SCHOOL of PUBLIC POLICY

5
Consortium For Longitudinal Studies

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/chapter5.html.

5
Consortium For Longitudinal Studies

The *Report from the Consortium of Longitudinal Studies* ("the Consortium Study") was a collaborative effort by eleven research groups with longitudinal studies of early childhood interventions that operated between 1962 and 1972. Only eight of the eleven projects were used for most of the analyses described here: the Early Training Project, the Experimental Variation of Head Start Curricula, the Harlem Training Project, the Mother-Child Home Program, the New Haven Follow-Through program, the Parent Education Program, the High/Scope Perry Preschool Project, and the Philadelphia Project.[1] (A ninth study, the Curriculum Comparison Study, was used in the analysis of a limited number of outcomes.) The evaluation sought to determine the long-term effects of selected infant and preschool intervention programs on child participants. The data for the final follow-up were collected through a youth survey, a parent survey, school records, and an IQ test.

Irving Lazar, then of Cornell University, headed the independent analytic team (the "Consortium team") that coordinated the collaboration. The Consortium Study statistically combined findings from selected early childhood education programs that had conducted long-term follow-ups. The studies included in the analysis had relatively rigorous designs, but most had small samples. By assessing the studies as a group, the Consortium team was able to increase statistical power. As a result, they often found statistically significant findings, even when many of the studies individually did not demonstrate such effects.

The Consortium Study has often been cited as evidence that early childhood intervention programs can improve cognitive outcomes and school performance. Its findings were largely consistent with other literature reviews, including the findings that initial IQ gains "fade out" and that early intervention leads to improvements in school performance. Despite the rigor of the study, some methodological issues remain: the small number of studies examined, relatively high rates of attrition, and potential selection bias. Moreover—although this concern was less important when the study was published—the projects themselves operated over forty years ago in a very different societal environment, which may limit their current applicability.

---

[1]For the findings reviewed here, three projects were excluded from the Consortium's analysis. The Institute for Development Studies project was excluded because of difficulties in obtaining school record data. The Curriculum Comparison Study (for most outcomes) and the Micro-Social Learning System programs were excluded due to the absence of a suitable control group.

Program Design

Program group. Of the eight projects primarily examined, four enrolled children when they were between ages four and five, three enrolled children between ages two and three, and one enrolled them as early as three months old.

The demographic characteristics of the children and families were similar across the projects. In the median project, 95 percent of the children were black, the children averaged 3.2 siblings and scored an average of 92 points on the Stanford-Binet IQ test.[2] The mothers had completed an average of 10.4 years of schooling and scored a 64.0 on the Hollingshead Two-Factor Index of Social Position. There were, however, some important differences among sites (for example, a small town in Tennessee vs. the Harlem neighborhood in New York City). In some cases, the eligibility criteria of the project itself affected the group's characteristics. For example, children in the High/Scope Perry Preschool Project had significantly lower IQs than children in the other programs (79 vs. 92), because participation in the program was restricted to children with IQs below 85.

Services. The programs studied were diverse (see table 1). Four were center-based programs that provided a nursery school-like preschool environment, using a variety of curricula; two were home-based programs, designed to change the behavior of the parent (typically the mother); and three offered a combination of both approaches. Most programs lasted just one or two years, but one project lasted up to four years (although most of the intervention period covered the school-age period).

The Evaluation. For its evaluation, the Consortium team used a modified version of a technique called "meta-analysis," a statistical method that combines or pools findings from a number of different individual studies. A meta-analysis involves several steps. First, the purpose of the analysis and the questions to be addressed are determined. Second, the evaluations that address the purpose are identified. Third, the data from each evaluation are collected and coded. This includes information on the outcomes to be examined, as well as the characteristics of the evaluations and programs themselves. Fourth, the outcomes are transformed into a common metric—an effect size—so that they can be compared across evaluations. The Consortium Study used meta-analysis to analyze the original raw data from the studies, but also collected a common set of data from participants during two, later follow-up periods.

The Consortium team notes that an important advantage of meta-analysis is that it can compensate for the shortcomings of a single study:

---

[2]This summary is based on all eleven studies and reflect the baseline characteristics of the 1976 follow-up sample. See Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 7.

> Every research and demonstration project has design problems that can threaten the validity of its findings to some degree. In the study of a single project one can rarely ascertain the extent to which design problems affect the results. In contrast, in a series of independently designed studies, no single design problem would be likely to affect all results.[3]

This reasoning assumes either that few studies share a particular design problem or that problems will cancel each other out. In some cases, however, problems across studies may generate biases in the same direction. For example, most projects were implemented as model programs and had special funding. It is not clear that the same benefits could be achieved if the programs were implemented on a large scale. Or, when combining experimental and quasi-experimental studies, if the former produced unbiased estimates but the latter produced biased estimates, it may not be the case that the inclusion of both types of studies would cancel out the bias inherent in the weaker studies. Indeed, studies with stronger research designs consistently generate smaller effect sizes than studies with weak research designs.[4]

The Consortium team used the following criteria for a study's inclusion in the meta-analysis: (1) completion of the programmatic aspect of the intervention before 1969 (so that it would be possible to examine long-term outcomes); (2) a large enough initial sample (over 100 subjects), so that even with an attrition rate of 65 percent, the follow-up sample would be large enough for analysis; and (3) a rigorous evaluation methodology, based on either random assignment or a carefully matched comparison group.

