Abstract Title Page

Title: To What Extent Do Head Start's Effects on Children's Language, Literacy, Mathematics, and Socio-Emotional Skills Vary Across Individuals, Subgroups, and Centers?

Authors and Affiliations:

Howard S. Bloom
MDRC

Christina Weiland
University of Michigan
School of Education

Abstract Body

Background / Context:
Head Start is the largest publicly funded preschool program in the U.S. and one of its primary goals is to improve the school readiness of low-income children. As has been widely reported, the first randomized trial of Head Start in the program's history found some evidence that it is achieving this goal. Receiving one year of Head Start had small impacts on children's cognitive outcomes, with impacts on cognitive impacts concentrated in the language and literacy domain. However, effects largely faded out by the end of first grade (U.S. Department of Health and Human Services, 2010). In explaining these results, some have pointed to the variation in quality across Head Start centers. For example, fewer than 1 in 20 4-year-olds in the treatment group were in centers with an "excellent" quality rating and only about half were in centers with recommended pupil/staff ratios (National Forum on Early Childhood Policy and Programs, 2010). Given this variation in center quality, by extension, there may be considerable variation in Head Start's impacts on children. That is, under certain conditions and for certain children, Head Start may be much more effective than it is on average.

Purpose / Objective / Research Question / Focus of Study:
In the current study, we use data for the first follow-up year of the Head Start Head Start Impact Study to examine variation in Head Start's impacts on children. Specifically, we examine whether there is statistically significant variation in Head Start's impacts on children's cognitive and socio-emotional outcomes across individual children, subgroups of children, and Head Start centers. To do so, we use a new and innovative methodology for estimating sources of variation in program impacts (Bloom, 2012; Bloom, Raudenbush, & Weiss, 2013).

Setting:
We use secondary data from the Head Start Impact Study, the first randomized trial of Head Start in the program's history. In the original study, Head Start centers were selected in 2002 to be representative of Head Start centers nationally and were located in 22 states.

Population / Participants / Subjects:
In total, 4,440 children in 202 center groups were randomized to treatment or control in the original Head Start Impact Study. Our sample is comprised of the subset of children within these 202 center groups for whom there was outcome data available during the study's first follow-up year (N=3,785 children, or 85% of the children originally randomized). Children in our sample were diverse in their background characteristics – 30% were Black, 38% were Hispanic, 30% spoke a non-English home language, 50% lived with both biological parents, 19% had a mother who was a recent immigrant, and 50% were male.

Intervention / Program / Practice:
Children randomized to treatment were offered a seat in a classroom in a Head Start program in fall 2002 for the 2002-2003 school year. In total, from our sample of 3,785 children, 86% took up the offered slot and enrolled in Head Start in the treatment year. Approximately 8% of children assigned to treatment enrolled in a non-Head Start center, 5% experienced parent care, 1% enrolled in family daycares, and 1% were cared for by a relative. Children assigned to treatment who took up the offered Head Start treatment were enrolled in programs that emphasized a "whole child" approach, meaning the program targeted children's cognitive,

academic, and socio-emotional skills, as well as the health and nutrition of enrolled children grade (U.S. Department of Health and Human Services, 2010). Head Start programs typically emphasized parent involvement and offered a wide array of comprehensive services for families.

Children randomized to control conditions were free to take up any available early childhood program except for that provided by the Head Start to which they had applied and had not won a seat. In practice, 14% of control group children enrolled in Head Start centers (some in the centers in which they had lost a lottery), 30% enrolled in a non-Head Start center, 39% experienced parent care, 7% enrolled in family daycare, and 12% were cared for by a relative.

Research Design:

Random assignment occurred prior to the beginning of the 2002-03 school year. Data collection began in the fall of 2002, after random assignment and continued through the spring of 2003. Children were randomized within Head Start center group which comprised individual Head Start centers or in some areas an amalgam of several Head Start centers.

Data Collection and Analysis:

In fall 2002 and in spring 2003, study children were tested by a trained child assessor on an extensive battery of cognitive and socio-emotional assessments (U.S. Department of HHS, 2010). Parents were also surveyed in fall 2002 and spring 2003 on basic demographic information and on dimensions of their children's behavior. Children were followed through third grade, and other data were gathered from various sources (e.g. additional child testing, teacher and center director interviews, direct observations of classroom quality). Our analysis is based on data from the fall 2002 and spring 2003 child testing and parent survey only.

