk, 2002). Consequently, empirical Bayes estimates *understate* the cross-site variance of true mean program effects and in this regard they do not properly represent the cross-site distribution of effects.[1]

The present appendix derives an adjustment that corrects for this fact. To begin, note that by definition, the sample variance of empirical Bayes estimates ($\hat{B}_j^{EB}$) around their estimated grand mean ($\hat{\beta}$) for J sites is:

$$\hat{Var}(\hat{B}_j^{EB}) \equiv \frac{\sum_{j=1}^{J}(\hat{B}_j^{EB}-\hat{\beta})^2}{J} \tag{H.1}$$

Then recall that by estimating Equations 1-3 in the present paper one can obtain an unbiased estimate of the cross-site variance of true mean ITT program effects $\hat{\tau}_{ITT}^2$. The problem is that

$$\hat{Var}(\hat{B}_j^{EB}) < \hat{\tau}_{ITT}^2 \tag{H.2}$$

or stated another way:

$$\hat{Var}(\hat{B}_j^{EB}) = \gamma \cdot \hat{\tau}_{ITT}^2 \tag{H.3}$$

where

$$\gamma \equiv \frac{\hat{Var}(\hat{B}_j^{EB})}{\hat{\tau}_{ITT}^2} \tag{H.4}$$

and

$$0 < \gamma < 1.$$

This implies that:

$$\hat{\tau}_{ITT}^2 = \frac{1}{\gamma}\hat{Var}(\hat{B}_j^{EB}). \tag{H.5}$$

Define a constrained empirical Bayes estimator ($\hat{B}_j^{CEB}$) with a sample variance:

[1] Raudenbush and Bryk (2002, p. 88) discuss these issues for a two-level hierarchical model.

$$\hat{Var}(\hat{B}_j^{CEB}) \equiv \frac{\sum_{j=1}^{J}(\hat{B}_j^{CEB} - \hat{\beta})^2}{J}. \quad \text{(H.6)}$$

Then specify that this sample variance should equal the model-based estimate of the true variance:

$$\hat{Var}(\hat{B}_j^{CEB}) = \hat{\tau}_{ITT}^2. \quad \text{(H.7)}$$

Substituting Equations H.5 and H.1 into Equation H.7 yields;

$$\begin{aligned} \hat{Var}(\hat{B}_j^{CEB}) &= \frac{1}{\gamma}\hat{Var}(\hat{B}_j^{EB}) \\ &= \frac{1}{J} \cdot \frac{1}{\gamma}\sum_j(\hat{B}_j^{EB} - \hat{\beta})^2 \\ &= \frac{1}{J} \cdot \sum_j(\frac{1}{\sqrt{\gamma}}(\hat{B}_j^{EB} - \hat{\beta}))^2. \end{aligned} \quad \text{(H.8)}$$

Equation H.8 indicates that multiplying the deviation of each empirical Bayes estimate from its grand mean by $\frac{1}{\sqrt{\gamma}}$ (which "stretches" these deviations) produces a sample variance that equals the estimated variance of true program effects. The resulting constrained empirical Bayes estimator for a given site *j* is

$$\hat{B}_j^{CEB} = \hat{\beta} + \frac{1}{\sqrt{\gamma}}(\hat{B}_j^{EB} - \hat{\beta}). \quad \text{(H.9)}$$

This approach is asymptotically equivalent to that suggested by Louis (1984).

References

Abbott-Shim, M., Lambert, R., and McCarty, F. (2003). A comparison of school readiness outcomes for children randomly assigned to a Head Start program and the program's wait list. *Journal of Education for Students Placed at Risk*, *8*, 191-214.

Achenbach, T. M., Edelbrock, C., and Howell, C. T. (1987). Empirically based assessment of the behavioral/emotional problems of 2- and 3-year-old children. *Journal of Abnormal Child Psychology*, *15*, 629-650.

Administration for Children and Families. (2014). Head Start center location datasets. Retrieved September 24, 2014, from http://eclkc.ohs.acf.hhs.gov/hslc/data/center-data.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455.

