SCHOOL of PUBLIC POLICY

14
Head Start Evaluation, Synthesis and Utilization Project

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/chapter14.html.

14
Head Start Evaluation, Synthesis and Utilization Project

The federal Head Start program, started in 1965, is designed to "break the cycle of poverty by providing preschool children from low-income families with comprehensive services to meet their emotional, social, health, nutritional, and psychological needs."[1] Head Start studies prior to 1980 varied widely in their findings prompting the U.S. Department of Health and Human Services to commission the Head Start Evaluation, Synthesis, and Utilization Project (the "Synthesis project") in 1981 to synthesize the findings.

Ruth McKey and her colleagues at CSR, Inc. (the "CSR team") conducted the study, integrating findings from 210 Head Start evaluation report to produce a comprehensive study of the effects of Head Start from 1965 to 1980.[2] The Synthesis project found strong initial cognitive and socioemotional gains from Head Start participation. These effects, however, faded by the end of the second year after Head Start participation. Many of the studies included in the assessment suffer from selection bias, small samples, and attrition. The CSR team attempted to address these problems in their analysis, but their conclusions did not change. Nevertheless, these limitations must be considered in examining their findings.

Program Design

Program group. The Synthesis project was limited to children (and their families) who participated in Head Start between its inception in 1965 through about 1980. Evaluations of other early childhood programs were purposely excluded.

Services. All studies included in the Synthesis evaluated programs that offered the standard Head Start services. There was, however, considerable variation in program operations over the period examined, including the organizations operating the programs and the curricula

---

[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]Ruth Hubbell McKey, Larry Condelli, Harriet Ganson, Barbara J. Barrett, Catherine McConkey, and Margaret C. Plantz, *The Impact of Head Start on Children, Families and Communities* (Washington, DC: CSR, Inc., June 1985).

used. Moreover, many of the evaluated programs operated before the Head Start Performance Standards set forth national goals and expectations for the programs in the mid-1970s.

The Evaluation. The CSR team collected 1,600 documents related to Head Start and identified those that examined the effects of Head Start. The final report synthesized findings from 210 evaluation reports.[3] Only seventy-six of these, however, included enough quantitative information to be analyzed using a statistical procedure known as "meta-analysis." (The remaining 134 studies were analyzed using "traditional narrative review methods.") Meta-analysis involves comparing outcomes across studies using a common metric, called an "effect size."[4] The CSR team considered effect sizes of 0.25 or greater to be "educationally meaningful," a term used by statisticians to indicate an effect that is observable in classroom performance.

The CSR team examined the impact of Head Start on children's cognitive and socioemotional development, their health, and their families. The number of studies within each category varied considerably: Seventy-two examined gains in cognitive ability, seventeen investigated changes in socioemotional functioning, and five assessed family impacts. Moreover, the number of studies used to assess a particular outcome or year might be considerably smaller.

The CSR team identified immediate effects (those occurring between four months before the program ended and six months afterwards), as well as those occurring one, two, and three or more years after Head Start participation.[5] For many outcomes, the researchers reported findings separately for (1) studies that compared Head Start children to children who did not attend Head Start (program/control) and (2) studies that compared the same children before and after Head Start participation (pre/post). In most cases, the program-control studies were based on comparison groups that were not formed by random assignment.[6]

Major Findings

[3]The excluded reports included descriptive studies, policy documents, and analyses of other early childhood interventions.

[4]An effect size is calculated by dividing the difference in mean scores between the program group and the control (or comparison) group by the standard deviation of the control (or comparison) group. Thus, only studies that provided sufficient information to calculate effect sizes, that is, means and standard deviations, were included in the quantitative analysis. For pre/post studies, the tendency for test scores to increase as children grow older meant that the traditional formula for calculating effect sizes would exaggerate Head Start's impact, so test norms were used to control for this tendency.

[5]"Three or more years" included periods anywhere from 28 to 168 months after Head Start participation.