Cognitive outcomes used in our analysis include the Peabody Picture Vocabulary Test-III (Dunn & Dunn, 1997), the Woodcock-Johnson Letter-Word Identification subscale, the W-J Oral Comprehension subscale, and the W-J Applied Problems subscale (Woodcock, McGrew, & Mather, 2001). These tests measure children's receptive vocabulary, early reading, oral comprehension, and early math, respectively, and all have strong psychometric properties, such as good internal consistency and inter-rater reliability (Dunn & Dunn, 1997; Woodcock et al., 2001). Children's externalizing behavior problems were assessed using a parent report based on the Child Behavior Checklist (Achenbach, Edelbrock, & Howel, 1987). Seven items regarding children's hyperactive and aggressive behaviors were combined to form the externalizing problems scale (alpha = 0.71). Children's self regulation skills were measured using items from the Leiter-R Assessor report (Roid & Miller, 1997); after testing children, assessors rated their task persistence, attention span, body movement, and attention to direction. Item scores were averaged to create a composite score for analysis (alpha=0.82).

Subgroups of sample members for our analysis are based on indicators of theoretically and empirically important characteristics that were the basis for subgroup analysis in the original Head Start Impact Study (U.S. Department of Health and Human Services, 2010). The subgroups that we examined are follows: lowest quartile of academic achievement vs. non-lowest quartile (as determined by children's Woodcock-Johnson Academic Skills composite scores); home language of English vs. some other language; child has special needs vs. has no special needs; age 4 vs. age 3 at baseline; male vs. female; and race/ethnicity (White/other, Black, Hispanic).

We address our research questions regarding impact variation across individuals, subgroups, and center groups using the following basic multi-level model specification:

Level 1: $Y_{ij} = \alpha_j + \beta_j T_{ij} + \sum_{k=1} \gamma_k X_{kij} + T_{ij} e_{1ij} + (1 - T_{ij}) e_{2ij}$

Level 2: $\alpha_j = \alpha_j$

$\beta_j = \beta_0 + r_j$

where *i* denotes children, *j* denotes center groups, *Y* is the child outcome, *T* indicates treatment-group assignment, *X* denotes the same set of baseline variables used in the original Head Start Impact study(gender, race/ethnicity, home language, mother's education, mother's marital status, language of child testing, when the child was tested or when the parent was surveyed, mother's age, whether the child resided with both biological parents, whether the child's mother was a recent immigrant, and whether the child had a teen mother) as well as a control for age cohort and the child's relevant pretest score, and e_1 and e_2 are residual outcomes for treatment and control group members, respectively. α_j is the fixed intercept for center group *j*, β_j is the mean intent-to-treat (ITT) effect for center group *j*, β_0 is the grand mean ITT effect, and $r_j$is the random component of β_j.

One important feature of our analysis was comparison of residual outcome variances for treatment and control group members. Given the remedial nature of the Head Start program, it was anticipated that it would compress the distribution of individual outcomes somewhat by mainly raising test scores for students who were at the low end of the cognitive achievement distribution. An F test identified whether the estimated residual outcomes for treatment versus control group members - e_1 and e_2 - were statistically significantly different from one another.

To examine impacts by subgroup membership, we first fitted our Level 1 and Level 2 specification within each subgroup (e.g. only males, then only females) and calculated the grand mean treatment effect size for each subgroup. To determine whether impacts were statistically significantly different by subgroup membership, we added main effects for the subgroup and an interaction term between the subgroup indicator and the treatment variable to our Level 1 specification and examined the t-test results on the interaction term.

The most novel contribution of our work is our analysis of variation in Head Start impacts across center groups (sites). To reflect this potential variation, the Level 2 equation of our model specifies that the β_j vary randomly across blocks with a grand mean of β_0 and a variance of $var(\beta_j)$. We will estimate the Level 1 and Level 2 equations using new capabilities in the software program HLM to produce an unbiased estimate of $var(\beta_j)$, which is the parameter of interest that identifies how much variation in impacts there is across center groups. To test whether the estimated cross-site variance of program effects, $\widehat{var}(\beta_j)$, is statistically significantly different from zero, we will use a Chi-Square test of the Q statistic, which is used widely in meta-analysis to test for heterogeneity of effects (e.g. Hedges & Olkin, 1985). It compares the squared deviation of each estimated $\hat{\beta}_j$ from the precision-weighted mean $\bar{\beta}$ of all $\hat{\beta}_j$ relative to the estimated error variance for each $\hat{\beta}_j$. Under the null hypothesis of no variation in effects, its uses a fixed-effect precision weight which reflects the estimated error variance $\hat{V}_j$, for each $\hat{\beta}_j$. In addition, using the new capabilities of the HLM software we will be able to report a non-symmetric confidence interval for our estimates of $\widehat{var}(\beta_j)$.