Arbour, M. C., Yoshikawa, H., Murnane, R., Weiland, C., Barata, M. C., and Snow, C. E. (2014). Testing for moderation of impact of the UBC preschool intervention by student absenteeism. Working paper.

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children*, 25-50.

Barnett, W. S. (2007). Revving up Head Start: Lessons from recent research. *Journal of Policy Analysis and Management*, *26*, 674-677.

Barnett, W. S., Yaroz, D. J., Thomas, J., Jung, K., and Blanco, D. (2007). Two-way immersion in preschool education: An experimental comparison. *Early Childhood Research Quarterly*, *22*, 277-293.

Bialystok, E. (2011). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology*, *65*, 229.

Bialystok, E., Luk, G., Peets, K. F., and Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *13*, 525-531.

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., Nelson, K. E., and Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI Program. *Child Development*, *79*, 1802-1817.

Bitler, M., Hoynes, H., and Domina, T. (2014). Experimental evidence on distributional effects of Head Start. Working paper.

Bloom, H., Raudenbush, S., Weiss, M., and Porter, K. (under review). Using multi-site evaluations to study variation in program effects. Working paper.

Bloom, H., and Weiland, C. (2014). *Replicating findings from the Head Start Impact Study*. New York: MDRC.

Brachet, T. (2007). Documentation for computing clustered standard errors for two-stage least squares in SAS. Retrieved January 5, 2015, from http://works.bepress.com/tbrachet/2/.

Burchinal, M., Kainz, K., and Kai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, and T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11-31). Baltimore, MD: Brookes.

Burke, L., and Sheffield, H. (2013). Universal preschool's empty promises. Retrieved September 24, 2014, from http://www.heritage.org/research/reports/2013/03/universal-preschools-empty-promises.

Campbell, S. B. (1995). Behavior problems in preschool children: A review of recent research. *Journal of Child Psychology and Psychiatry, 36*, 113-149.

Card, N. A, Stucky, B. D., Sawalani, G. M., and Little, T. D. (2008). Direct and indirect aggression in children and adolescents: A meta-analytic review of gender differences, inter-correlations, and relations to maladjustment. *Child Development, 79*, 1185-1229.

Carlson, S., and Meltzoff, A. (2008). Bilingual experience and executive functioning in young children. *Developmental Science, 11*, 282-298.

Cicirelli, V. G. (1969). *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development*. Athens, OH: Westinghouse Learning Corporation.

Clements, D. H., and Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education, 38*, 136-163.

Clements, D. H., Sarama, J. H., and Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology, 28*, 457-482.

Clements, D. H., Sarama, J., Wolfe, C. B., and Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies persistence of effects in the third year. *American Educational Research Journal, 50*, 812-850.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Currie, J. (2001). Early childhood education programs. *Journal of Economic Perspectives, 15*, 213-238.

Currie, J. (2007). How should we interpret the evidence about Head Start? *Journal of Policy Analysis and Management, 26*, 681-684.

Currie, J., and Thomas, D. (1995). Does Head Start make a difference? *American Economic Review, 85*, 341-364.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*, 111-34.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., and Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428-1446.

Duncan, G. J., and Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives, 27*(2), 109-132.

Dunn, L. M., and Dunn, L. M. (1997). Peabody Picture Vocabulary Test-Third Edition. Bloomington, MN: Pearson Assessments.

Early, D. M., Iruka, I. U., Ritchie, S., Barbarin, O. A., Winn, D. M. C., Crawford, G. M., Frome, P. M., Clifford, R. M., Burchinal, M., Howes, C., Bryant, D., and Pianta, R. C. (2010). How do pre-kindergarteners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 25*, 177-193.

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., and Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin, 132*, 33-72.

Engel, M., Claessens, A., and Finch, M. A. (2013). Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35,* 157-178.

Epstein, J. L., and Sheldon, S. B. (2002). Present and accounted for: Improving student attendance through family and community involvement. *Journal of Educational Research, 95*, 308-318.

Feller, A., Grindal, T., Miratrix, L., and Page, L. (2014). Compared to what? Variations in the impacts of early childhood education by alternative care-type settings. Working paper.

