SCHOOL of PUBLIC POLICY

15
Head Start Impact Study

Douglas J. Besharov
Peter Germanis
Caeli A. Higney
and
Douglas M. Call

September 2011

Maryland School of Public Policy
Welfare Reform Academy
www.welfareacademy.org

Part of a forthcoming volume
Assessments of Twenty-Six Early Childhood Evaluations
by Douglas J. Besharov, Peter Germans, Caeli A. Higney, and Douglas M. Call

Note: This report is open to public comments, subject to review by the forum moderator. To leave a comment, please send an email to welfareacademy@umd.edu or fill out the comment form at http://www.welfareacademy.org/pubs/early_education/chapter15.html.

15
Head Start Impact Study

The federal Head Start program, begun in 1965, is designed to "break the cycle of poverty by providing preschool children of low-income families with a comprehensive program to meet their emotional, social, health, nutritional, and psychological needs."[1] After reviewing the full body of Head Start research, a 1997 U.S. Government Accountability Office (GAO) report concluded, "The body of research on current Head Start is insufficient to draw conclusions about the impact of the national program."[2] The Head Start Amendments of 1998 mandated that the Department of Health and Human Services (HHS) determine the impact of Head Start on the children it serves.

Westat, in collaboration with the Urban Institute, the American Institutes for Research, and Decision Information Resources (the "Westat team"), was selected as the principal evaluation contractor[3] and conducted a random assignment evaluation to determine the impact of Head Start. Data collection began in the fall of 2002 and ended in the spring of 2006 when the children had finished first grade. HHS awarded another contract to Westat to continue to track the children through third grade. In the first-year findings of the Impact Study,[4] published in June 2005, the Westat team reports that Head Start has "small to moderate" impacts on fourteen out of thirty measures in the cognitive, socioemotional, parenting, and health domains for three-year-olds and

---

[1]U.S. Department of Health and Human Services, Administration for Children and Families, Head Start Bureau, "Head Start History," (Washington, DC: HHS, 2002), http://www.acf.hhs.gov/programs/hsb/about/history.htm (accessed November 8, 2005).

[2]U.S. Government Accountability Office, *Head Start: Research Provides Little Information on Impact of Current Program* GAO/HEHS-97-59 (Washington, DC: GAO, April 15, 1997), 8.

[3]Michael Puma, Stephen Bell, Gary Shapiro, Pam Broene, Ronna Cook, Janet Friedman, and Camilla Heid, *Building Futures: The Head Start Impact Study. Research Design Plan* (Washington, DC: Administration for Children and Families, March 31, 2001), http://www.acf.dhhs.gov/programs/core/ongoing_research/hs/research_plan_4-251.pdf (accessed July 31, 2003). The cost of the evaluation is $28.3 million. *See* U.S. Government Accountability Office, *Early Childhood Programs: The Use of Impact Evaluations to Assess Program Effects* (Washington, DC: U.S. Government Accountability Office, April 2001), 13.

[4]Michael Puma, Stephen Bell, Ronna Cook, Camilla Heid, and Michael Lopez *Head Start Impact Study: First Year Findings* (Washington, DC: U.S. Department of Health and Human Services, June 2005), http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf (accessed November 8, 2010).

on six out of thirty measures for four-year-olds. In the first-grade follow-up, published in January 2010, the Westat team found basically no differences between the program and control groups.[5] Overall, these findings can be viewed as highly suggestive that Head Start had a very minimal impact, if any, on Head Start children and their parents.

Program Design

Program group. The program group for the Impact Study is, according to the evaluators, a nationally representative sample of low-income three- and four-year-old children who were first-time applicants to Head Start in the fall of 2002. Excluded from the evaluation were Head Start programs that were "very new," those serving only Native American or migrant children, and those determined to be "saturated" (i.e., serving all of the eligible children in their community). The latter requirement was imposed "due to ethical concerns about the potential denial of services to families in locations with relatively few unserved, eligible families."[6] (Below, we discuss the impact of these constraints on the representativeness of the sample.)

Services. Head Start provides comprehensive early childhood services to eligible low-income children and their families. There are four main components of Head Start services—education, parent involvement, social services, and health—but because programs are implemented at the local level, exact services can vary from site to site. Head Start has both full-time and part-time program options, with about 453,885 children enrolled in full-time programs and 383,395 children enrolled in part-time programs at the time of the study.[7] (Full-day enrollment is considered six or more hours per day and part-day enrollment is considered less than six hours per day.)

Education services consist of classroom instruction from a Head Start teacher. Such services usually include language, literacy, and math activities, as well as arts and crafts, games,

---

[5]Michael Puma, Stephen Bell, Ronna Cook, and Camilla Heid, *Head Start Impact Study: Final Report* (Washington, DC: U.S. Department of Health and Human Services, January 2010), http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/hs_impact_study_final.pdf (accessed May 12, 2010).

[6]U.S. Department of Health and Human Services, Administration for Children and Families, "National Head Start Impact Research, Second Report to Congress June 2002," 6, http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/nhs_impact/nhs_impact.pdf (accessed February 22, 2006).

[7]U.S. Department of Health and Human Services, Head Start Bureau, "Head Start Program Information Report for the 2003–2004 Program Year," (Washington: U.S. Department of Health and Human Services, undated). These counts are funded enrollment in center-based Head Start programs. The PIR does not provide separate full-time or part-time enrollment counts for the 86,370 children in home-based and other forms of non-center-based Head Start.

and sports. According the Impact Study, the language and literacy activities most frequently used by Head Start teachers included naming letters, discussing new words, reading to children, and showing children how to read a book. Common math activities included counting aloud, working with shape blocks, learning the days of the week, and counting small toys. The majority of Head Start teachers used either the High/Scope[8] or the Creative curriculum,[9] both of which emphasize "hands on" activities.[10]

Another way to understand the nature of Head Start services is to compare them to the services provided by other center-based child care programs. About 25 percent of three-year-olds in the first year, 36 percent of three-year-olds in the second year, and 35 percent of four-year-olds in the control group attended other center-based programs.[11] The Impact Study details the differences between the services provided by Head Start and these center-based programs. For instance, Head Start teachers were more likely to use either the High/Scope or Creative curriculum than were teachers from other center-based programs. Additionally, the teachers for each group of children were asked how often they generally engaged children in certain types of language, literacy, and math activities. Compared to teachers in other center-based classroom settings, Head Start teachers for the three-year-old group provided significantly more instruction in five out of eleven specific language and literacy activities and in six out of eight math activities. Head Start teachers for the four-year-old group were more likely to use math games, music, and dance to help children learn math concepts, but there were no significant differences in other areas of instruction.

The Impact Study also describes certain characteristics of Head Start and other center-based programs that may be related to the quality of care they provide. Teacher behavior was measured using the Arnett Scale of Lead Teacher Behavior. The Arnett Scale measured teacher's sensitivity, harshness, detachment, permissiveness, and encouragement of independence for both age groups. The Impact Study found that three-year-old Head Start children had teachers who were rated as more sensitive and promoted more independence than did three-year-olds in other center-based programs. Four-year-old children in Head Start had teachers who were rated as less

---

[8]See generally, The High/Scope Educational Research Foundation, "The High/Scope® Curriculum" http://www.highscope.org/EducationalPrograms/EarlyChildhood/homepage.htm (accessed November 8, 2005).