Importantly, our analysis of the statistical significance of across center variance in impacts on children distinguishes between: (1) variation in program effect *estimates* and (2) variation in program *effects*. Because of estimation error, variation in estimates of program effects is often many times the variation in true program effects (Bloom, 2012; Bloom, et al., 2013). Accordingly, we will be able at SREE to characterize the variation in impacts on children across Head Start centers, accounting for the error variation in *estimates* of program effects.

Findings / Results:
As shown in Table 1, Head Start had small but statistically significant average impacts on children's receptive vocabulary, early reading, early numeracy, and externalizing skills. There was no average impact of Head Start on children's oral comprehension or self-regulation skills.

We also found that the individual residual variance was statistically significant for treatment versus control group children for three outcomes – receptive vocabulary, early numeracy, and oral comprehension, with differences especially pronounced for receptive vocabulary and early numeracy (see Table 2). There were no such differences for children's early reading, externalizing, or self-regulation.

In terms of impacts on student subgroups, we found statistically significantly different and larger impacts for dual language learners for receptive vocabulary and early numeracy (see Table 3). Not surprising given the strong correlation between Hispanic and dual language learners, there were statistically significant results by race/ethnicity for these same two outcomes, with larger impacts among Hispanic students. There were statistically significant impacts on externalizing by students' academic quartile, special needs, and gender. There were no other statistically significantly different impacts by student subgroups across the outcomes.

Analyses regarding whether impacts vary across Head Start center groups are currently underway. New capabilities with the HLM software that will facilitate these analyses will be released in October 2013. At SREE, we will present evidence on whether there is such impact variation. We will also characterize the distribution of impacts across center groups and the confidence intervals around the variation in impacts across center groups.

Conclusions:
Our findings suggest that Head Start had impacts on individual residual variance that have not yet been recognized in analyses of these data. Specifically, Head Start reduced the individual residual variance for treatment group members versus control group members on two important early skills – receptive vocabulary and early math. Notably, these are the two skills for which there were also consistent, larger differences in impacts by subgroup for dual-language learners and Hispanic students. Taken together, these findings suggest that Head Start was successful in improving the vocabulary and early numeracy skills of children with lower levels of such skills at entry, with such effects concentrated among dual-language learners and Hispanic children.

At SREE, we will present additional findings regarding whether the average impacts on the six examined outcomes vary across Head Start centers. These findings will complement the individual residual variance and subgroup analyses reported here. For outcomes in which there is statistically significant variation in impacts across center groups, these findings will suggest that some Head Start centers are substantially more (or less) effective than others in improving children's skills and will allow us to characterize the likely distribution such impacts. Findings of no impact variation across center groups will suggest the opposite; for those outcomes, Head Start centers were no more (or less) effective than the average impact findings suggest.

Our findings will provide a methodological example of applying this new methodology for studying impact variation within a randomized trial with small numbers of participants randomized at each site. Our study will also offer a more in depth understanding of the impacts of Head Start on participating children. For example, we will be able to describe for which (if any) child skills Head Start centers were more versus less effective at improving. We expect these findings will be tied to the Head Start model – e.g. we expect less variance in impacts across centers for outcomes in which the targeting of the intervention was more consistent.

Appendices

Appendix A. References

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

Bloom, H.S. (2102). *Impact variation: How do you know it when you see it?* Paper presented at the Society for Research in Educational Effectiveness Conference, Washington, DC.

Bloom, H.S., Raudenbush, S.W., & Weiss, M. (2013). *Using multi-site evaluations to study variation in program effects*. Manuscript in preparation.

Dunn, L.M., Dunn, L.L., and Dunn, D.M. (1997). *Peabody picture and vocabulary test, third edition (PPVT)*. Circle Pines, MN: American Guidance Service.

National Forum on Early Childhood Policy and Programs. (2010). *Understanding the Head Start Impact Study*. Retrived March 28, 2013 from http://www.developingchild.harvard.edu/

Roid, G. H., & Miller, L. J. (1997). Social emotional rating scale – Examiner version. Leiter International Performance Scale – Revised (Leiter-R). Wood Dale, IL: Stoelting Co.

U.S. Department of Health and Human Services. (2010). *Head Start Impact Study: Final report*. Washington, DC: Administration for Children and Families, Office of Planning, Research and Evaluation.