[6]Treatment/control studies typically compared Head Start children to comparable, but not randomly assigned, children who did not participate in Head Start. The pre/post studies involved comparisons of the same children before and after their participation in Head Start.

The Synthesis project found strong initial cognitive and socioemotional gains from Head Start participation. These effects, however, faded by the end of the second year after Head Start.[7] There were also positive findings with respect to school readiness, socioemotional development, and children's health, but the effects were either small or subject to uncertainty due to limitations surrounding the research.

Cognitive. The CSR team reported that Head Start produces significant, immediate gains in cognitive test scores (see table 1). The average weighted effect size was 0.59 for program/control studies and 0.43 for pre/post studies, both well above the 0.25 criterion considered "educationally meaningful." Similar findings were reported for "readiness" and "achievement" measures. Regardless of the measure, these gains diminished rapidly among treatment/control studies and were no longer educationally meaningful by the end of the second year. Long-term effects for pre/post studies did not diminish after the first year, but the CSR team did not present longer-term follow-up because only one of the evaluations had data beyond the first year. They explain that the small increase in effect sizes in the first year may have been "due to insufficient control for maturation in the method used for computing effect sizes in pre/post studies."[8]

Table 1. Synthesis Project: Cognitive Development

Period after Head Start participation	Effect sizes for:					
	Treatment/control studies			Pre/post studies		
	IQ	Readiness	Achievement	IQ	Readiness	Achievement
Immediate	.59	.31	.54	.43	.59	.37
1st year	.09	.21	.20	.65	.69	NA
2nd year	-.03	.02	.13	NA	NA	NA
3rd year	-.20	NA	0	NA	NA	NA

Source: Ruth Hubbell McKey, Larry Condelli, Harriet Ganson, Barbara J. Barrett, Catherine McConkey, and Margaret C. Plantz, *The Impact of Head Start on Children, Families and Communities* (Washington, DC: CSR, Inc., June 1985), III-9–III-13.
Note: "NA" indicates that the data were not available, because only one study using the pre/post design examined effects beyond the first year after Head Start participation had ended.

[7]McKey et al., 1985.

[8]McKey et al., 1985, III-14.

School readiness/performance. Head Start children were found to perform better on various measures of school performance. They were less likely to be retained a grade or to be placed in special education classes. The CSR team, however, cautioned that these findings were based on just three studies.

Socioemotional development. The findings with respect to socioemotional development were mixed, depending on the measure used (see table 2). There were immediate effects for three measures: self-esteem (0.17 SD), achievement/motivation (0.22 SD), and social behavior (0.35 SD). Only the social behavior measure, however, was considered educationally meaningful. In addition, it was the only effect that persisted beyond the immediate postintervention period, and it too disappeared after the second year.

Health. The CSR team also looked at health data, but determined that most of the studies examining health effects were qualitative and could not be used for the meta-analysis. Upon examining the available evidence, they concluded that Head Start children were more likely to receive a range of medical services, which "appeared" to improve their general health. In particular, one random assignment study conducted by Abt Associates Inc. and evaluated by the Synthesis project found that Head Start children showed a decrease in pediatric problems when compared to non-Head Start children, and showed improved motor coordination, nutritional intake, and dental health.

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.

Table 2. Synthesis Project: Findings on Socioemotional Development

	Effect sizes for:		
Period	**Self-esteem**	**Achievement/ motivation**	**Social behavior**
Immediate	.17	.22	.35
1st year	-.20	-.11	.16
2nd year	.01	.06	.63
3rd year	-.14	.08	-.10

Source: Ruth Hubbell McKey, Larry Condelli, Harriet Ganson, Barbara J. Barrett, Catherine McConkey, and Margaret C. Plantz, *The Impact of Head Start on Children, Families and Communities* (Washington, DC: CSR, Inc., June 1985), III-9–III-13.