Friedman-Krauss, A. H., Connors, M. C., and Morris, P. A. (2014). Unpacking the treatment contrast in the Head Start Impact Study: How did assignment to treatment impact quality of care? Working paper.

Friedman-Krauss, A. H., Connors, M. C., Morris, P. A., Feller, A., and McCoy, D.C. (2014). Program level variation in Head Start impacts: The moderating role of classroom quality. Presentation at the annual meeting of the Association for Public Policy Analysis and Management, Albuquerque, N.M.

Fuller, B., and Kim, A.Y. (2011). Latino access to preschool stalls after earlier gains. Retrieved from the Institute of Human Development at the University of California-Berkeley website: http://ihd.berkeley.edu/Latino%20preschool%20decline%20-%20NOLA-NJLC-Brief-2011-FINAL.pdf.

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

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., and Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development, 4*, 1343-1359.

Gibbs, C. (2014). Experimental evidence on early intervention: The impact of full-day kindergarten. Working paper.

Ginsburg, H. P., Lee, J. S., and Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report*, *22*(1), Society for Research in Child Development

Gormley Jr., W. T., Gayer, T., Phillips, D., and Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*, 872-884.

Gormley Jr., W. T., Phillips, D., Adelstein, S., and Shaw, C. (2010). Head Start's comparative advantage: Myth or reality? *Policy Studies Journal, 38*, 397-418.

Gormley Jr., W. T., Phillips, D. A., and Gayer, T. (2008). Preschool programs can boost school readiness. *Science, 320*, 1723-1724.

Gormley Jr., W. T., Phillips, D. A., Newmark, K., Welti, K., and Adelstein, S. (2011). Social-emotional effects of early childhood education programs in Tulsa. *Child Development, 82*, 2095-2109.

Grindal, T., Bowne, J. B., Yoshikawa, H., Schindler, H., Duncan, G. J., Magnuson, K., and Shonkoff, J. (2013). The added impact of parenting education in early childhood education programs: A meta-analysis. Paper presented at the 2013 Partners Summit of the Alliance for Early Success and the Ounce of Prevention Fund, Boston.

Head Start Improvement Act of 2014, S. 2119 (2014).

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics, 54*, 3-56.

Hedges, L.V., and Olkin, I. (1985) Statistical methods for meta-analysis. San Diego, CA: Academic Press.

Hill, C. J., Bloom, H. S., Black, A. R., and Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111*, 127-155.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., and Levine, S. (2002). Language input and child syntax. *Cognitive Psychology, 45*, 337-374.

Improving Head Start for School Readiness Act of 2007, H.R. 1429, 110th Cong.

Jenkins, J. M., Farkas, G., Duncan, G. J., Burchinal, M., and Vandell, D. L. (2014). Head Start at ages 3 and 4 versus Head Start followed by state pre-k: Which is more effective? Working paper.

Johnson, R. (2013). School quality and the long-run effects of Head Start. Goldman School of Public Policy working paper.

Justice, L. M., McGinty, A. S., Zucker, T., Cabell, S. Q., and Piasta, S. B. (2013). Bi-directional dynamics underlie the complexity of talk in teacher-child play-based conversations in classrooms serving at-risk pupils. *Early Childhood Research Quarterly, 28*, 496-508.

Kelchen, R., Magnuson, K. A., Duncan, G. J., Schindler, H. S., Shager, H., and Yoshikawa, H. (2012). Do the effects of early childhood education programs differ by gender? A meta-analysis. Manuscript under review.

Kraemer, S. (2000). The fragile male. *British Medical Journal, 321*, 1609-1612.

Leak, J., Duncan, G., Li, W., Magnuson, K., Schindler, H., and Yoshikawa, H. (2012). Is timing everything? How early childhood education program cognitive and achievement impacts vary by starting age, program duration and time since the end of the program. Manuscript submitted for publication.

Lee, V. E., Brooks-Gunn, J., Schnur, E., and Liaw, F. R. (1990). Are Head Start Effects sustained? A longitudinal follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Child Development, 61*, 495-507.