The Consortium Study collected data from four periods: (1) at program entry (between 1962 and 1972, when the children ranged in age from birth to five years); (2) shortly after program termination; (3) in 1976–1977, when the children were between ten and nineteen years old, depending on the project; and (4) 1979–1981, when the children were between thirteen and twenty-one years old, depending on the project.[5] The first two phases of data collection were independently conducted by each project investigator, but all of the original data were sent to Cornell, where an independent research team checked for accuracy and internal consistency and, if necessary, corrected the data. The data for the third follow-up were obtained from a youth

[3]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 8.

[4]*See*, for example, Mark W. Lipsey, "Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects," in *Meta-Analysis for Explanation: A Casebook* (New York: Russell Sage Foundation, 1992), 119.

[5]The age range at final follow-up for all eleven projects was from ten to twenty-two.

survey,[6] a parental survey,[7] school records (special education and grade retention), and the Wechsler Intelligence Scale for Children-Revised (WISC-R) IQ test. The data for the final follow-up were obtained from a youth survey[8] and school records. A common protocol was used, with follow-up data collection on as many of the original sample members as possible. (In addition, several of the investigators collected additional data and published independent analyses of their own.)[9]

The Consortium team divided the studies into two categories: the "more nearly randomized" (or close to experimental) and the "less randomized" (or quasi-experimental).[10] The Consortium team's analysis of effects included all studies, but it also examined the impacts separately, based on the strength of the research design, in order to determine whether such a categorization affected their conclusions.

Data from the studies were pooled, as is standard for a meta-analysis. The Consortium team measured the impact of early childhood interventions on various school performance and cognitive outcomes in two ways. First, they compared the mean outcomes for program and control (or comparison) groups, with appropriate tests of statistical significance, to see if there were differences between the groups. They then used multiple regression to control for background characteristics "measured at entry to each project" for both the child (for example, sex and pre-test IQ) and the family (for example, number of siblings, family structure, father's presence or absence, and maternal education). As Lazar and his colleague, Richard Darlington, also at Cornell University, describe:

> Each technique compensated for a disadvantage of the other. The simpler techniques (e.g., cross-tabulation, *t* tests) were used because multiple regression often reduced

---

[6]The youth data included information on "the child's status in school, his educational and occupational aspirations, leisure time activities and interests, employment status, and integration into his peer group and the larger community."

[7]The parent data included information on "household composition, socio-economic status, parental attitudes toward, aspirations for, and evaluations of their child, information on the child's medical history, school educational history, the parent's current relationship with the child, and parental assessment of the intervention program."

[8]The 1980 youth survey focused on post-high school educational, employment, and family-related outcomes.

[9]In some cases, the findings of the Consortium team differed from those of the original investigators due to different samples and analytic techniques.

[10]The "more nearly randomized" projects included: the Early Training Project; the Harlem Training Project; the Parent Education Program; and the Perry Preschool Project. The "less randomized" projects included the Experimental Variation of Head Start Curricula; the Mother-Child Home Program; the New Haven Follow-Through Study; and the Philadelphia Project.

> sample size (and therefore the power) due to missing data. Multiple regression, on the other hand, helped to ensure that any program/control differences on outcome measures were not accounted for by the slight differences on background characteristics which could exist even with an experimental design. Thus, achieving the same results with both techniques would prove far more convincing than achieving a result from one technique alone."[11]

To ensure that no single study was responsible for the pooled result, the Consortium team recalculated effects, excluding the project with the strongest result. If a significant result remained, the finding was considered "robust."[12]

[11]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 25.

[12]In a meta-analysis, researchers typically transform findings into effect sizes and then compare them across the studies. The Consortium study appears to take a median finding and a pooled-z score, which is the sum of the t-statistics divided by the square root of the number of studies. This calculation is an indicator of statistical significance, but the use of a median experimental-control difference in lieu of an effect size is somewhat unusual. One reason for this approach may be because the study was done when meta-analytic techniques were just being developed. The practical significance of this difference is uncertain.