Effects on parents. There were too few studies with quantitative effects on most parental outcomes to conduct a meta-analysis, so the analysis of family effects was based on a narrative review of Head Start studies. The CSR team reported that parents valued the Head Start experience, but concluded that many other effects were uncertain, such as whether parents changed their child-rearing practices or whether Head Start changed parents' attitudes toward their own lives.

Benefit-cost findings. Apparently a benefit-cost analysis was not performed. In 1984, Head Start cost approximately $4,400 per child (in 2005 dollars).

Overall Assessment

Many of the studies included in the assessment suffer from selection bias, small samples, and attrition. The CSR team attempted to address these problems in their analysis, but their conclusions did not change. Nevertheless, these limitations must be considered in examining the findings.

Program theory. The CSR team describe the developmental approach guiding the Head Start program with a quote from the Head Start Performance Standards:

> The overall goal of Head Start is to enhance the social competence of children from low-income families. By social competence is meant . . . the child's everyday effectiveness in dealing with both present environment and later responsibilities in school and life. Social competence takes into account the interrelatedness of cognitive and intellectual development, physical and mental health, nutritional needs, and other factors that enable a

developmental approach to helping children achieve social competence.[9]

The studies included in the Synthesis project explore Head Start's impact on cognitive development, socioemotional development, and child health, as well as its impact on families and communities, which is appropriate within this context.

Program implementation. The studies included in the meta-analysis cover different periods of Head Start operations, beginning in the 1960s through the early 1980s. Thus, there is considerable variation in program operations over the period examined, including the organizations operating the programs and the curricula used. In addition, many evaluations cover programs that were operating before HHS implemented Head Start Performance Standards in the mid-1970s. Although the CSR team did not have sufficient information to assess "the content of the programs or to evaluate their quality," they examined findings to see if they varied for different time periods.[10] They found some evidence that the quality of programs improved throughout the 1970s, resulting in slightly larger short-term impacts. They conclude, however, that the "changes did not produce test scores that lasted beyond two years."[11]

Assessing the randomization. Two types of studies were included in the Synthesis Project: (1) Pre/post studies, which compare the performance of the same group of children before and after their Head Start involvement; and (2) Treatment/control studies, which compare a group of children with Head Start experience (the treatment group) to a group of children without Head Start experience (the control group).

Among the treatment/control group studies, the evaluators examined the degree to which the treatment and control groups were comparable, although, they did not indicate whether the groups were actually randomized. They found six studies in which the two groups were comparable, six studies in which they were not comparable and the controls were of higher socio-economic status than the Head Start group, three studies in which the treatment and controls were from the same neighborhood, eight studies in which the controls were also Head Start-eligible, two studies in which comparable and non-comparable children were mixed, and seven studies in which comparability could not be determined.

The evaluators then grouped the studies into three categories—comparable, not comparable, and unknown—and analyzed them to determine if systematic differences existed

---

[9]U.S. Department of Health and Human Services, Administration from Children, Youth, and Families, Head Start Program Performance Standards (Washington, DC: HHS, 1975), 1.

[10]Improvements in statistical models over the years also affected the estimates, making it more difficult to isolate the impact of the program during different time periods.

[11]McKey et al., 1985, E-6.

among these categories. While the studies that had comparable groups showed higher effects at one and two years after Head Start participation, these effects disappeared at three years after the program ended. Thus, the evaluators conclude, "it does not appear that non-equivalence of control groups is a serious problem in the analysis of long-term effects."[12]

Assessing statistical controls in experimental and nonexperimental evaluations. There are two types of selection bias to consider in the Synthesis study: selection bias within individual studies and "publication bias" among the studies selected for inclusion in the analysis.