[9]See generally, Diane Trister Dodge, Laura Colker, Cate Heroman, *The Creative Curriculum® for Preschool, 4th Edition* (Washington, DC: Teaching Strategies, Inc., 2002).

[10]Puma et al., 2005, 3-17.

[11]Children attending "other center-based programs" are those enrolled in preschool or child care *classrooms* that did not receive federal Head Start funding, although the center may have received such funding, Puma et al., 2010, 3-8--3-13. This number does not include control group children who were enrolled in classrooms receiving federal Head Start funds and were deemed "crossovers," as discussed under "Assessing the Randomization."

harsh and who promoted more independence than did four-year-olds in other center-based programs.[12]

In 1998, Congress mandated that, by September 30, 2003, 50 percent of all Head Start teachers have a minimum of an associate's degree in early childhood education or in a related field with teaching experience.[13] The Westat team found that, for both the three- and four-year old cohort, about 30 percent of the teachers had an A.A. and another 30 percent had a B.A. However, "Less than half the children in both cohorts had lead teachers who had received 25 hours of training in the last year or received mentoring at least once a month."[14]

Parent education and training, as well as the involvement of parents in Head Start classrooms, are often included as part of Head Start services. The Impact Study describes the parent involvement component of the program: "Historically, Head Start programs have reached out to families in a variety of ways, by encouraging parent involvement in their child's classroom, providing parent education to help strengthen parents' childrearing knowledge and skills, and providing referrals to address family needs so that parents can be more effective in their role as caregiver."[15] The Impact Study does not detail specific parental activities provided by the participating Head Start centers.

The health component of Head Start involves the organization and administration of some health services, including medical, dental, mental health, and nutrition elements. Yet, the main objective is to provide Head Start parents with the necessary skills to assume responsibility for their family's health after the children leave the program.[16] This involves creating linkages between Head Start and other community and state health resources, such as Medicaid, WIC, and the USDA's school meals program. The Impact Study does not describe the specific health services provided by the participating Head Start centers.

---

[12]Puma et al., 2005, 3-14.

[13]U.S. Government Accountability Office, *Head Start: Increased Percentage of Teachers Nationwide Have Required Degrees, but Better Information on Classroom Teachers' Qualifications Needed*, GAO-04-05 (Washington, DC: GAO, October 1, 2003).

[14]Puma et al., 2010, 3-7.

[15]Puma et al., 2005, xvi.

[16]Michael J. Keane, Robert W. O'Brien, David C. Connell, Nicole C. Close, "Executive Summary" in *Descriptive Study of Head Start Health Services: Volume I Summary Report*, prepared for the US Department of Health and Human Services, Administration for Children and Families (Washington, DC: HHS, December 1996), 6, http://www.acf.hhs.gov/programs/opre/hs/descriptive_stdy/reports/descrip_stdy_vol1/hshealth_vol1.pdf (accessed April 3, 2006).

The Evaluation. The Impact Study legislation called for comparisons of "individuals who participate in Head Start programs with control groups (including comparison groups) composed of (i) individuals who participate in other early childhood programs (such as public or private preschool programs and day care); and (ii) individuals who do not participate in any other early childhood program."[17] Congress further directed that the Impact Study examine effects on a broad range of child and family outcomes and that such assessments be conducted upon program completion, at the end of kindergarten, and then again at the end of first grade. Congress also called for information on the circumstances in which Head Start produces its greatest impacts and the types of children who benefit most.

The Impact Study design is based on the random assignment of children and families newly entering Head Start at the start of the 2002–2003 program year to either a Head Start (program) group or a non-Head Start (control) group. Head Start grantees and centers were selected from around the country to represent the diversity of the communities in which Head Start operates and thus reflect the range of quality offered by local Head Start grantees. To obtain a nationally representative sample of Head Start programs (and the children they serve), the Impact Study used a multi-stage sampling process. First, grantee/delegate agencies were clustered, stratified, and randomly selected; next, eligible centers were stratified and randomly selected; and finally, eligible children within these centers were selected and randomly assigned to the Head Start program group or the control group. Some Head Start grantees and centers were deemed ineligible and excluded from the Impact Study—those that were "very new," serving only Native American children, or serving all eligible children (saturated). Hence, children attending these centers are not represented by the sample.

In total, the Head Start Impact Study included 4,667 newly entering children from 383 randomly selected centers in twenty-three different states. About 60 percent of these children were assigned to the Head Start group and about 40 percent were assigned to the non-Head Start group.[18] The non-Head Start group is not a "no service" group, because many of the children in this group (about 19 and 23 percent of three- and four-year-olds, respectively) attended other center-based programs.[19] Thus, the Impact Study compares the impact of Head Start with a range of alternatives that are available to parents in the absence of Head Start, including center-based-programs, parent care, and other care settings. (As discussed under "Assessing the

[17] *Human Services Reauthorization Act of 1998*, sec 117(2)(6)(D).

[18] According to Ronna Cook, the 60/40 split was used because it allowed them to "achieve nearly universal cooperation with random assignment" by minimizing the number of children at each center assigned to the control group. Ronna Cook, Westat, e-mail message to Caeli Higney, February 9, 2006. According to the report, "this imbalance reduces the precision of the impact estimates by less than 2 percent (compared to a balanced 50-50 design)," Puma et al., 2005, 1-11.

[19] This does not include control group children who attended Head Start ("crossovers"), Puma et al., 2005, viii.

Randomization," this could dilute Head Start's measured impact.)

The legislative mandate reflected a concern that possible variation in program impact could be related to "the length of time a child attends a Head Start program (and) the age of the child on entering the Head Start program."[20] Hence, the Impact Study included two separate samples: a newly entering three-year-old group (to be studied through two years of Head Start participation, kindergarten, and first grade), and a newly entering four-year-old group (to be studied through one year of Head Start participation, kindergarten, and first grade).[21] Of the 4,667 children participating in the Impact Study, 2,559 are in the three-year-old group and 2,108 are in the four-year-old group.[22]

Not all of the questions posed by Congress can be answered experimentally. For example, determining how program characteristics affect program impacts requires nonexperimental analysis, because children were not randomly assigned to different program models. There is considerable debate among researchers about the effectiveness of doing so, as Larry Orr of Abt Associates explains:

> The variation in program characteristics we observe is natural variation and may therefore be correlated with other unobserved factors that affect the outcomes of interest. If so, the variation in impacts across sites may be due to these unobserved factors rather than to the program. That is, our estimates of the influence of program features on impacts are potentially subject to selection bias. We cannot, therefore, have the same confidence in them that we have in the overall impact estimates, which are fully experimental. For this reason, if testing alternative program features is an important objective of the research, the experimenter should consider designs . . . in which multiple program variants are implemented at the same site (to hold site effects constant) and participants are randomly assigned to alternative variants (to eliminate systematic differences in participant characteristics across treatments).[23]

At this writing, the evaluation is still underway. In the fall of 2002, baseline data collection began and included in-person interviews with the parent/primary caregiver and one-on-one child

---

[20]As quoted in Puma et al., 2005, 1-10.