Table 1. Consortium Studies: Program and Evaluation Summary

Program	Description	Age of entry	Duration	Research design/sample size (initial/follow-up)
Early Training Project	Center-based half-day summer program, with home visits during the school year	4 years	1–2 years	Random assignment (n = 92/77)
Experimental Variation of Head Start Curricula	Center-based, compared different preschool programs	4 years	1 year	Comparison group (n = 271/141)
Harlem Training Project	Center-based program, with one-on-one tutoring and child-directed play	2-3 years	1–2 years	Comparison group (n = 309/228)
Mother-Child Home Program	Home visits twice weekly focused on improving mother-child verbal interaction	2 or 3 years	1–2 years	Comparison group (n = 250/186)
New Haven Follow-Through	Center-based	5 years	4 years	Comparison group (n = 156/144)
Parent Education Program	Home visits weekly by paraprofessionals	3-24 months	1–3 years	Random assignment (n = 309/107)
High/Scope Perry Preschool Project	Center-based half-day preschool combined with home visits	3 or 4 years	1–2 years	Random assignment (n = 123/123)
Philadelphia Project	Center-based half-day nursery school, 4 days per week, with weekly home visits	4 years	1 year	Comparison group (n = 170/126)
Total	Three center-based, two home-based, and three mixed			Four random assignment and four comparison group (n=3,593/2,008)

Note: This table excludes three Consortium projects: The Institute for Development Studies (because of difficulties in obtaining school record data) and the Curriculum Comparison Study and the Micro-Social Learning System programs (because of the absence of a suitable control group).

Major Findings[13]

The Consortium Study has often been cited as evidence that early childhood intervention programs can improve cognitive outcomes and school performance. Its findings were largely consistent with other literature reviews, including the findings that initial IQ gains "fade out" and that early intervention leads to improvements in school performance.

Cognitive. The Consortium Study included both IQ effects (which the Consortium team concluded were not permanent) and some statistically significant findings on achievement.

IQ. The programs in the Consortium analyses administered IQ tests prior to the children's entry into the program, shortly after program completion, and in 1976, when the children were ages nine to nineteen. The Stanford-Binet and Peabody Picture Vocabulary Test (PPVT) were used for the early measurements and the WISC-R was used for the final follow-up in 1976. The analysts used multiple regression to control for children's background characteristics, including sex, pre-program IQ scores, number of siblings, maternal education, and whether their father lived in the home.[14]

The pooled IQ results from seven projects (four with "more nearly randomized designs") were statistically significant and robust at the time the children completed the program, with a median gain of over 7 points. These results remained significant and robust for two years after program participation, although their magnitude diminished to a median gain of about 4 points. At three and four years after program participation, the pooled results remained statistically significant, with a median gain of 3 points, but were no longer robust.

The Consortium team then examined IQ scores in 1976, but analyzed each project separately, because the children differed considerably in age and years since program completion across the projects. They found only one statistically significant gain in the six projects they examined, leading them to conclude that, "the effect of early education on intelligence test scores was not permanent."[15]

---

[13]All findings are from Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 1–151; or Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 411–459.

[14]The analyses for most outcomes reported simple program and control (or comparison) group differences and then presented the findings controlling for background differences. Only the latter approach was used for the IQ findings, because the two methods produced very similar results.

[15]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 47. The only study to report a gain was the Mother-Child Home Program. The evaluation was based on a comparison

Achievement. The Consortium team examined children's test scores for each grade between the third and sixth grade in seven projects (four with "more nearly randomized designs"). Because different school systems administered these tests, the tests used and their frequency varied from project to project.[16] The Consortium team controlled for children's sex, pre-program IQ scores, and age (when the relevant data were available).

Program impacts were larger in the earlier grades and for math subtests than reading subtests. For example, in third grade, program children did significantly better on both math and reading, although only the math finding was robust. In the fourth and fifth grades, only the math result was significant, but it was not robust. By the sixth grade, the findings were no longer statistically significant for either subtest.

A number of qualifications related to the data and attrition analyses should be mentioned. For example, because the amount of data at each grade level varied by project, the same children were not compared in all of the grades. Thus, the Consortium team cautions:

> Differences in results across grades may be more a function of which projects were in the analysis than of a genuine temporal effect. Second, in cases where a project had data for several grades, there were often attrition effects in one grade but not in another—thus biasing any between-grade comparison within a project. In sum, the data were sufficient to indicate that, in general, program graduates tended to do somewhat better than their controls but were not sufficient to draw conclusions about effects across time.[17]

School readiness/performance. The Consortium team evaluated whether students met school standards of adequate performance using the incidence of special education placement and grade retention as measures. Recognizing that schools often emphasized one of these responses to inadequate performance rather than the other, they also included a composite measure: *either* placement in special education or grade retention.

Special education. Lazar and Darlington examined six projects (three with "more nearly randomized designs") in the analyses of assignment to special education classes. The findings were based on children's outcomes when they were between the third grade and the twelfth

---

group design and subject to selection bias problems, so even this effect is uncertain. Moreover, the comparison group used for the long-term follow-up was not the same as the one used for the program's earlier IQ findings, but rather one that was selected about five years after program participation. The Consortium report does not explain why the program's other control group was not administered the WISC-R for the final follow-up.

[16]The tests administered included the California Achievement Test (CAT), Metropolitan Achievement Test (MAT), Stanford Achievement Test (SAT), and Peabody Individual Achievement Test (PIAT).