Selection bias within individual studies. The studies included in the Synthesis project included a mix of experimental and quasi-experimental studies, some of which may have suffered from selection bias problems serious enough to bias the entire study. As Richard Berk, professor of Criminology and Statistics at the University of Pennsylvania, and Peter Rossi, former professor at the University of Massachusetts (Amherst), caution:

> Although meta-analysis can be a very useful tool and certainly has its champions, our assessment is rather cautious. First, everything depends on the quality of underlying studies. If they have weak validity overall, even the fanciest of meta-analyses cannot save the day. Meta-analysis cannot correct for fundamental flaws in the original research.[13]

The CSR team examined the comparability of the program and comparison groups and, where possible, identified those that were well-matched and those that were not. They then presented the findings for each separately. Even with the better matched groups, however, there may be unobservable differences, so selection bias remains a concern.

"*Publication bias.*" The second form of selection bias surrounds the way that the studies were selected. All the programs selected had been the subject of published studies. Journals favor studies that show significant results, yet this preference may present a one-sided view of the evidence, commonly referred to as "publication bias" or the "file-drawer problem." For every study with significant results that gets published, there may be many more with insignificant results that languish in file drawers, unpublished. In addition, some studies with disappointing findings early on may have been rejected or abandoned, so that long-term follow-up was not possible. For example, given the disappointing findings of the more recently completed Comprehensive Child Development Program[14] (see chapter 3), it is unlikely that any long-term

[12]McKey et al. 1985, II-15.

[13]Richard A. Berk and Peter H. Rossi, *Thinking About Program Evaluation 2* (Thousand Oaks, CA: Sage Publications, 1999), 105.

[14]Robert G. St.Pierre, Jean I. Layzer, Barbara D. Goodson, and Lawrence S. Bernstein, *National Impact Evaluation of the Comprehensive Child Development Program: Final Report* (Cambridge, MA: Abt Associates

follow-up would be considered.

Meta-analyses of the effect of early childhood programs may be particularly vulnerable to the file-drawer phenomenon, because there are relatively few published studies and few of these reveal large effects. According to the science writer Morton Hunt, "The file-drawer problem poses a particular threat to a small meta-analysis; the smaller it is, the greater the chance that its conclusions, even if very strong, could be weakened or voided if a body of nonsignificant results were to come to light."[15]

Sample size. There are two issues related to sample size: the sample size within individual studies and the number of studies that examined a specific outcome.

The size of the samples varied considerably among individual studies. The CSR team notes that:

> The number of children in a study is of concern because meta-analysis generally makes no allowance for sample size. Therefore, an effect size based on the performance of a group of fifty children is treated equally to that of a group of five children even though the scores comprising the former effect size would be much more stable than the latter.[16]

When the CSR team grouped the studies by sample size, however, they found little difference in the magnitude and pattern of effects across studies.

For the meta-analysis itself, the relevant sample size is the number of studies examined for each outcome. While the number of studies used to assess cognitive gains was fairly large, the number used for other outcomes was considerably smaller and often too small to assess some effects at all.

Attrition. The CSR team examined the studies to assess their dropout rates and whether the characteristics of the children who dropped out were different from those who remained. Although they found relatively low attrition rates, most studies did not provide sufficient information to assess the comparability of the groups over time. After grouping the studies into those in which attrition posed a threat to validity and those in which it did not, the CSR team found little difference in the magnitude and patterns of effect sizes.

Inc., June 1997).

[15]Morton M. Hunt, *How Science Takes Stock: The Story of Meta-Analysis* (New York: Russell Sage Foundation, 1997), 119.

[16]McKey et al., 1985, III-8.

The large drop-off in the number of studies included in the analysis—as the follow-up period lengthened—posed another attrition-related problem. For example, the number of studies with IQ findings fell from forty-one immediately after program participation, to twelve one year later, to six three years later. The CSR team cautioned that, "Findings from these [later] studies are less stable than those focusing on earlier years since the later studies are subject to idiosyncracies in the few studies upon which effect sizes are based."[17]

Data collection. The CSR team collected more than 1,600 Head Start related documents through on-line searches of bibliographies, written requests to grantees, and contacts with government personnel and private researchers. Of these, 210 reported the results of research on Head Start and formed the data base for the synthesis.[18] For 176 of the studies, traditional narrative reviews were used to synthesize the information. For the remaining thirty-four studies, enough information was on hand to allow for the use of meta-analysis. The data sources were appropriate for the questions be studied.