[21]Puma et al., 2005,1-10.

[22]Puma et al., 2005, v, note: "The sample of 3-year-olds is slightly larger than the sample of 4-year-olds to protect against the possibility of higher study attrition resulting from an additional year of longitudinal data collection for the younger children."

[23]Larry L. Orr, *Social Experiments: Evaluating Public Programs with Experimental Methods* (Thousand Oaks, CA: Sage Publications, 1998), 208.

assessments. In June 2005, the U.S. Department of Health and Human Services issued the first report on the findings, covering the study's design, methodology, and the impact of Head Start after one program year. In January 2010, HHS released the first-grade follow-up which collected data on the impact of Head Start on program group children after they had completed Kindergarten and first grade. Data collection for the third-grade follow-up is expected to be completed in September 2010.[24]

Major Findings

The preliminary findings after the first Head Start program year indicate that, for both age groups, the actual gains associated with Head Start participation were in limited areas and disappointingly small. By the first grade follow-up, any gains had disappeared and there no were no significant differences between the program and control group on almost every measure.

Cognitive. In the first year follow-up, Head Start's measured impact on cognitive development was small and limited to the areas of pre-reading skills, pre-writing skills, vocabulary knowledge, and parent perceptions of children's emergent literacy skills. No significant effects were found on oral comprehension, phonological awareness, or early math skills. In the first-grade follow-up, almost all these cognitive gains had disappeared.

First-year report. Both three- and four-year-old children who attended Head Start showed statistically significant, but small, gains in their pre-reading skills, compared to the control group children. By the end of the program year, the three-year-old Head Start group scored 5.65 points (0.24 standard deviations) higher on the Woodcock-Johnson III Letter-Word Identification test than the control group and could identify an average of 1.3 more letters (out of 26). The four-year-old Head Start group scored 5.74 points higher on the Letter-Word Identification test (0.22 SD) and could identify an average of 2.28 more letters. After the first year of the program, however, the pre-reading skills of Head Start children were still about one-third of a standard deviation behind the average scores for all U.S. children.

Head Start also had a small impact on the pre-writing skills of three- and four-year-olds. Among Head Start children, three-year-olds showed a statistically significant, but very small, difference (0.13 SD) in their average scores on the Draw-A-Design test compared to the control group, but no significant difference was found on the more advanced Spelling test. Four-year-old Head Start children showed a small difference (0.16 SD) on the Spelling test, but no difference on the Draw-A-Design test.

[24]U.S. Department of Health and Human Services, Administration for Children and Families, Office of Program, Research, and Evaluation, "Head Start Impact Study: Overview," http://www.acf.hhs.gov/programs/opre/hs/impact_study/imptstudy_overview.html#third_grade (accessed June 16, 2010).

Head Start had a moderate impact (0.34 SD) on parental reports of children's emergent literacy skills. However, the Parent-Reported Emergent Literacy Scale, which was used to measure parent perceptions, relies entirely on a parent's response to questions such as how many letters of the alphabet the child knows and whether the child can count.

Head Start children showed limited to no gains in vocabulary knowledge. Three-year-old Head Start children scored 4.23 points higher on the Peabody Picture Vocabulary Test, Third Edition (0.12 SD) and could name about one more color (0.1 SD) than the control group. Head Start also did not have a significant impact on the vocabulary measures of four-year-old children.

Oral comprehension and phonological awareness—skill areas that relate to children's emergent literacy and later academic achievement—were also tested.[25] Head Start did not have a statistically significant effect on these skills for either age group. Head Start also did not have a significant impact on the early math skills of three- and four-year-olds. Math skills were measured with two tests—the Woodcock-Johnson III Applied Problems test and the Counting Bears test—that "assess basic skills and understandings that are essential for the development of more advanced quantitative capabilities and are predictive of mathematics achievement in kindergarten and first grade."[26]

Within the areas where statistically significant effects were found, the evaluators deemed the impacts "small" to "moderate" in size. But it seems unlikely that they are large enough to result in meaningful gains for Head Start children. There are several reasons these findings could be considered disappointing.

For Head Start four-year-olds, there were statistically significant impacts in only six out of thirty measures (itself a statistically suspect result). Of these six measures, only three measures—the Woodcock Johnson Letter-Word Identification test, the Spelling test and the Letter Naming Task—directly test cognitive skills, and they show a slight improvement in only one of three major predictors of later reading ability (letter identification). Head Start four-year-olds were able to name about two more letters than their non-Head Start counterparts, but there were no significant differences on vocabulary and oral comprehension or sensitivity to sounds.

Head Start produced more impacts for three-year-olds (statistically significant results on fourteen out of thirty measures); however, the measures that showed the most improvement tended to be superficial. For instance, Head Start three-year-olds were able to identify one and a half more letters and they showed a small gain in vocabulary. Yet, Head Start three-year-olds came only 8 percent closer to the national norm in vocabulary tests—a very small relative gain—and they showed no improvement in oral comprehension, phonological awareness, or early

[25]Puma et al., 2005.

[26]Puma et al., 2005, 5-9.

math skills. This is concerning because, as Jean Layzer of Abt Associates points out, vocabulary and oral comprehension skills tend to be more indicative of later reading comprehension.[27] The Impact Study also notes that vocabulary tests are "strongly predictive of children's general knowledge at the end of kindergarten and first grade."[28]

First-grade follow-up. The findings from the first grade follow-up of the Head Start Impact Study, released in January 2010, were even more disheartening. The earlier, minimal (or "modest") gains had disappeared—dashing the hopes of those who thought that the early gains would result in later increases in school achievement. According to the Westat report: "Yet, by the end of 1st grade, there were few significant differences between the Head Start group as a whole and the control group as a whole for either cohort."[29]

For the three-year-old cohort, by the end of Kindergarten, statistically significant differences were found on only two of the nineteen cognitive measures. Children in the program group who were tested in Spanish had statistically significantly higher vocabulary scores than the control group (an effect size of 0.26 SD). More surprising, children in the program group had significantly *lower* scores in math ability than the control group (an effect size of -0.19 SD). By the end of the first grade, significant differences were only found on one of twenty-two measures (oral comprehension) and the effect size was so small (0.08 SD) as to be practically meaningless.

For the four-year-old cohort, by the end of Kindergarten, there were *zero* statistically significant differences on nineteen cognitive measures. By the end of first grade, there was only one statistically significant difference; children in the program group had significantly higher vocabulary scores, but the effect size was only 0.09 SD.

Nevertheless, the design of the Impact Study is a reasonable approach for addressing the main congressionally mandated questions concerning the impact of Head Start on school readiness compared with children who participate in another preschool program or in no preschool program. The first-year findings of the Head Start Impact Study suggest that the program has limited impact on newly entering three- and four-year-olds.

School readiness/performance. See above under "Cognitive" findings.

---

[27]Jean Layzer, Abt Associates, e-mail message to Caeli Higney, July 29, 2005.

[28]Puma et al., 2005, 5-7.

