SCHOOL of PUBLIC POLICY

16
The High/Scope Perry Preschool 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/chapter16.html.

16
The High/Scope Perry Preschool Project

The High/Scope Perry Preschool Project, which operated in Ypsilanti, Michigan between 1962 and 1967, sought to break the cycle of poverty and school failure by providing a combination of home visits and a part-time preschool program to three- and four-year-olds. The evaluation of the High/Scope Perry Preschool Project was a pioneering effort to evaluate rigorously the long-term impact of an early childhood education program. Not only did the project's creators implement a program highly regarded by developmental experts, but they also thought ahead to do so in a way that would be conducive to later experimentation, following program participants into the late-1990s and early 2000s, through age forty. Additionally, it was included in the Consortium Study (see chapter 4), which statistically combined findings from selected early childhood education programs that had conducted long-term follow-ups.

Lawrence J. Schweinhart and other staff of the High/Scope Perry Educational Research Foundation associated with the project (the "High/Scope team")[1] evaluated the project using a random assignment design. According to the High/Scope team, the project increased the high school graduation rate and earnings of program participants and reduced criminal behavior. In some cases, however, these findings were inconsistent or varied by the measure used, raising some uncertainty about their validity. Moreover, despite the obvious care the High/Scope team devoted to the evaluation, several major issues related to the random assignment process and data inconsistencies limit the confidence that can be placed in these findings. Although there have been replications of the project, they have not been rigorously evaluated.

Program Design

Program group. The High/Scope Perry Preschool Project targeted three- and four-year-old, low-income children with IQs between 70 and 85. Children were recruited from a predominantly black neighborhood on the south side of Ypsilanti, Michigan. Potentially eligible children were identified by surveying families of students attending the neighborhood elementary schools, referrals by community groups, and door-to-door canvassing. Only three families refused to participate, ruling out self-selection as a problem. At the time of enrollment in the project, 45 percent of the program families were headed by a single parent, 58 percent were on welfare, and in 47 percent of the families neither parent worked. There were five cohorts. Most children in the

[1] In this assessment, there are a number of publications by various members of the High/Scope group. Each publication is fully cited in the footnotes, but, for ease of reading, the institutional affiliation is used in the text.

program group participated for two years, but thirteen children in the first cohort (beginning preschool in 1962) participated for just one year.

Services. Children in the program group attended a preschool program for two and one-half hours a day, five days a week, from October to May. The preschool program, operated by the school district, was center-based, with a staff-to-child ratio of 1:6. The staff was composed of teachers certified to teach in elementary, early childhood, and special education. The curriculum was age-appropriate and designed to enhance cognitive and social development. The staff was particularly interested in Piagetian theory, emphasizing children as active learners. They attended seminars on Piaget and throughout the five years of the program, attempted to introduce Piagetian principles into the curriculum. It was not until the fifth and final year of the program that they were able to consolidate this approach fully. Emphasis was placed on a stable daily routine that included a "plan-do-review" sequence in which children were encouraged to plan and carry out their own activities, while the adults supported them in doing so—asking questions and working with them to extend their activities into developmentally appropriate experiences. In addition, staff carried out weekly home visits, during which teachers involved mothers in the educational process and supplemented the children's preschool education.

The Evaluation. Using a random assignment-like procedure (see "Assessing the Randomization" below for a description of how the procedure varied from random assignment), the High/Scope team assigned fifty-eight children to the program group and sixty-five to the control group.[2] It then followed the children for a number of years, collecting data on a range of outcomes, beginning at program intake at age three or four, with periodic follow-ups through age forty. The evaluators examined data from standardized tests, school records, criminal records, and surveys of both parents and children. Group differences were considered statistically significant at the 5 percent level.[3]

Major Findings

The evaluation found that the High/Scope Perry Preschool Project increased the high school graduation rate and earnings of program participants and reduced criminal behavior.[4] In some cases, however, these findings were inconsistent or varied by the measure used, raising some uncertainty about their validity.

---

[2]The evaluation cost through the age twenty-seven follow-up was $2.4 million.

[3]According to the High/Scope team to "round out the picture" of their findings, they called differences between 5 and 10 percent "nearly significant" and those that are between the 10 and 25 percent level "noticeable." Lawrence J. Schweinhart, Helen V. Barnes, and David P. Weikart, *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27* (Ypsilanti, MI: The High/Scope Press, 1993), 45.