Lee, V. E., Burkam, D. T., Ready, D. D., Honigman, J., and Meisels, S. J. (2006). Full-day versus half-day kindergarten: In which program do children learn more? *American Journal of Education, 112*, 163-208.

Lindley, D. V., and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-41.

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., and Bilbrey, C. (2013). Evaluation of the Tennessee Voluntary Prekindergarten Program: End of pre-k results from the randomized control design. Research report, Vanderbilt University, Peabody Research Institute, Nashville.

Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., and Hofer, K. G. (2014). The prekindergarten age-cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis*, doi: 10.3102/0162373714547266.

Liu, J. (2004). Childhood externalizing behavior: theory and implications. *Journal of Child and Adolescent Psychiatric Nursing, 17*(3), 93-103.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., and Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review, 26*, 52-66.

Loeb, S., Fuller, B., Kagan, S. L., and Carrol, B. (2004). Child care in poor communities: Early learning effects of type, quality, and stability. *Child Development, 75*, 47-65.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and Empirical Bayes methods. *Journal of the American Statistical Association, 79*, 393-398.

Ludwig, J., and Miller, D. L. (2007). Does Head Start improve children's life chances: Evidence from a regression discontinuity design. *Quarterly Journal of Economics, 122*, 159-208.

Ludwig, J., and Phillips, D. (2007). The benefits and costs of Head Start. Society for Research on Child Development, Social Policy Report. Volume XXI, Number 3.

Ludwig, J., and Phillips, D. A. (2008). Long-term effects of Head Start on low-income children. *Annals of the New York Academy of Sciences, 1136*, 257-268.

Mancilla-Martinez, J., and Lesaux, N. K. (2011). Early home language use and later vocabulary development. *Journal of Educational Psychology, 103*, 535-546. doi: 10.1037/a0023655

Magnuson, K., Lahaie, C., and Waldfogel, J. (2006). Preschool and school readiness of children of immigrants. *Social Science Quarterly, 87*, 1241-1262.

Magnuson, K. A., Ruhm, C., and Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26*, 33-51.

Magnuson, K. A., and Waldfogel, J. (2005). Early childhood care and education: Effects on ethnic and racial gaps in school readiness. *The Future of Children, 15*, 169-196.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., and Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732-749.

Matthews, J. S., Ponitz, C. C., and Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology, 101*, 689-704.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674-701.

Moiduddin, E., Aikens, N., Tarullo, L., West, J., and Xue, Y. (2012). *Child outcomes and classroom quality in FACES 2009* (No. 7622). Princeton, NJ: Mathematica Policy Research.

Morris, P., Millenky, M., Raver, C. C., and Jones, S. M. (2013). Does a preschool social and emotional wellbeing intervention pay off for classroom instruction and children's behavior and academic skills? Evidence from the Foundations of Learning project. *Early Education and Development, 24*, 1020-1042.

NICHD Early Child Care Research Network. (2002). Early child care and children's development prior to school entry: Results from the NICHD Study of Early Child Care. *American Educational Research Journal, 39*, 133-164.

NICHD Early Child Care Research Network. (2003). Social functioning in first grade: Prediction from home, child care and concurrent school experience. *Child Development, 74*, 1639-1662.

Office of Head Start. (2013). A national overview of grantee CLASS™ scores in 2013. Retrieved September 24, 2014, from https://eclkc.ohs.acf.hhs.gov/hslc/data/class-reports/class-data-2013.html.

Office of Head Start. (2014a). Head Start program facts: Fiscal year 2013. Retrieved September 21, 2014, from https://eclkc.ohs.acf.hhs.gov/hslc/data/factsheets/2013-hs-program-factsheet.html.

Office of Head Start. (2014b). History of Head Start. Retrieved December 20, 2014, from http://www.acf.hhs.gov/programs/ohs/about/history-of-head-start.

Peck, L. R., and Bell, S. H. (2014). The role of program quality in determining Head Start's impact on child development. OPRE Report #2014-10, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., and Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*, 1534-1553.