[29]Michael Puma, Stephen Bell, Ronna Cook, and Camilla Heid, *Head Start Impact Study: Final Report* (Washington, DC: U.S. Department of Health and Human Services, January 2010), xxv, http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/hs_impact_study_final.pdf (accessed May 12, 2010).

Socioemotional development. The Impact Study reports that Head Start had a small, positive effect on a few social and emotional outcomes for three-year-olds in the program compared to the control group. No statistically significant effects on socioemotional development outcomes were found for Head Start four-year-olds. The analysis relied solely on behavior reports from parents and did not include reports from teachers or other caregivers, because such reports were not available for the non-Head Start group children who were in parental care.

Head Start three-year-olds' average score on the Total Behavioral Problems scale at the end of the program year was 0.5 points lower (-0.13 SD) than that of the children in the control group. Head Start three-year-olds also scored 0.3 points lower than the control group (-0.18 SD) on the Hyperactive Behavior Scale.

Head Start had no impact on the social skills and approaches to learning of either three- or four-year-olds, as reported by parents, the Social Skills and Positive Approaches to Learning (SSPAL) of children, and the Social Competencies Check List (SCCL).

First-grade follow-up. In the three-year-old cohort, by the end of Kindergarten, parents reported that children in the program group were less likely to have hyperactive behavior (an effect size of 0.12 SD) and more likely to have better social skills (an effect size of 0.14 SD) compared to the control group. There were no statistically significant differences on seven other parent-reported measures and nine teacher-reported measures. By the end of first grade, the previously significant differences had disappeared, but two other significant differences appeared. Parents reported that children in the program group were more likely to show closeness (an effect size of 0.10 SD) and to have positive relationships (0.10 SD). There were no statistically significant differences on the other parent-reported measures or on the nine teacher-reported measures.

In the four-year-old cohort, by the end of Kindergarten, there were no statistically significant differences on the nine parent-reported measures or the eleven teacher-reported measures. By the end of first grade, parents reported that children in the program group were less likely to have withdrawn behavior (an effect size of -0.13 SD), while teachers reported that children in the program group were more likely to be shy or reticent (an effect size of 0.19 SD) and more likely to have problems with teacher interaction (an effect size of 0.13 SD). There were no other statistically significant differences on the other seventeen measures.

Health. In the first-year report, the Impact Study found that Head Start had positive, moderate effects on some indicators of children's health. Yet, as the Impact Study notes, "The analysis of Head Start's impact on children's health is based solely on reports from parents. No direct measurement of children's actual health status, or their receipt of health care services, was

undertaken for this study."[30]

There was a statistically significant difference between the receipt of dental care by Head Start children and the control group, both in the three-year-old (17 percentage difference, or 0.34 SD) and four-year-old groups (16 percentage difference, or 0.32 SD). Among three-year-olds, Head Start had a statistically significant, but small, impact on parental reports of the child's health status being excellent or very good (0.12 SD). Head Start did not have a significant impact on whether the child had health insurance, needed ongoing care, or had care for an injury in the last month.

As with other indicators, by the first year follow-up, many of these gains had disappeared. For the third-grade cohort, by the end of Kindergarten, children in the program group were more likely to have health insurance than children in the control group, an effect size of 0.14 SD. There were no other statistically significant differences on four other health measures. By the end of first grade, there were no statistically significant differences on any of the five health measures

For the four-year-old cohort, by the end of Kindergarten, children in the program group were more likely to have health insurance (an effect size of 0.11 SD) and more likely to have excellent or good health (effect size of 0.13 SD), however both differences are quite small. There were no significant differences on the other three health measures. By the end of first grade, only the difference in health care status remained (an effect size of 0.11 SD).

Behavior. See socioemotional development.

Crime/delinquency. Data apparently either not collected or not reported.

Early/nonmarital births. Data apparently either not collected or not reported.

Economic outcomes. Data apparently either not collected or not reported.

Effects on parents. The Impact Study found that Head Start had small, positive effects on parenting practices. Once again, however, these were self-reported effects and therefore should be viewed with caution. Although there was a statistically significant difference between the control and the program group, it is possible that program group parents became more aware of appropriate behaviors, such as the importance of reading to children, and were thus more likely to report such practices, without actually changing their behaviors.

For both three- and four-year-olds, Head Start had a positive average effect on the amount of time parents reported reading to their child (0.18 SD and 0.13 SD, respectively). For children

[30]Puma et al., 2005, 7-1.

in the three-year-old group, Head Start also had a statistically significant effect on the Family Cultural Enrichment Scale, meaning that Head Start parents in this group were more likely (according to their own reports) to provide children with enrichment activities than were the parents in the control group. Parents of three-year-old Head Start children were also 14 percent less likely to report the use of spanking in the last week. Head Start did not have an impact on parents' child safety practices at home.

First-grade follow-up. For the three-year-old cohort, by the end of Kindergarten, parents of children in the program group were less likely to have spanked their child in the last week (an effect size of -0.09 SD) and less likely to have used time out in the last week (an effect size of -0.13). There were no other statistically significant differences on the other nine parenting measures. By the end of first grade, the effect for usage of time out persisted (an effect size of -0.11 SD) and parents of children in the program group were more likely to use an authoritarian style of parenting (an effect size of -0.11 SD). There were no other statistically significant differences on the other parenting measures.

For the four-year-old cohort, there were no statistically significant differences on any of the eleven parenting measures by the end of both Kindergarten and first grade.

Subgroup effects. The Impact Study examines Head Start's impact on a number of different subgroups and moderating factors, including whether the child has special needs, child race/ethnicity, child gender, language of child assessment, caregiver depression, caregiver marriage status, and home language. As the Impact Study notes, however, "Because data are limited, the study cannot decisively answer all questions about Head Start's impact on different subpopulations."[31]

The only subgroup characteristic that exhibited an overall difference in the impact of Head Start was a mother's depression symptoms. According to the Impact Study, Head Start's measured impact consistently decreased as the levels of a caregiver's reported depressive symptoms increased.

Within child language subgroups, more effects were found for English-speaking children than for Spanish-speaking children. Head Start had a small, statistically significant effect on the pre-reading, pre-writing, and vocabulary skills of English-speaking children, but had a statistically significant effect only on the vocabulary skills of Spanish-speaking children. For instance, Head Start had a positive effect on the Woodcock-Johnson III Letter-Word Identification test for both three- and four-year-old English speaking children, but no effect on Spanish-speaking children. Rather, Spanish-speaking children exhibited positive effects on the Peabody Picture Vocabulary Test (PPVT) III and the Color Naming task. They moved 13 percent closer to the national norm

[31]Puma et al., 2005, 4-22.

on the PPVT III and could name two more colors than the Spanish-speaking children in the non-Head Start group.

There was no consistent trend in the cognitive gains exhibited by different racial/ethnic groups, although the evaluators argue that "there is particularly strong evidence that Head Start is having a positive impact on the cognitive development of minority children."[32] Specifically, for three-year-old Hispanic children, Head Start had positive impacts on the Letter Word Identification test, the Letter Naming test, the PPVT-III, and the Spelling test. For three-year-old African American children, positive impacts were found on the Letter Word Identification test, phonological awareness, and the Draw-a-Design pre-writing task, while for four-year-old African American children impacts were found on the Letter-Word Identification test and the Spelling test. Among white children, Head Start had a positive impact on the oral comprehension of three-year-olds and on the Letter Naming Task for four-year-olds.