[17]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 42–43.

grade, depending on the study. Program children were considerably less likely to be placed in special education, with a median rate of assignment to special education of 14 percent compared to 29 percent for the control (or comparison) group.[18] The finding was statistically significant and robust, and was even stronger for the "more nearly randomized designs."[19]

The Consortium team extended the consortium analysis to a seventh project and focused on findings when the children were in the seventh and twelfth grades. (The latter analysis was based on four projects.) Program children were less likely to be placed in special education, with an average rate of placement of 15 percent, compared with 35 percent for the control (or comparison) group.[20] This finding was statistically significant and robust.[21] In the twelfth grade, the average rate of special education placement remained lower, 13 percent compared to 31 percent, and was statistically significant.[22] (The twelfth grade finding is not directly comparable to the seventh grade finding, because is based on just four projects.)

Grade retention. The findings for grade retention were "less striking," in the words of the Consortium team, than those for special education placement. Eight projects (four with "more nearly randomized designs") were included in this analysis. The median rate of grade retention was 25 percent for the program group and 31 percent in the control (or comparison) group. Although seven of the eight projects showed reductions in grade retention among program group children, only one of the differences was statistically significant. The pooled results were statistically significant, but were not robust.[23] The findings were somewhat stronger among the four nearly randomized projects, where the median grade retention rate was 26 percent for the program group compared to 37 percent for the control (or comparison) group. Although these

[18]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 32.

[19]This finding held up after the researchers controlled for the background characteristics of the children and their families.

[20]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 433.

[21]This finding held up after the researchers controlled for the background characteristics of the children and their families.

[22]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 436.

[23]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 34.

findings were statistically significant, they were not robust.[24]

The Consortium team found a somewhat larger differential when the children were in the seventh grade, where the median grade retention was 20 percent for the program group compared to 32 percent for the control (or comparison) group.[25] The finding was statistically significant and robust. After controlling for background characteristics, it remained statistically significant, but was no longer robust. By twelfth grade, the differences were no longer statistically significant. (The twelfth grade findings are not directly comparable to the seventh grade findings, because they are based on just four projects.)

None of the projects had significant effects on both special education placement and grade retention, however. A possible reason may be that schools differ in their reliance on one or the other option when children are not meeting minimal requirements.

Percent who failed to meet school requirements. The Consortium team also examined a composite variable for inadequate performance that combined the special education and grade retention measures. Across eight projects, 44 percent of control (or comparison) group children failed to meet school requirements, compared to just 25 percent of the program children. The findings were statistically significant and "very robust." Consistent with the findings for special education placement and grade retention, these results were stronger for the "more nearly randomized research" projects.[26]

Royce and her colleagues found similar effects when the children were in the seventh grade, with 30 percent of program group children failing to meet school requirements, compared to 45 percent of control (or comparison) group children.[27] The finding was statistically significant and robust. In the twelfth grade, the differential was 44 percent compared to 62 percent, a finding

[24]After controlling for background differences, the findings "were in the same direction but weaker." Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 36.

[25]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 433.

[26]After controlling for background characteristics, the results remained statistically significant and were no longer robust (but were "close to robust").

[27]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 434.

that was statistically significant.[28] (The twelfth grade findings are not directly comparable to the seventh grade findings, because they are based on just four projects.)

High school graduation. High school graduation is another school performance indicator. Royce and her colleagues reported that, across four projects, the program group was more likely to have completed high school than the control (or comparison) group, 65 percent compared with 52 percent.[29] Although the difference was statistically significant for only one project, the pooled results were also significant.[30]

Socioemotional development. Relevant tests apparently not administered or results not reported.

Health. Data apparently either not collected or not reported.

Behavior. Data apparently either not collected or not reported.

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

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

Economic outcomes. Royce and her colleagues examined impacts for a range of labor market variables in three projects, including the labor force participation rate, the unemployment rate, or the employment ratio, hours worked, and earnings. They found no statistically significant impacts.[31] They also found no effect on the receipt of public assistance. (The program group members were between nineteen and twenty-two years of age at the time of the follow-up.)

Effects on parents. Data apparently either not collected or not reported.

---

[28]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 439.

[29]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 440.

[30]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 440.

[31]Jacqueline M. Royce, Richard B. Darlington, and Harry W. Murray, "Pooled Analyses: Findings Across Studies," *As the Twig is Bent . . . Lasting Effects of Preschool Programs* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1983), 443.

Benefit-cost findings. The Consortium Study did not investigate the benefit-cost ratio directly.

Overall Assessment

Despite the rigor of the study, some major methodological issues remain: the small number of studies examined, relatively high rates of attrition, and potential selection bias among the children in the program and among the studies included in the analysis. Moreover, although this concern had less force when the study was published— the projects themselves operated over thirty-five years ago in a very different environment. For example, fewer children were in child care and fewer mothers were employed.

Program theory. Apparently, there is no specific theory detailed beside the general expectation that early intervention programs promote school readiness and improve developmental outcomes for children.