Phillips, D. A., and Meloy, M. E. (2012). High-quality school-based pre-k can boost early learning for children with special needs. *Exceptional Children, 78*, 471-490.

Puma, M., Bell, S. H., Cook, R., and Heid, C. (2010a). Head Start Impact Study: Final report. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Puma, M., Bell, S. H., Cook, R., and Heid, C. (2010b). Head Start Impact Study: Technical report. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Puma, M., Bell, S. H., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A., and Downer, J. (2012). Third grade follow-up to the Head Start Impact Study final report, OPRE Report #2012-45, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Reardon, S. F., and Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness, 5*, 303-332.

Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Champion, K. M., and Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly, 23*, 10-26.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., and Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development, 82*, 362-378.

Raver, C. C., Jones, S. M., Li-Grining, C. P., Zhai, F., Metzger, M., and Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology, 77*, 302-316.

Resnick, G., and Zill, N. (1999). Is Head Start providing high-quality educational services? "Unpacking" classroom practices. Paper presented at the biannual convention of the Society for Research in Child Development, Albuquerque, NM.

Robin, K. B., Frede, E. C., and Barnett, W. S. (2006). Is more better? The effects of full-day vs half-day preschool on early school achievement. National Institute for Early Education working paper.

Roid, G. H., and Miller, L. J. (1997). Leiter international performance scale — revised. Wood Dale, IL: Stoelting.

Sameroff, A. J., and Chandler, M. J. (1975). Reproductive risk and the continuum of caretaker casualty. In F. D. Horowitz (Ed.), *Review of child development research* (Vol. 4, pp. 187-244). Chicago: University of Chicago Press.

Schindler, H. S., Kholoptseva, J., Oh, S. S., Yoshikawa, H., Duncan, G. J., Magnuson, K., and Shonkoff, J. P. (2013). Maximizing the potential of early childhood education to prevent externalizing behavior problems: A meta-analysis. Working paper.

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

Starkey, P., and Klein, A. (2000). Fostering parental support for children's mathematical development: An intervention with Head Start families. *Early Education and Development, 11*, 659-680.

Starkey, P., Klein, A., and Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly, 19*, 99-120.

Strong Start for America's Children Act of 2013, H.R. 3461 (2013).

U.S. Department of Health and Human Services. (2004). Head Start Impact Study spring 2004 cohort B parent interview. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., and Vandergrift, N. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD study of early child care and youth development. *Child Development, 81*, 737-756.

Votruba-Drzal, E., Coley, R. L., and Chase-Lansdale, P. L. (2004). Child care and low-income children's development: Direct and moderated effects. *Child Development, 75*, 296-312.

Vygotsky, L. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

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

Weiland, C., Ulvestad, K., Sachs, J., and Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly, 28*, 199-209.

Weiland, C., and Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*, 2112-2130.

White House. (2013). Early learning. Retrieved September 24, 2014, from http://www.whitehouse.gov/issues/education/early-childhood.

Wong, V. C., Cook, T. D., Barnett, W. S., and Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27*, 122-154.

Woodcock, R. W., McGrew, K. S., and Mather, N. (2001). Woodcock-Johnson tests of achievement. Itasca, IL: Riverside Publishing.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Magnuson, K. A., Phillips, D., and Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. New York: Foundation for Child Development, Society for Research in Child Development.

Zaslow, M. (2008). Issues for the learning community. *Infants and Young Children, 21*, 4-17.

Zaslow, M. S., and Hayes, C. D. (1986). Sex differences in children's responses to psychosocial stress: Toward a cross-context analysis. In M. Lamb and B. Rogoff (Eds.), *Advances in developmental psychology* (Vol. 4, pp. 289-337). Hillsdale, NJ: Erlbaum.

Zhai, F., Brooks-Gunn, J., and Waldfogel, J. (2011). Head Start and urban children's school readiness: A birth cohort study in 18 cities. *Developmental Psychology, 47*, 134.

Zhai, F., Brooks-Gunn, J., and Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology, 50*(12), 2572-2586.

Zigler, E., and Styfco, S. J. (2010). *The hidden history of Head Start*. New York: Oxford University Press.