An examination of Head Start's effect on special needs children compared to other children did not find positive impacts in pre-reading and pre-writing skills for special needs children.[33]

Benefit-cost findings. Apparently a benefit-cost analysis was not performed. In 2005, Head Start cost approximately $7,287 per child (in 2005 dollars).[34]

Overall Assessment

Program theory. The Head Start Impact Study notes that, "Since its beginning in 1965, Head Start's goal has been to boost the school readiness of low-income children. The premise underlying the program is that low-income children do not receive the same level of intellectual stimulation at home as middle-class children."[35] To this end, Head Start aims to close the gap between low-income children and other children by promoting school readiness through intellectual stimulation and social development.

In the past decade, Head Start has adopted the "whole child" theory of child development,

[32]Puma et al., 2005, 5-12.

[33]Puma et al., 2005, 5-13.

[34]U.S. Department of Health and Human Services, Administration for Children and Families, Head Start Bureau, *Head Start Program Fact Sheet: Fiscal Year 2005* (Washington, DC: HHS, 2006), http://www.acf.hhs.gov/programs/hsb/research/2006.htm (accessed January 19, 2005).

[35]Puma et al., 2005, 1-1.

as recommended by the Goal One Technical Planning Group.[36] The panel identified five development domains that are important to the child's readiness for school: physical well-being and motor development, social and emotional development, approaches to learning, language usage and emerging literacy, and cognition and general knowledge.[37] To address these domains, Head Start provides comprehensive services including preschool education; medical, dental, and mental health care; nutrition; and parental involvement.

The legislation guiding the Impact Study mandated that the evaluation measure the "school readiness" of Head Start participants (and the control group), which seems appropriate within the context of the program's theory.

Program implementation. The actual implementation of Head Start services was not discussed in the first-year findings of the Impact Study, nor in earlier reports from the evaluation. Based on other recent studies of Head Start, it seems reasonable to assume that there were no major problems with program implementation, although program services and quality vary from site to site.[38]

Assessing the randomization. The Impact Study staff were able to implement random assignment successfully in all but five of the 383 selected Head Start centers, resulting in a final sample of 378 centers.

In order to obtain a nationally representative sample of Head Start programs (and the children they serve), the Impact Study used a multi-stage sampling process. First, a list of Head Start grantee/delegate agencies was developed, excluding "very new," migrant, Tribal Organization, and Early Head Start-only agencies. The remaining grantee/delegate agencies were organized into geographic clusters and then stratified on the basis of state prekindergarten and

---

[36]The Goal One Technical Planning Group was established to address the first of the National Education Goals issued by President George H.W. Bush in 1990—that "by the year 2000 all children in America will start school ready to learn." See U.S. Department of Health and Human Services, Administration for Children and Families, *Head Start FACES 2000: A Whole-Child Perspective on Program Performance* (Washington, DC: U.S. Department of Health and Human Services, May 2003), 2.

[37] U.S. Department of Health and Human Services, Administration for Children and Families, *Head Start FACES 2000: A Whole-Child Perspective on Program Performance* (Washington, DC: U.S. Department of Health and Human Services, May 2003), 2.

[38]The Head Start FACES 2000 study notes, "Head Start quality has been observed to be consistently good over time, using a variety of indicators including child-adult ratio, teacher-child interactions, and classroom activities and materials. Few classrooms scored below minimal quality." U.S. Department of Health and Human Services, *Head Start FACES 2000: A Whole-Child Perspective on Program Performance* (Washington, DC: U.S. Department of Health and Human Services, May 2003), iv, http://www.acf.hhs.gov/programs/opre/hs/faces/reports/faces00_4thprogress/faces00_4thprogress.pdf (accessed November 8, 2010).

child care policy; child race/ethnicity; urban/rural location; and region. One cluster of programs was then randomly selected from each of the strata with a probability proportional to Head Start enrollment.

According to the Impact Study, "To be eligible for inclusion in the study sample, grantee/delegate agencies had to have enough 'extra' or additional newly entering applicants beyond their number of funded slots to allow for the creation of a non-Head Start control group."[39] Thus, grantee/delegate agencies that were "saturated" (where all eligible Head Start children are being served) were excluded, resulting in a 11 percent reduction in the number of grantee/delegate agencies available to the study.[40]

The remaining grantee/delegate agencies were then stratified along several dimensions including urban location, auspice (school based compared to all other agency types), percentage of Hispanic and African American enrollment, program options offered, and the percentage of newly entering three-year-olds. Approximately three grantee/delegate agencies were randomly selected from each of the twenty-five strata.

After all selected grantee/delegate agencies were visited (three out of ninety agencies were dropped from the study), the next step was to identify operating Head Start centers. The selected grantee/delegate agencies provided a list of all Head Start centers that could be included in the Impact Study. About 11 percent of the identified, operating Head Start centers were dropped because they were determined to be saturated. Center groups were then created by combining small centers with nearby centers, and the groups were stratified using the same characteristics used for the selection of grantee/delegate agencies. An average of three center groups were selected from each eligible grantee/delegate agency, resulting in a main sample of 448 centers in 84 grantee/delegate agencies. It was later determined, however, that some of these centers were ineligible for inclusion in the Impact Study.[41] Thus, 103 centers were dropped from the Impact Study and thirty-eight replacement centers were added, resulting in the final sample of 383 centers.

There were four points in the sample selection process when grantees/delegate agencies or centers were dropped due to saturation. In fact, a comparison of the characteristics of saturated

---

[39]Puma et al., 2005, 1-7. According to Ronna Cook, "Saturation was determined by the evaluators after site visits to the programs and centers by senior project staff, telephone calls with community agencies, and the collection of program and center enrollment data for several years," Ronna Cook, Westat, e-mail message to Caeli Higney, February 9, 2006.

[40]Puma et al., 2005, 1-7.

[41]On the basis of more in-depth or up-to-date information, it was found that some centers either had recently closed or merged; served only Early Head Start children; were collaborations between Head Start and private preschool programs; or were saturated. Puma et al., 2005, 1-8.

versus non-saturated grantees/delegate agencies reveals various differences between the two. Saturated grantees/delegate agencies were much smaller, with an average newly entering enrollment of 113 children compared to 388 children in non-saturated programs. They were also more likely to be school-based and had smaller percentages of Hispanic enrollment than the non-saturated grantees/delegate agencies. Likewise, saturated Head Start centers were significantly smaller, had fewer Hispanic children, and had a larger percentage of newly entering three-year-olds than did non-saturated Head Start centers. The Westat team notes, "While some differences were found between the saturated and non-saturated sites, non-response adjustments were made to the weights. Subsequent sensitivity analyses have shown that we can be comfortable with the conclusion that our estimated impacts are not likely to have been different in an important way if we had been able to include saturated sites in the experimental study."[42] Of course, the evaluators cannot control for unobserved differences between the saturated and unsaturated grantee/delegate agencies and centers.