Program implementation. The Consortium team presents a "matrix of program descriptions," which includes variables such as the adult-child ratio, staff qualifications, and degree of structure in teaching activities. But implementation issues are not discussed in detail. In addition, the programs differed in a number of ways, including age of children, program duration, parental involvement, staff training, curriculum models, and delivery system. As a result, even with good implementation, it would be difficult to isolate the more effective service strategies and the program groups that seem to benefit the most from the intervention.

Assessing the randomization. Four of the eight studies in the Consortium Study of school performance and cognitive outcomes were described as "more nearly randomized." (The ninth study, included in some analyses, was not randomized.) As noted above, however, one of these, the Harlem Training Project, should more properly have been considered a quasi-experiment. Of the remaining three, the Consortium Study briefly describes the procedures for random assignment and possible problems with them. For example, they describe the randomization in the Perry Preschool Project as follows:

> Assignment of treatment and control groups was essentially random, with the exception that in cases where a child assigned to the treatment group could not attend due to lack of transportation or maternal employment (preventing scheduling of home visits), the child was exchanged with a matched child assigned to the control group. This occurred approximately once in each of the five waves. . . . This study has been classified with the more nearly randomized designs as this one exception was deemed a relatively minor departure from experimental design.[32]

[32]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 75.

This is how they describe the assignment process for the Parent Education Program:

> Three waves of children were involved, and the assignment procedures varied among waves. All three waves were randomly assigned to treatment or control groups, but the assignment in one wave was not on an individual basis.[33]

The study indicates that the final wave was assigned on a group basis. Although the Consortium team considered the process to be "essentially random," they presented no evidence to this effect. The final "more nearly randomized" project, the Early Training Project, was classified as having a "strong" experimental design. Our assessment suggests a more ambiguous situation (see chapter 7 in this volume).

Assessing statistical controls in experimental and nonexperimental evaluations. An assessment of the Consortium Study must consider two types of selection bias: bias within individual studies and the so-called "file-drawer problem" or "publication bias."

Selection bias within individual studies. The studies included in the Consortium team's analyses 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.[34]

The Consortium team characterized four of the eight projects in their analysis as having "less randomized designs." (A ninth study, included in the analysis by Royce and her colleagues, was also based on a comparison group methodology.) Although the projects made efforts to match program and comparison group members, and some applied statistical means to control for measured differences, the groups may nevertheless have differed in some unmeasured ways. For example, in the Philadelphia Project, parents were invited to apply to have children enroll in nursery school. Later, children from kindergarten who had not been in nursery school were selected on the basis of age, sex, and ethnicity. The groups appeared to be similar, based on pre-test IQ and ten demographic variables. Nonetheless, the volunteer group may have differed from

[33]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 69.

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

the other group on unmeasured dimensions, such as motivation. This might result in misstating or misrepresented the impact of the intervention.

Some of the other "less randomized" projects reported significant baseline differences. For example, the matched comparison group in the Mother-Child Home Program was of a higher socioeconomic status, and the comparison group in the Experimental Variation of Head Start Curricula project was more likely to consist of white families with fathers present. These differences suggest that children in the comparison group were more advantaged, perhaps biasing the program effects downward. The Consortium team attempted to control for these differences, but there were probably other differences that were not be controlled for, such as parent interest in a child's development. Thus, selection bias remains a concern in the four quasi-experimental studies. (Sophisticated statistical models that could attempt to deal with such unobserved differences, for example, instrumental variables or fixed effects models, were not available at the time of the study.)

The Consortium team dealt with the issue of bias by analyzing results from just the "more nearly randomized" designs. In general, these findings tended to support the overall analysis. However, even with the "more nearly randomized studies," there were differences among the program and control groups that raise some concerns. In particular, the categorization of one of the four projects, the Harlem Training Project, as a "more nearly randomized design" is questionable. The Consortium team describes the assignment procedure for that project as follows:

> Children born in the months of August-October 1964 were randomly assigned to a particular treatment group. Children born in November and December 1964 were recruited specifically as controls. It is likely that this selection procedure did not introduce serious bias since the project staff emphasized the benefits of a total of 4.5 weeks of testing in recruiting the controls. One could thus see the control parents as volunteering for a less extensive program.[35]

In this case, however, the method was clearly not random assignment and, because the research sample was limited to volunteers, the Consortium team could only speculate that "serious bias" had not been introduced by the selection method. Moreover, the Consortium team indicated that the pre-test IQ scores differed significantly, with the program group scoring 9 points higher (94 vs. 85). Part of this difference may have been due to missing data, a problem in itself, but no evidence was provided to indicate how comparable the two groups actually were. (Other

[35]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 74.

reviewers have described the research approach as a comparison group design.)[36] In other projects, such as the High/Scope Perry Preschool Project, there were some deviations from random assignment (described in chapter 14). As a result, selection bias cannot be ruled out in even the "more nearly randomized" projects.