[4]Unless otherwise noted, all findings are from Schweinhart, Barnes, and Weikart.

Cognitive. The evaluation examined impacts on IQ, nonverbal intellectual performance, vocabulary, psycholinguistic abilities, and achievement.

IQ. Intellectual performance was measured using the Stanford-Binet IQ test through age eleven and the Wechsler Intelligence Scale for Children (WISC) at age fourteen (see table 1). By the end of the first preschool year, children in the High/Scope Perry Preschool Project outscored their control group counterparts by about 13 IQ points. These initial gains faded, however, after participation in the project ended. By the time the children were eight years old, they were no longer statistically significant. Similar patterns of "fade out" have been found in many other early childhood interventions.

Nonverbal intellectual performance. Nonverbal intellectual performance was measured using the Adapted Leiter International Performance Scale. The High/Scope team found significant differences between the program and control group after the initial completion of the program, but these gains faded out from ages six to eight. However, by age nine, the differences were statistically significant.

Vocabulary. The children's vocabulary was measured using the Peabody Picture Vocabulary Test (PPVT). The High/Scope team found statistically significant differences between the program and control group after the initial completion of the program, but these gains had faded out by age six and the differences continued to not be statistically significant through age nine.

Pscycholinguistic abilities. The children's psycholinguistic abilities were measured using the Illinois Test of Psycholinguistic Abilities. As with the PPVT, the High/Scope team found statistically significant differences between the program and control groups after initial completion of the program, but the gains had faded out by age six and the differences continued to not be statistically significant through age nine.

Achievement. The High/Scope team found mixed effects on children's scores on the California Achievement Test (CAT) (see table 2). At the ages of seven, eight, and eleven, there were no statistically significant differences between program and control group children on the mean CAT total achievement score, but at the ages of nine and ten, the High/Scope team found that the program group children had significantly higher CAT total achievement scores compared to the control group children. At age fourteen, however, the program group registered significantly higher scores than the control group in total achievement, as well as in each component (reading, arithmetic, and language).[5]

[5]The High/Scope Perry Preschool group scored at the thirteenth percentile, while the control group scored at the sixth percentile on total achievement.

Table 1. High/Scope Perry Preschool Project: Effects on IQ

	IQ scores		
When administered	**Program group**	**Control group**	**Differences (percentage point)**
Program entry (age 3 or 4)	80	79	—
End of first program year	95	84	11
End of second program year	95	84	11
Age 6	91	87	4
Age 7	92	88	4
Age 8	88	87	—
Age 9	87	87	—
Age 10	85	85	—
Age 14	80	81	—

Source: Lawrence J. Schweinhart, Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Barfield, and Milagros Nores, *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40* (Ypsilanti, MI: The High/Scope Press, 2005).
Notes: Only significant differences are reported. "—" indicates difference is not statistically significant at the 5 percent level. IQ was measured using the Stanford-Binet IQ test through age eleven and the Wechsler Intelligence Scale for Children (WISC) at age fourteen.

The High/Scope team hypothesized that one reason for this finding at age fourteen is that the test that year was more difficult and demanding than the achievement tests administered when the children were younger:

> It is puzzling that a large group difference appeared at age 14 on the achievement test, but not on the IQ test. While both kinds of tests were administered by trained testers employed by the High/Scope Foundation, IQ tests were given individually, while the achievement tests were given in groups. The IQ testers personally maintained the attention and persistence of the test-takers—waiting for each response, writing it down, and minimizing the number of items presented which the test-takers could not answer correctly. In contrast, the achievement testers provided only general test instructions at the beginning of each section of the test, monitored completion of the items in a passive way, and made no attempt to shield test-takers from items they could not answer correctly. Thus achievement test performance, but not IQ test performance, was dependent on the degree of attention and task persistence of the person taking the test. We conclude that teenagers who had preschool education showed greater persistence on academic tasks

without continuous attention from teachers.[6]

However, Charles Locurto of the Department of Psychology at the College of the Holy Cross observed in a 1991 analysis that it is unusual for differences in achievement test scores to appear just as the IQ differences fade: "Given the substantial correlation between achievement tests and IQ . . . for achievement test differences to emerge just when IQ differences disappear is, of course, unusual, particularly because there were no significant differences in overall achievement in first grade ($p > .05$) when IQ differences were still present."[7] Locurto raised the possibility that these differences in the Perry Preschool Program arose due to differential attrition.