Once the participating sites were selected, the goal was to randomly assign children to a program group (sixteen children in each site) or a control group (eleven children in each site). Local staff were allowed to exclude a small number of "very high risk" children from the random assignment process. Out of a total of about 18,000 applications only 276 exclusions were requested, so the evaluators believe that their omission had a negligible effect on the representativeness of the sample. A comparison of Head Start and non-Head Start children (using weighted data) shows no statistically significant differences between the two randomly assigned groups.

The Impact Study, however, does reveal some possible problems with the makeup of the program and control groups, which arose after the random assignment took place.

Within the program group, there were some children who did not participate in Head Start at any time during the year ("no-shows"). Among the three- and four-year-old Head Start groups, no-shows accounted for 15 and 20 percent of the full randomly assigned sample, respectively (12 and 17 percent of the analysis sample). It is possible that the no-shows could reduce the measured impact of Head Start on the program group. The Impact Study notes, however, that "the best way to estimate Head Start's impact on the average participant does not require that one knows anything about why no-shows arise, or how they differ from other families and children in the sample."[43] The assumption that Head Start has zero impact on the children and families who never receive services (no-shows) makes it possible to translate the measured effect of the program on the entire Head Start sample into the average effect of Head Start on just the participants. The overall impact of Head Start on all children is calculated as a weighted average of the impact on participants and the impact on the no-shows (which is zero).

---

[42]Ronna Cook, Westat, e-mail message to Caeli Higney, February 9, 2006.

[43]Puma et al., 2005, 4-30.

Hence, the Impact Study calculated the impact of Head Start participation with the no-show adjustment for the detailed subgroups. Among both the three- and four-year-old groups, all of the statistically significant impacts for various subgroups remained significant after the no-show adjustment. The size of the impacts sometimes increased, but usually the increase was small. The largest change in impact was among the African American subgroup. For example, four-year-old African-American scores on the Letter-Word Identification test increased by 3.5 points, with the effect size increasing from 0.4 to 0.53 SD after the no-show adjustment was made.[44]

The different racial/ethnic makeup of the three- and four-year-old Head Start groups is a concern. Within the three-year-old Head Start group, there were similar percentages of Hispanics and blacks (37.4 percent and 32.8 percent, respectively). The four-year-old-group, however, was composed of notably more Hispanics (51.6 percent) than blacks (17.5 percent). According to the Westat team, the difference in the racial/ethnic composition of the four-year-old group is partially due to sampling error from the sampling of the primary sampling units (PSUs) and programs.[45] The FACES 2000 data also reveal more Hispanics than blacks in the four-year-old group (30 percent and 17 percent, respectively) compared to the three-year-old group (39 percent and 27 percent, respectively). The Westat team notes that "differences in race/ethnicity reporting procedures on the PIR (program information report), HSNRS (Head Start National Reporting System), and the Impact Study, and in the population being compared (newly entering versus all children) also contribute to the difference being observed."[46] Thus, they conclude that "These differences do not necessarily indicate there is a systematic bias in the Impact Study sample with respect to race/ethnicity."[47]

Within the control group, some children attended other center-based programs (preschool or child care classrooms that did not receive federal Head Start funding) and some children attended Head Start, two occurrences that also could have the potential to dilute the impact of Head Start.

As noted earlier, the control group is not a "no service" group. Thus, it could be argued that the Impact Study found minimal effects because the control group included children who were in other developmentally appropriate settings, usually center-based programs. If these settings provided services similar to those provided by Head Start, they might be expected to produce similar effects, thus reducing the measured effects of Head Start. Some members of the Advisory Committee on Head Start Research and Evaluation expressed this concern: "Some Head

[44]Puma et al., 2005, Exhibit A.5.1.1, 5.1-3 –5 and Exhibit A.5.1.2, 5.1-6–7.

[45]Ronna Cook, Westat, e-mail message to Caeli Higney, February 9, 2006.

[46]Ronna Cook, Westat, e-mail message to Caeli Higney, February 9, 2006.

[47]Ronna Cook, Westat, e-mail message to Caeli Higney, February 9, 2006.

Start programs (particularly the best) are likely to have influenced other child care and prekindergarten programs available to low-income children, so that the environments of control group children have been influenced (or, in research terms, "contaminated") by the Head Start treatment."[48]

About 25 percent of three-year-olds and 35 percent of four-year-olds in the control group attended other center-based programs, however, these percents are not adjusted for children who attended Head Start (crossovers).[49] Either these children would have to exert a very strong influence on the rest of the control group, or the other care settings attended by the control group (including parental care, relative care, and non-relative care) would also have to provide services that could produce similar positive effects. If the control group children who attended center-based care did substantially dilute the sample, this might imply that low cost center-based child care can produce similar effects as a high cost program such as Head Start, drawing into question Head Start's cost-effectiveness.

In addition to children who attended other center-based programs, some control group children actually participated in Head Start at some time during the year ("crossovers"). Among the full randomly assigned sample, crossovers accounted for about 18 percent of three-year-olds and about 14 percent of four-year-olds in the first year. In the three-year-old cohort's second year, the embargo on Head Start participation was lifted on children in the control group. In that year, about 50 percent of control group children participated in Head Start. [50]

The occurrence of crossovers can affect the integrity of the sample and, thus, the resulting average impact of program participation. As Richard Berk, professor of Criminology and Statistics at the University of Pennsylvania, and Peter Rossi, former professor at the University of Massachusetts (Amherst), note, "Crossovers produce control groups that are not no-treatment

[48]Advisory Committee on Head Start Research and Evaluation, 1993, v.

[49]We think it appropriate to remove the crossovers from this group because after the Impact Study makes adjustments for the crossovers, almost all of Head Start's effect sizes remain in the same range. The Impact Study's procedure for making this crossover adjustment is detailed under "Assessing the Randomization" and, more thoroughly, in chapter four of the Impact Study (pp. 4-33–7).

[50]The Westat team explained that control group participation in Head Start in the three-year-old cohort's second year did not constitute "crossover." They explain, "The control group children in this age cohort were not supposed to represent outcomes in a world entirely without Head Start, but rather (as discussed below) a world where Head Start only becomes available at age 4." Michael Puma, Stephen Bell, Ronna Cook, and Camilla Heid, *Head Start Impact Study Technical Report* (Washington, DC: U.S. Department of Health and Human Services, January 2010), 5-35, http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/hs_impact_study_tech_rpt.pdf (accessed July 2, 2010).

groups but instead are combinations of no-treatment and participation in competing programs."[51]

In order to deal with the possible contamination of the non-Head Start sample by the children who participated in Head Start, the Impact Study removed the crossover cases from the non-Head Start sample and recalculated the impact of the program without them. This method, however, raised a new possibility of selection bias because the crossovers may have had particular characteristics that were different from those of the control group as a whole. Thus, the researchers recalculated the analysis weight adjustments (which were originally used to account for non-responses) and applied these weights to the crossovers. The crossovers were then statistically "put back" into the non-Head Start sample by increasing the analysis weights of other observations with outcome data that had similar background characteristics. As the Westat team admits, this was an imperfect method for dealing with the problem posed by crossovers because "there is no assurance that the combination of these methodologies effectively or completely offsets the potential bias of using an incomplete sample."[52] Nevertheless, this certainly seems to have been the most reasonable thing to do under the circumstances.