"*Publication bias.*" The second form of selection bias may have arisen due to 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[37] (see chapter 3), it is unlikely that any long-term follow-up would be considered. Thus, such a study would have been excluded from the Consortium team's analysis, even though it undoubtedly would have weakened the findings.

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."[38]

Sample size. Perhaps the most important contribution of the Consortium Study is the statistical power gained by combining a number of studies with relatively small samples. Most of the studies in the Consortium Study had relatively small samples, with follow-up samples ranging from 77 to 228 in the third follow-up period and from 46 to 219 in the fourth follow-up period. These sample sizes were based on the number of cases with data for at least one outcome, however, so the sample size for any particular outcome may have been much smaller. For example, in the Harlem Training Project (in the third follow-up period), school record data was available for 223 children, but IQ data (in 1976) for only 141 children.

---

[36]*See*, for example, W. Steven Barnett, "Long-Term Effects on Cognitive Development and School Success," in *Early Care and Education for Children in Poverty: Promises, Programs, and Long-Term Results* ed. W. Steven Barnett and Sarane S. Boocock (Albany, NY: State University of New York Press, 1998), 32.

[37]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 Inc., June 1997).

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

The use of meta-analysis led to stronger findings of program effect for some outcomes than would a vote-counting procedure, because—in the individual studies—the small samples often produced insignificant findings even if program and control (or comparison) group differences were fairly large.

Attrition. The Consortium team examined a number of potential attrition-related problems, including the overall rate of attrition; rates of attrition across program and control (or comparison) groups; whether the final samples differed from those who dropped out; whether there was differential attrition, that is, whether the characteristics of program children who dropped out were different across program and control (or comparison) groups; and whether the final program and control samples differed on some characteristics.

The eight studies included in the third follow-up reported data for 68 percent of the original sample in the follow-up.[39] This is a low attrition rate considering the long period between program participation and follow-up, but it nevertheless raises questions about potential attrition-related biases. Moreover, attrition was more serious for some outcomes than others. For example, the IQ findings in the 1976 follow-up were based on just 41 percent of the original sample of the six projects with such data. For grade retention, the sample consisted of 51 percent of the original total of participants (from eight projects). Thus, the risk of attrition-related bias was greater for some outcomes than others.

The Consortium team examined whether there was a difference in the rate at which program and control (or comparison) children dropped out for each of the major outcomes examined. They found no significant differences.

The Consortium team also compared the final follow-up sample to those who dropped out. This analysis affects the degree to which findings can be generalized to the original sample. The characteristics analyzed included maternal education, head of household socioeconomic status, and child pre-test IQ score. Again, because attrition could vary depending on the outcome, separate analyses were performed for each outcome. The Consortium team concluded that there was no "systematic" attrition.

Using a similar procedure, the Consortium team compared dropouts and the final sample by treatment status, that is, program or control (or comparison) group. Again they concluded that there was no "overall differential attrition effect."

The Consortium team also tested for the equivalence of the final program and control (or comparison) groups. Here, they found some statistically significant differences, but these were attributed to "differences in the original samples."

[39]The attrition rate was somewhat lower in the final follow-up, but this section focuses on the third-follow-up, because it included a more detailed discussion of potential attrition-related biases.

The Consortium team clearly paid a great deal of attention to attrition.[40] The lack of statistically significant differences between groups due to attrition is not surprising, however, because the sample sizes of individual projects were small, so the differences would have had to be very large for them to be statistically significant. Despite considerable success in locating many families, any time that attrition approaches 50 percent—as it did for some outcomes—only limited confidence can be placed in the findings.

Data collection. The data collection relied on a wide range of tests, school records, and surveys. The data sources were appropriate for the questions being studied, but their completeness varied significantly from study to study.

Measurement issues. The Consortium team utilized a wide range of outcomes and data sources, including school records, interviews, and standardized test scores. They "made every effort to insure the reliability and validity of our measures at every stage of data collection and processing. The instruments were precoded where possible to minimize errors, videotapes were used to train interviewers at the field sites, and frequent contact was maintained with field supervisors."[41]

The use of school records for information about grade retention and special education placement raises several problems. First, schools vary considerably in the degree to which they assign students who are not meeting school requirements to special education or retain them in grade. If one school relies almost exclusively on grade retention and the other almost exclusively on special education placements, a study of early childhood education programs may show strong effects on one variable, but not the other, depending on the school district. In a meta-analysis, these varying patterns might weaken the findings for the use of any one measure. Yet, even a program that is effective in promoting school performance would probably not show statistically significant effects on the measure (retention or special education placement) that a school rarely uses anyway. The composite measure was designed to address this concern.

Another problem, however, was that, given the relatively long period of follow-up, many of the families had moved to other areas and other schools and were thus exposed to a large variety of school policies. It is unclear whether this led to any systematic bias. In some cases, the Consortium team addressed this problem by comparing program and control children within the same school districts. (Although this method would improve the comparability of the data, it would exacerbate problems related to attrition and statistical power.)

---

[40]For all the attention to attrition, it is disappointing that little attention was given to the comparability of the groups in the first place.