Measurement issues. Many of the cognitive outcomes were measured using nationally recognized IQ or achievement tests. The CSR team noted, however, "It is difficult to measure socioemotional development, and instruments assessing this domain are generally not as refined, valid, or reliable as those for evaluating cognitive development."[19] In particular, many of the measures relied on a child's self-report, which raises concerns about validity due to children's limited verbal skills or desires to provide "socially desirable" responses. In addition, there were limited standardization procedures, creating a particularly serious problem for the pre/post studies.

Generalizability. The evaluations included in the project cover many geographic areas and a wide span of Head Start's early operational period, but they do not constitute a nationally representative sample of Head Start sites. Nevertheless, the findings could be considered suggestive of the impact of Head Start nationally for the period of time covered by the studies.

Replication. There have been no subsequent meta-analyses on the scale of the Synthesis project.

Evaluator's description of findings. The CSR team concluded that Head Start produced immediate cognitive gains that diminished by the end of the second year after Head Start participation. They reached this conclusion by comparing effect sizes immediately after Head Start participation and then again one, two, and three years later. But, the number of studies included

[17]McKey et al., 1985, II-13.

[18]McKey et al., 1985, 4.

[19]McKey et al., 1985, IV-1.

for each time period differed. Thus, it is possible that part of the decline in effect sizes results not from a diminution of Head Start's effects, but from a difference in the composition of studies examined.

In any event, the CSR team concludes that the apparent "fade out" of Head Start effects indicates that "even more program improvements are warranted."[20]

Evaluator's independence. CSR, Inc., was selected through a competitive procurement process to act as the independent contractor for this study.

Statistical significance/confidence intervals. The CSR team did not conduct tests of statistical significance. They explain, "Our analysis is descriptive . . . Thus, it is not necessary to draw statistical inferences."[21]

Effect sizes. Effect sizes were calculated for each study in the meta-analysis, and then effect sizes measuring the same type of impact were grouped together and averaged to determine the average impact of Head Start on that aspect of development.[22]

At the conclusion of Head Start participation, cognitive effect sizes were in the range of 0.3 to 0.6 SD. One year after Head Start participation had ended, effect sizes fell to between 0.1 and 0.2 SD among the treatment/control studies. Two years after Head Start participation, effect sizes fell further, to between 0 and 0.1 SD. Three or more years after Head Start participation had ended, very few studies showed any significant effects.

The CSR team deemed an effect size of 0.25 or greater to be "educationally meaningful," explaining, "Educators and researchers in early childhood education commonly consider an effect size in the range of 0.25 or greater (either positive or negative) to be educationally meaningful."[23] Thus, "educationally meaningful" cognitive effects were found only at the immediate conclusion of Head Start participation, and not thereafter. (See Appendix 1 for a further discussion of effect sizes and their interpretation.)

Sustained effects. A few of the studies included in the meta-analysis examined the long-term effects of Head Start (one to three or more years after the program ended). An analysis of the immediate and long-term effects among the treatment/control studies suggests that Head Start

[20]McKey et al., 1985, 23.

[21]McKey et al., 1985, II-9.

[22]McKey et al., 1985, 6.

[23]McKey et al., 1985, 5.

does not have a significant impact when measured three or more years after Head Start participation.[24]

Benefit-cost analysis. Apparently not performed.

Cost-effectiveness analysis. Apparently not performed.

[24]McKey et al., 1985, III-15.

Commentary

Editor's Note: For each evaluation included in this report, we attempted to contact the senior evaluators to offer them the opportunity to respond to our assessment. Dr. Ruth Hubbell McKey said that she was willing to provide comments, but she has not done so.

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/chapter14.html.