Table 2. High/Scope Perry Preschool: Effects on Achievement

	Achievement Test Scores		
Age of child (years)	**Program group**	**Control group**	**Difference (percentage point)**
7	94	87	—
8	141	129	—
9	170	148	22
10	226	199	27
11	252	243	—
14	119	98	21

Source: Schweinhart, Montie, Xiang, Barnett, Barfield, and Norest.
Notes: Only significant differences are reported. "—" indicates difference is not statistically significant at the 5 percent level. California Achievement Test.

Although there is generally a positive correlation between IQ and achievement, it is far from unity, so the differential impacts may not be that surprising. The magnitude of attrition, however, differed by group—84 percent (49/58) of the program group took the test, but only 71 percent (46/65) of the control group took it. Hence, the possibility of differential attrition should be considered seriously. As described later, the High/Scope team conducted statistical tests to determine whether attrition or differences in group characteristics could have biased the findings. It concluded that the differences were not due to differential attrition. It is possible that there were

[6]Lawrence J. Schweinhart and David P. Weikart, *Young Children Grow Up: The Effects of the Perry Preschool Program on Youths Through Age 15* (Ypsilanti, MI: High/Scope Educational Research Foundation, 1980), 44.

[7]Charles Locurto, "Beyond IQ in Preschool Programs?" *Intelligence* 15 (1991): 300.

unmeasured differences not accounted for in the assessment of potential attrition-related biases. The strong finding at age fourteen and the pattern of findings, mostly positive but statistically insignificant at younger ages, however, suggests that the intervention had modest effects on achievement test scores.[8]

At age nineteen, the program group outscored the control group on the APL Survey, a test of general literacy. But there was no difference at age twenty-seven.

At the age forty follow-up, the High/Scope team administered the Kaufman Functional Skills Test (K-FAST) to both the program and control groups to "measure study participants' competency in reading and mathematics applied to realistic situations of daily life."[9] They found no statistically significant differences between the program and control groups. However, only 68 percent of study participants completed the test.

School readiness/performance. The High/Scope team found no statistically significant impact on either grade retention or the overall use of special education services. Program participation did result in a reduction in placements for "educable mental impairment" (MI) services of 1.7 years (1.1 vs. 2.8 years), but increased compensatory education placements by 0.6 years (1.0 vs. 0.4 years). Since the MI classes cost more than compensatory education classes, the cost savings from the intervention were larger than suggested by the impacts in years spent in these classes alone.

Overall, the program group had a higher "on-time" graduation rate (65 percent vs. 45 percent),[10] but there were large differences in graduation impacts for males and females. For young females, the difference in on-time graduation rates was 52 percentage points (84 percent for the program group vs. 32 percent for the control group). Among young males, however, the control group actually did better than the program group by 4 percentage points (50 percent vs. 54 percent), although the difference was not statistically significant. Why the program had such a large high school graduation impact for females and not for males is somewhat puzzling, since the preschool experience did not result in larger gains on achievement test scores for females.[11]

---

[8]Even restricting the assessment to age fourteen, the effect size of 0.68 for total achievement suggests a "modest" to "large" effect. Many researchers consider an effect size of 0.2 as small, 0.5 as modest, and 0.8 as large.

[9]Lawrence J. Schweinhart, Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Barfield, and Milagros Nores, *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40* (Ypsilanti, MI: The High/Scope Press, 2005), 65.

[10]There was no statistically significant difference between the groups when the measure was expanded to include high school graduation or its equivalent.

[11]Similar differences in the impact on high school graduation rates by gender have been found in the Early Training Project (see chapter 7).

There were no statistically significant differences on several measures reflecting participation in postsecondary education.

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

Health. At the age forty follow-up, the High/Scope team collected data on a number of health factors. Out of fourteen indicators, only one ("Health stopped respondent from working for 1 or more weeks in the past fifteen years") found a statistically significant difference between the program and control groups.