[41]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 20.

To measure cognitive outcomes, the Consortium team relied on a number of nationally recognized tests. For the final analysis of IQ scores, because the children differed considerably in their ages, the Consortium team relied on a vote-counting method and abandoned their use of meta-analysis.

Generalizability. The Consortium team cautions that, "these data derive from a specific period in our nation's history—a specific generation of youngsters and scientists, of popular beliefs, scientific theories, and psychological instruments."[42] Moreover, over 90 percent of the children studied were black. Thus, it would probably be a mistake to apply these findings to children of all races.

The diversity of program approaches, combined with the common outcome measures, gives strength to the study. At the same time, this diversity makes it more difficult to sort out the aspects of the programs that were most effective and the program groups that benefitted most. Thus, the findings cannot be easily generalized to any one approach or target population.

In addition, all of these studies were performed thirty to forty years ago in a different early childhood education environment.

Replication. Other researchers have summarized the early childhood literature without applying meta-analysis and reached similar conclusions.[43] In addition, the Head Start Synthesis and Utilization Project (see chapter 11) used meta-analysis to examine Head Start's short-term impacts. The results tend to be essentially the same.

Evaluator's description of findings. The Consortium team carefully summarizes the pattern of positive impacts they found: "early education programs had significant effects in the four outcome areas studied: school competence, developed abilities, children's attitudes and values, and selected family outcomes."[44] In most cases, the authors' interpretations are based on a careful analysis of the findings, taking care not to generalize beyond the research. Their recommendation about the expansion of such programs, however, seems to go beyond what is warranted:

> The sum of our work indicates that children from low-income families derive measurable educational benefits from diverse well-run early education programs. . . . In addition, early childhood education programs can mean dollar savings to school districts.

---

[42]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): xi.

[43]*See*, for example, Barnett, 1998, 25–50.

[44]Irving Lazar and Richard Darlington, "Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies," *Monographs of the Society for Research in Child Development* 47, no. 2/3 (1982): 1–151.

> Consequently, we believe that such programs should be expanded, whether at the national, state, or local level.[45]

Although these programs may have produced savings, it is unclear whether these savings offset the initial program costs. The Consortium team summarized findings from one such study, but did not examine the reasonableness of the assumptions, identify the beneficiaries of such savings, indicate the time frame over which such savings are produced, or present other information to substantiate an expansion of such programs on the grounds that they save money. Of course, there may also be nonmonetary reasons for expanding such programs.

Evaluator's independence. The Consortium Study was produced by an independent team of analysts in collaboration with the original project investigators.

Statistical significance/confidence intervals. Statistical significance was measured and reported at the 5 percent level.

Effect sizes. Apparently, effect sizes were either not calculated or not reported. Instead, in their analysis of IQ data, the Consortium team pooled the significance levels of comparisons between the mean posttest IQ scores of program and control groups for each project. They assert, "This technique gives an indication of whether, given all available evidence, there is a significant difference in IQ scores between program and control children a given number of years after the program."[46] It does not, however, allow readers to gauge the size of these effects.

Sustained effects. One of the criteria for including studies in the meta-analysis was the availability of long-term follow-up data. Children in the study were between thirteen and twenty-one years of age at the time of the final follow-up.

Benefit-cost analysis. Apparently not performed.

Cost-effectiveness analysis. Apparently not performed.

---

[45]Lazar and Darlington, 66.

[46]Irving Lazar, Virginia Ruth Hubbell, Harry Murray, Marilyn Rosche, and Jacqueline Royce, *The Persistence of Preschool Effects: A Long-Term Follow-Up of Fourteen Infant and Preschool Experiments* (Washington, DC: U.S. Department of Health, Education, and Welfare, September 1977), 47.

Commentary

Irving Lazar*

What is of immediate interest in this chapter is what is missing. The only report which is cited is the first of them, the SRCD Monograph. Missing are any citations of the final report, a volume entitled "As The Twig is Bent" and published by Erlbaum Associates in 1983, or of the chapter on social and motivational effects of preschool intervention in "Education for Values" edited by David McLelland and published by Irvington Press in 1982. Some of the data from the Erlbaum volume were selected for inclusion in this review, but the book as a whole, which integrates the findings, is not cited. [Editors' note: This commentary was written prior to the final version of the chapter which does cite *The Twig is Bent*.]

Let me first address the chapter's implication that we might have "cherry-picked" the studies included in the membership of the Consortium. In 1974, we undertook a thorough review of the literature in early childhood education, in developmental psychology, in compensatory education, and in related fields to identify every study of low-income American children that met certain specific requirements. We looked for studies that had at least one hundred subjects from low-income families because we anticipated a high rate of attrition; studies that had well-defined cognitive goals and curricula that were written and specific; studies in which there were reasonably selected control or comparison groups; and studies in which standard descriptive and measurement data were collected immediately prior to and immediately after the intervention experience.