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

Appendix B. Tables and Figures

Table 1: Main impacts of Head Start on child outcomes at the end of the Head Start year (spring 2003)

	Receptive Vocabulary (PPVT)	Early reading (W-J LW)	Oral Comprehension (W-J OC)	Early numeracy (W-J AP)	Externalizing	Self regulation
Grand mean impact (in effect size)	0.15***	0.15***	0.01	0.12***	-0.05*	0.03
P-value	<0.01	<0.01	0.68	<0.01	0.09	0.39
N children	3668	3675	3612	3649	3681	3654
N center groups	198	198	198	198	198	198

Note: Models were fitted using children with available outcome data, included the standard HSIS covariates, use fixed intercepts for center groups, used the appropriate non-residualized pretest, used data from both cohorts, included a control for age cohort, and used multiple imputation to adjustment for a small amount of missing covariate data. *p<.10; **p<.05; ***p<.01

Table 2: Individual residual variance by treatment or control group assignment status and statistical significance of treatment-control group differences

	Treatment Residual variance	Control residual	Treatment – Control Difference	Treatment Control Group Ratio
Receptive Vocabulary (PPVT)	544***	671***	-127***	0.811***
Early reading (W-J LW)	435***	433***	22.0	1.01
Oral Comprehension (W-J OC)	115***	127***	-12.1**	0.905**
Early numeracy (W-J AP)	465***	563***	-98.3***	0.825***
Externalizing	0.104***	0.106***	-0.002	0.981
Self-regulation	0.389***	0.405***	-0.016	0.960

Note: Models were fitted using children with available outcome data, included the standard HSIS covariates, use fixed intercepts for center groups, used the appropriate non-residualized pretest, used data from both cohorts, included a control for age cohort, and used multiple imputation to adjustment for a small amount of missing covariate data. *p<.10; **p<.05; ***p<.01

Table 3: Variation in impacts across child subgroups, expressed in effect sizes

	Receptive Vocabulary (PPVT)	Early reading (W-J LW)	Oral Comprehension (W-J OC)	Early numeracy (W-J AP)	Externalizing	Self regulation
Academic skills						
Lowest Quartile	0.08*	0.18***	-0.05	0.09	**0.03**	0.07
Non-lowest quartile	0.16***	0.13**	0.01	0.13***	**-0.13*****	0.01
Language						
Dual-language learner	**0.27*****	0.20***	-0.01	**0.30*****	-0.08	0.03
English only	**0.09*****	0.13***	0.01	**0.06****	-0.06	0.01
Special Needs						
Has special needs	0.21**	0.07	0.02	0.12	**-0.09****	-0.09
No special needs	0.14***	0.18***	0.02	0.13***	**-0.01**	0.05
Age Cohort						
Age 3	0.16***	0.19***	0.03	0.14***	-0.14***	0.02
Age 4	0.11***	0.12***	-0.02	0.11***	-0.01	0.00
Gender						
Male	0.17***	0.13***	0.03	0.12***	**-0.03**	-0.01
Female	0.13***	0.21***	0.01	0.15***	**-0.13*****	0.10**
Child's race/ ethnicity						
Black	**0.05**	0.16***	-0.03	**0.03**	-0.05	0.04
Hispanic	**0.23*****	0.15***	0.02	**0.22*****	-0.10*	0.02
White/other	**0.12*****	0.12***	0.02	**0.04**	-0.03	-0.05

Note: Within each relevant subgroup, models were fit using children with available outcome data, included the standard HSIS covariates, use fixed intercepts for center groups, used the appropriate non-residualized pretest, used data from both cohorts, and included a control for age cohort. Effect sizes were calculated by dividing by the SD of the outcome for the control group. Statistically significant differences between subgroups on a given outcome are indicated in bold and within boxes ($p<.10$). 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 non-missing outcome data, missing data were imputed once, except for the relevant subgroup characteristic which was not imputed. The exception is the Academic Skills subgroup. Because this variable was created based on a continuous pretest and to be consistent with our work on other pretests, we used the imputed version of the academic skills pretest to create the dichotomous Academic Skills variable. N's for the PPVT were 3668 (academic skills, age cohort, and gender subgroups), 3652 (special needs and child race/ethnicity), and 3640 (language). N's for the W-J LW were 3675 (academic skills, age cohort, and gender subgroups), 3568 (special needs), 3659 (child race/ethnicity), and

3647 (language). N's for the W-J OC were 3612 (academic skills, age cohort, and gender subgroups), 3506 (special needs), 3596 (child race/ethnicity), and 3584 (language). N's for the W-J AP were 3649 (academic skills, age cohort, and gender subgroups), 3542 (special needs), 3633 (child race/ethnicity), and 3621 (language). N's for Externalizing were 3681 (academic skills, age cohort, and gender subgroups), 3577 (special needs), 3665 (child race/ethnicity), and 3655 (language). N's for Self-regulation were 3654 (academic skills, age cohort, and gender subgroups), 3546 (special needs), 3638 (child race/ethnicity), and 3626 (language).
*p<.10; **p<.05; ***p<.01