Behavior. There were no statistically significant findings on a range of behavioral outcomes during childhood and adolescence, including: teacher ratings of personal and school misconduct when the children were ages six to nine (although they were significant at the 0.10 level); self-reports of total misconduct by age fifteen; and juvenile arrests.

Crime/delinquency. Data from criminal records, supplemented by self-reports, showed that the children who attended the preschool had significantly fewer lifetime arrests (2.3 vs. 4.6) at age twenty-seven. This finding appears to be driven by a difference in adult arrests (1.8 vs. 4.0), since there was no difference in juvenile arrests (0.5 vs. 0.6). Many who cite the High/Scope Perry Preschool Project's success in reducing criminal behavior focus on the reported reduction in arrests. However, these results should be interpreted with caution. For one, there were no statistically significant differences on other important measures, such as convictions for adult felonies, the percent sentenced to prison for adult felonies, or the percent who served time in prison or jail. Moreover, there was no statistically significant difference in the total self-reported acts of misconduct by age twenty-seven (18.8 vs. 19.4).

At age forty, the children who attended the preschool had significantly fewer lifetime arrests, but this was driven by differences in adult arrests between the ages of eighteen and twenty-seven; there were no statistically significant differences in the number of arrests when the study participants were juveniles or between twenty-eight and forty. There continued to be no statistically significant differences on other important measures, such as convictions for adult felonies, the percent sentenced to prison for adult felonies, or the percent who served time in prison or jail. Moreover, there was no statistically significant difference in the total self-reported acts of misconduct by age forty.

The relationship between felony arrests and convictions between the two groups is puzzling. For example, at age twenty-seven, the mean number of felony convictions for the program group was 0.7, as was its mean number of felony arrests, suggesting that all arrests resulted in subsequent convictions. In contrast, the mean number of felony convictions was 0.8 for the control group, compared to a mean of 1.5 felony arrests. It is unclear why this relationship is so different, which raises questions about the accuracy of the arrest measure.

It also seems unusual that differences in criminal behavior would tend to manifest themselves mainly when the children became adults. The difference in lifetime arrests for males, in particular, seems anomalous, since there were no differences in juvenile arrests or in high school completion. Indeed, in one of its early reports, the High/Scope team noted:

> There is a wealth of data which demonstrates a strong relationship between school failure and delinquency. As Gold has said: "a major provocation for delinquent behavior is incompetence in the role of student and its adjunct roles in the school."[12]

For males, there was little evidence of greater school failure among the control group (except for being classified as being in programs for mental impairment) or juvenile delinquency, so the higher adult arrest rate is perplexing given this explanation.

Finally, not all early childhood interventions have achieved reductions in delinquency and crime. For example, no such impacts were found for the Abecedarian Project (see chapter 1).[13] Schweinhart offered one possible explanation why the High/Scope project may have been successful in this regard, while some other early childhood programs were not:

> Unlike the High/Scope project, the Abecedarian project did not try to improve children's initiative and responsibility, a likely explanation why it did not reduce crime. Even so, the High/Scope study establishes the real *possibility* of an early childhood program preventing crime, whereas the negative results of the Abecedarian project in this regard do not disprove it.[14] [emphasis added]

Early/nonmarital births. Among females, the program group was more likely to be married at the age twenty-seven follow-up (40 percent vs. 8 percent). Several additional outcomes were significant at the 0.10 level, including a reduction in out-of-wedlock births (1.0 vs. 1.7) and a lower rate of reported abortions (4 percent vs. 23 percent). For males, there was little difference in the outcomes by group and none were statistically significant.

There was not a statistically significant difference in marital status. However, at age forty, the program group was more likely to have had more marriages than the control group. Several additional outcomes were significant at the 0.10 level, including three or more out-of-wedlock births (26 percent vs. 35 percent) and a lower rate of reported abortions (16 percent vs. 46

---

[12]Schweinhart and Weikart, 1980, 13.

[13]Frances A. Campbell, Craig T. Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson, "Early Childhood Education: Young Adult Outcomes from the Abecedarian Project," *Applied Developmental Science* 6, no. 1 (January 2002): 42-57.