We identified thirteen studies which met these criteria and invited them to participate. One investigator refused to allow us—or even his federal sponsor—to examine his raw data or observe his intervention. He was not included in the group and, indeed, has essentially vanished from the literature. As it happened, I had visited all but one of the remaining twelve while they were in the midst of their interventions. These visits were in the role of a monitor or as a consultant to the federal or philanthropic organizations that were financing their work. What we found was unambiguous. We expected high attrition rates but, in fact, found over 80 percent of the original subjects, often in over a ten year period. Less than 2 percent of the parents refused to participate. Indeed many of the control/comparison parents had interpreted the occasional testing and interviews as positive interventions in themselves. We could find no systematic differences in the socio-economic and other characteristics of the children we found versus those we did not find. It is notable that while these studies took place in fourteen communities, we found the subjects in forty states and five other countries.

*Irving Lazar is Professor Emeritus, Cornell University.

The major findings were clear: The experimentals were significantly more likely to finish high school than were the controls, were more likely to go on to post-secondary education, were more likely to be employed, and were less likely to be known to juvenile court or child welfare agencies.

There are two findings that speak to a possible explanation of why these programs appear to have sustained positive effects into adulthood. First, the difference between the experimental and control groups in their assignment to special or remedial classes, or retention in grade were significant at the 0.0002 level of probability. This is not an insignificant finding. Second, Beller's systematic interviews with parents and children during their adolescent years found that parental anticipations for the future of their children were strikingly different in the experimental and control groups. Where the parents of the experimentals, enthused by the academic success of their children, saw them becoming doctors, lawyers, engineers, and successful businessmen, the typical expectation of a mother of one of the control girls was "I hope she doesn't marry a man who beats her." Almost the reverse was true of the youngsters. The control youngsters seemed sure they would become successful sports players or movie stars. The experimentals thought they might become auto mechanics or beauticians—actual, rational positive hopes at a time when the average white man with a high school diploma earned more than the average black man with a bachelor's degree.

My explanation of our positive findings over so long a time-span derives principally from the findings I mention in the above paragraph. After all, all of the curricula represented in these studies were apparently effective, given the long-term outcomes. To be sure the most highly structured "academic" programs and the least structured free-play programs were less effective than those in which there was a balance between child-directed and teacher-directed activities; however, these differences were not significant. Similarly, we found that the earlier the intervention, the greater the positive long-term effects. Beller specially designed his study to test whether this was so. I do not believe that preschool education is a vaccination against the intellectual blahs. What I think is at work is a change in the valuing of education by the child and the parents. When three-year-old Suzy came home from preschool able to do things her mother had never seen a three-year-old do, she was enthusiastic about Suzy's accomplishments. The child, in turn, began asking of her parents and siblings that they play the same kinds of games that she played in school. This started a mutually-reinforcing system of interaction, in which the value of learning carried on through the school years. Thus, the control girls went back to school after having a baby, the boys and girls went on to post-secondary education, and both aspired to better jobs and careers than their parents had achieved.

I wonder why you separated the Perry Preschool, which was part of the Consortium research, from the others for a separate chapter. Certainly, from an experimental design point of view, it could be perceived as the "weakest" of the earliest batch of early intervention studies. Originally conceived as a program to "prevent" mental retardation, the children were selected on the basis of their liability to become retarded. The mothers were women who typically had IQ

scores in the lower 70s and the children had similarly low scores on infant tests. The number of subjects was very small—it barely made the cut for inclusion in the Consortium. We later learned that the ninety-two subjects came from only fifty-seven families. Indeed, my original concern about Weikart's first findings was that what we were seeing was simply a case of regression to the mean. It is the concordance of his findings with those of larger and more diverse samples that convinced me that his findings were not a statistical fluke. Certainly Weikart did a thorough job of publishing his findings and marketing his curricula. Unlike all but one other of the Consortium members (Levenstein), he was not an academic and needed to support his family with the marketing of his materials. Still, I think that the ingenuity of Gordon's achieving randomization of his sample, and of Palmer's selecting whole cohorts for his experimental and control groups and going to great lengths to make sure that every child had the same number of sessions, are greater contributions to the field.

Finally, I was disturbed by the authors' apparent belief that IQ scores measure some inherent trait. While most scientists, myself included, prefer numbers to words in understanding nature, we have long recognized that the IQ is, at best, a measure of, if I can coin a word, "middleclassness." Test scores of children under six years of age are notably unreliable predictors. Most tests establish their claims of validity by finding high correlations with Binet scores. The Binet itself claims its validity by the agreement of their scores with teacher judgments. Most of the students in the standardization groups were middle class students, and the items themselves were selected on their ability to distinguish different age levels. The vocabulary score correlates most highly with the total score on the leading tests, and it is high dependent on exposure to words. With the universal availability of television, Neisser has shown a steady increase in IQs across our whole population, largely because our children hear more words and experience richer notions of how and where people live. The declining correlation between performance and IQ is not due to the fading away of knowledge; it is instead a reflection of the irrelevance of IQ scores.