The Impact Study provides a set of estimates of Head Start's affect on participants adjusted to remove the influence of both no-shows and crossovers, but notes that these estimates "may be subject to uncorrected selection bias up or down."[53] These adjustments were made only on the findings for the entire age group, not for the more detailed subgroups and moderating factors. After the adjustments were made for the three-year-old group, all statistically significant findings but one, the impact on the Color Naming/Identification test, remained statistically significant.[54] In some cases, the size of the impacts increased, but only slightly, and all of the effect sizes remained in the small to moderate range. For instance, Head Start's effect on the number of letters an average three-year-old child could name increased from 1.3 letters to about 2.1 letters, and the impact on the PPVT-III rose from 4.23 points to 4.5 points after the adjustments were made.[55] Among the four-year-old group, all statistically significant findings remained significant. Head Start's impact on the number of letters an average child could name increased by 0.11 letters, from 2.28 letters to 2.39 letters, and its impact on the Spelling test rose from 4.14 to 4.74 points after the adjustments were made.[56] Again, the effect sizes remained in the same range,

---

[51]Robert G. St.Pierre and Peter H. Rossi, "Randomize Groups, Not Individuals: A Strategy for Improving Early Childhood Programs," *Evaluation Review,* forthcoming.

[52]Puma et al., 2005, 4-36.

[53]Puma et al., 2005, 4-37.

[54]Puma et al., 2005, Exhibit A.5.1.1, 5.1-3–5.

[55]Puma et al., 2005, Exhibit A.5.1.1, 5.1–3.

[56]Puma et al., 2005, Exhibit A.5.1.2, 5.1-6–7.

never increasing by more than 0.06 SD.

The Westat team does not believe the incidences of crossovers and no-shows seriously weaken their findings: "At worst, violations of random assignment that extend Head Start's services to some children in the non-Head Start group and reduce the exposure to Head Start among the treatment group make it harder to detect the average impact of the program . . . These considerations should increase the confidence that any observed statistically significant impacts are real and important."[57] The estimates of Head Start's impact after the no-show and crossover adjustments seem to support this statement, but also indicate that effects sizes would not be significantly larger in the absence of crossovers and no-shows.

In sum, the random of assignment of children to the program and control groups was conducted with relatively few problems, and there were no statistically significant differences between the randomly assigned Head Start and non-Head Start children (using weighted data). The Westat team notes, "This suggests that the initial randomization was done with high integrity and that the samples can provide the necessary confidence in the validity of the impact estimates."[58] Deviations from the random assignment occurred (no-shows and crossovers), and the Impact Study provides estimates to account for both.

Assessing statistical controls in experimental and nonexperimental evaluations. The evaluation was based on random assignment, so selection bias was not a serious problem. As describe above, however, possible bias may have arisen due to deviations from the random assignment. Covariates were used to increase the precision of the reported impact estimates.

Sample size. A total of 4,667 three- and four-year-old newly entering children were randomly assigned and are included in the Impact Study sample. According to the evaluators, an analysis of the characteristics of the sample indicates that it is nationally representative of new Head Start enrollees.

Attrition. Based on the fall 2002 and spring 2003 data collection, there do not seem to be attrition problems. In fact, the response rate for spring 2003 was higher than the response rate for fall 2002. Overall, 83 percent of parents who had children participating in the Impact Study completed interviews at both points in time, and 82 percent of the children were assessed.

Despite the relatively high response rate, bias can occur as a result of differences between the responding and nonresponding centers and parents; thus, adjustments were made to account for nonresponse. A statistical "hot-deck" procedure was used to impute missing background

[57]Puma et al., 2005, 2-6.

[58]Puma et al., 2005, 2.3.

variables for children with either no fall 2002 parent interviews or incomplete fall 2002 data caused by item nonresponse.[59] To compensate for nonrespondents, the analysis weights (which initially were based on the probability of selection into the study sample during random assignment) were increased for responding children with similar background characteristics on variables that were measured for all randomly assigned cases in 2002 (prior to random assignment).[60]

The year one analysis sample, consisting of the 3,898 children for whom data was collected, reveals that the Head Start and non-Head Start children, in both the three- and four-year-old groups, differ in two respects (using weighted data). The primary caregivers of children in the three-year-old Head Start group are, on average, 0.9 years older than the caregivers in the non-Head Start group. The three-year-old Head Start children are also somewhat (1.9 percent) more likely to have a grandparent living with them. Among the children in the four-year-old group, mothers of the Head Start children are 6.5 percent more likely to receive education beyond high school and are 4.5 percent less likely to receive assistance through the Temporary Assistance for Needy Families (TANF) program. The Impact Study notes that "these differences may arise from the lag in fall 2002 data collection after the point of random assignment."[61]

Data collection. The data collection relied on a wide range of standardized tests and surveys. The data sources were appropriate for the questions being studied.

Measurement issues. The data collection plan does not raise any apparent measurement issues. The data collection consists of twice yearly in-person interviews with parents in the first year of the Impact Study (i.e., 2002–03) and fall telephone interviews and in-person interviews with parents in the spring in subsequent years. In-person child assessments were conducted in the fall and spring of the first year and will be conducted annually in the spring of future years. The second spring data collection (spring 2004) has been completed, but had not been analyzed as of this writing. The final components consist of annual surveys of providers and teachers, direct classroom observations, and teacher ratings of children. This includes child data related to school readiness, such as physical well-being and motor development, social and emotional development, approaches to learning, language usage and emerging literacy, cognition and general knowledge. Additionally, information was collected on parenting practices, family resources and risk factors, and the socioeconomic characteristics of participating families.

Generalizability. Practical considerations make it difficult to obtain a truly nationally representative sample of Head Start centers and, hence, Head Start children. As noted above,

[59]Puma et al., 2005, 4-9.

[60]Puma et al., 2005,. 4-9.

[61]Puma et al., 2005, 2-7.

Head Start centers that were "very new," or were "saturated" were excluded from the random assignment process. The excluded centers represent an estimated eight percent of the Head Start center universe, approximately 5 percent of the total number of three- and four-year-old children enrolled in Head Start, and 4.7 percent of newly-entering three- and four-year-old applicants.[62] These exclusions were reasonable and affected a relatively small number of children, and they do not substantially limit the generalizability of the findings (although they may have affected measured impacts).

Ideally, all newly entering three- and four-year-olds would have the possibility of being included in the study, thus the "coverage rate" would equal 100 percent. Newly entering Head Start children in saturated communities, however, were not eligible for participation in the Impact Study and, therefore, are not represented in the sample. Taking this into account, the evaluators estimate a coverage rate of 84.5 percent. In other words, they estimate that the Impact Study sample is representative of 84.5 percent of all newly entering Head Start three- and four-year-olds across the country.[63]

Replication. This study is a more methodologically rigorous replication of past studies that attempted to measure the impact of Head Start due to its use of random assignment to create statistically equivalent Head Start and non-Head Start groups. It is too early for possible replication of this study.

Evaluator's description of findings. The Westat team describes the first year findings from the Impact Study: "For children in the three-year-old group, the preliminary results from the first year of data collection demonstrate small to moderate positive effects favoring the children enrolled in Head Start for some outcomes in each domain." They note, "Fewer positive impacts were found for children in the four-year-old group." The Westat team describes the Impact Study's methods and findings in a straightforward manner and were careful to point out the limited size and significance of some effects. Yet, they report effect sizes as small as 0.1 standard deviations, even though other analysts do not consider such effects to be meaningful (see Appendix 1 for a further discussion of effect sizes and their interpretation.). They are careful to present the underlying data behind these effect sizes (such as the difference in the number of letters recognized), so that readers can draw their own conclusions. The Westat team also acknowledges that potential bias could have arisen during the randomization process[64] (as a result of the exclusion of saturated grantees and centers).

Referring to the gains that the Head Start children made on the Woodcock-Johnson III

[62]Puma et al., 2001, 13–14.

[63]Puma et al., 2005, 2-2.

[64]Puma et al., 2005, 2.1-1.

Letter-Word Identification test, the Westat Team notes, "At the end of one year, Head Start was able to nearly cut in half the achievement gap that would be expected in the absence of the program."[65] In reaching this conclusion, the Westat team compared the gap between the average score of the three-year-old Head Start group and the national average and the gap between the non-Head Start group and the national average. The Head Start group's gap was 4 points and the control group's gap was 7.6 points, meaning that the Head Start group had a deficit that was 3.6 points or 47 percent smaller (3.6/7.6=0.47). Another way to examine Head Start's impact on this outcome measure is to compare the groups' mean scores on the test. The Head Start group had a mean score of 307, while the non-Head Start group had a mean score of 300.5—a difference of 6.5 points. Thus, the Head Start group's mean score was only about 2 percent higher than the non-Head Start group (6.5/300.5=0.021). The Westat team discusses only the former calculation, which implies a larger difference than the latter calculation would suggest.

For the first-grade follow-up, the Westat team writes: "Yet, by the end of 1st grade, there were few significant differences between the Head Start group as a whole and the control group as a whole for either cohort."[66]

Evaluator's independence. The evaluators—Westat, the Urban Institute, the American Institutes for Research, and Decision Information Resources—were selected through a competitive process. They were, however, under contract with the U.S. Department of Health and Human Services.

Statistical significance/confidence intervals. Statistical significance was measured and reported at the 0.1 percent, 1 percent, and 5 percent levels. All effects that were reported as "statistically significant" were significant at least at the 5 percent level.

Effect sizes. Effect sizes were calculated by taking the difference in mean scores between the Head Start group and the control group and dividing by the standard deviation (of the measure of interest) and were reported in standard deviation (SD) units.

After approximately one year of Head Start participation, most cognitive effect sizes fell within the range of 0.1 to 0.4 SD. The Westat team describes effect sizes smaller than 0.2 SD as

[65]Puma et al., 2005, xi.

[66]Michael Puma, Stephen Bell, Ronna Cook, and Camilla Heid, *Head Start Impact Study: Final Report* (Washington, DC: U.S. Department of Health and Human Services, January 2010), xxv, http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/impact_study/hs_impact_study_final.pdf (accessed May 12, 2010).

"small" and those between 0.2 and 0.5 SD as "moderate."[67]

This reflects a move away from the traditional demarcations, which would consider effect sizes in the range of 0.2 to 0.5 SD to be "small." For instance, the Westat team describes effect sizes in the range of 0.19 to 0.24 SD as "modest but meaningful,"[68] while under traditional demarcations, these effects sizes would be considered small and not relevant to policy. In some cases small effects may be important, but in most circumstances we agree with the traditional demarcations. (See Appendix 1 for a further discussion of effect sizes and their interpretation.)

Sustained effects. The evaluation examined short-term post-intervention impacts.

Benefit-cost analysis. Apparently not performed.

Cost-effectiveness analysis. Apparently not performed.

[67]Puma et al., 2005, vi.

[68]Puma et al., 2005, 5-5.

Commentary

Michael Puma, Ronna Cook, Stephen Bell, and Camilla Heid

We appreciate the opportunity to comment on the chapter describing the design of the Head Start Impact Study and the findings reported to date. We respect the authors' right to present their interpretation of the findings, and appreciate their overall favorable critique of the study's approach and methods. However, we stand by the original interpretation of the results as provided in the Head Start Impact Study: First Year Findings.

The preliminary results presented from the first year findings show that Head Start increases 3-year-old children's cognitive and social emotional development and children's health, as well as positive parenting practices (all the domains examined in the study). Domains consisted of multiple measures and impacts were found on some of the measures in each of these domains. Findings were also positive, though less prevalent, for four-year olds. Readers are encouraged to review the report for greater detail regarding the methodology of this study, and the specifics of the results. The strong scientific basis for these conclusions is thoroughly documented in the report, reflecting a study the authors of Chapter 12 characterize as "soundly designed and implemented."

The methodology and conclusions presented in the first year report were the result of intensive, scientific analysis, including careful examination of the "state of the art" in the fields of early childhood development and policy research. For instance, the characterization of effect sizes was based on reviews of the body of evidence on early childhood education and developed within this broader context of findings. Our team followed the advice of experts in the field in interpreting the effect sizes in relation to findings from other similar early childhood intervention studies[69] and in characterizing the findings as "small to moderate" in size. However, the characterization of effect sizes is an issue of continued debate among experts in the field.

The study also carefully examined the possibility of potential bias introduced into the sample from the exclusion of saturated Head Start centers, a possibility to which the authors of Chapter 12 repeatedly refer. While some differences were found between the saturated and non-saturated sites (which were accounted for in the sampling weights), sensitivity analyses in the report show that these differences could not substantially bias the results. Readers should refer to Appendix 2.1 of the report for details of this approach and the findings.

[69]Gene V. Glass, Barry McGraw, and Mary Lee Smith, *Meta-Analysis in Social Research* (London: Sage, 1981); and Kathleen McCartney and Robert Rosenthal, "Effect Size, Practical Importance, and Social Policy for Children," *Child Development* 71 (2000): 173–180.

It is also important to note that the measures used in the study were carefully chosen, with input from a committee of experts in the field, to accurately assess all major aspects of children's school readiness and well-being. For example, letter and word recognition, the importance of which the authors of chapter 12 repeatedly question, are key predictors of children's basic reading skills as they enter the early school years and, as such, are central indicators of early literacy. Positive impacts were found for both the 3-and 4- year old cohorts for these constructs. While it is essential to take note of areas in which no significant impacts were found, such as oral comprehension and early math skills, the significance of impacts that were found need not be minimized.

Finally, and most importantly, this report presents impacts after less than a full year of Head Start. We are currently working on a final report that will provide a longer-term picture of how Head Start affects the children and families it serves. This report will allow us to answer, empirically, questions regarding whether impacts fade out and whether impacts occur on a broader array of measures (including teacher reports of children's behavior). Readers are encouraged to await findings from the final report for the full picture of how Head Start affects the children and families it serves.

Note: This report is open to public comments, subject to review by the forum moderator. To leave a comment, please send an email to welfareacademy@umd.edu or fill out the comment form at http://www.welfareacademy.org/pubs/early_education/chapter15.html.