[14] E-mail message from Larry Schweinhart to Peter Germanis, May 30, 2001.

percent). For males, there was little difference in the outcomes by group and none were statistically significant.

Most interventions that have directly attempted to reduce out-of-wedlock births and abortions and encourage marriage have achieved little or no impact, so it is surprising that a preschool intervention not so directed seems to achieve such large effects. Moreover, these effects were not found by the Abecedarian Project, even though it offered a longer and more intensive intervention.

Economic outcomes. At age twenty-seven, monthly earnings were higher for the program group than for the control group ($1,219 vs. $766, or 59 percent higher). The difference in annual earnings, however, was smaller (19 percent higher for the program group) and was not statistically significant. The large discrepancy in impacts between the previous month's earnings and the previous year's earnings raises questions about which is the most appropriate measure to use in assessing the program's impact. (For the benefit-cost analysis, described below, the more conservative annual estimates were used.) Self-reported employment histories also showed relatively small differences that were not statistically significant.

For males at age twenty-seven, the High/Scope group found no differences in employment rates in the previous month between the two groups, but those in the program group had significantly higher monthly earnings ($1,368 vs. $830). A significantly higher proportion of females in the program group was employed in the previous month (80 percent vs. 59 percent) and they also had significantly higher monthly earnings ($1,047 vs. $651). Given the correlation between education and earnings, the large increase in male earnings seems anomalous.[15]

At age forty, median monthly earnings were higher for the program group than for the control group ($1,856 vs. $1,308, or 42 percent higher). The difference in annual earnings, however, was smaller ($20,800 vs. $15,300, or 36 percent higher for the program group), but still statistically significant. The discrepancy in impacts between the previous month's earnings and the previous year's earnings raises questions about which is the most appropriate measure to use in assessing the program's impact. (For the benefit-cost analysis, described below, the more conservative annual estimates were used.) Self-reported employment histories also showed

[15]Similarly, Barnett in his benefit-cost analysis wrote:

> Program males attained about the same highest year of education as no-program males; so little or no difference in their earnings might be expected on this account. Thus, it is surprising that the preschool program appears to have had a large effect on males' monthly earnings at the age-27 interview. . . . By contrast, the preschool program's estimated effect on annual earnings for males from age 25 to age 27 is quite small ($1,490 per year) and more consistent with expectations based on educational attainment.

W. Steven Barnett, "Cost-Benefit Analysis," in *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27* (Ypsilanti, MI: The High/Scope Press, 1993), 156.

relatively small differences that were not statistically significant.

At age forty, the High/Scope team found significant differences in employment rates in the previous month between the two groups, driven by male employment (70 percent vs. 50 percent).

Welfare. At age twenty-seven, a significantly lower proportion of program group members reported receiving assistance from the government (15 percent vs. 32 percent) at the time of the interview, and a significantly lower proportion reported receiving any social services[16] in the previous ten years (59 percent vs. 80 percent). According to combined survey responses and social services records, however, there were no statistically significant differences in the receipt of AFDC, food stamps, or General Assistance over the five previous years or in the mean number of months on assistance in the previous ten years. The differences in estimated impacts, depending on the data source (that is, the survey or combined survey and administrative records) and time horizon make it difficult to determine the practical significance of the findings.

At age forty, there were no statistically significant differences in overall receipt of any social services, receipt of social services in the previous seven years, the number of years spent on assistance, or the type of assistance.

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

Benefit-cost findings. Early childhood intervention programs are often justified on the assertion that they produce savings that exceed their costs. W. Steven Barnett, currently a professor at the Graduate School of Education at Rutgers University and also a former research associate at the High/Scope Educational Research Foundation, conducted detailed benefit-cost analyses of the program at ages twenty-seven and forty. He estimated the effects of the High/Scope Perry Preschool Project on participants, taxpayers, and society as a whole. (This review will focus on the findings for taxpayers only.) The analysis involved many detailed, carefully done calculations, and included projections through age sixty-five. As is the case with most benefit-cost studies, however, it is based on many assumptions.

Since program costs are incurred "up front," while some benefits and costs appear only later, the rate at which society is willing to tradeoff future benefits and costs for current benefits and costs (the discount rate) affects the estimated "present value" of benefits and costs. For purposes of this analysis, all estimates are adjusted to 2005 dollars, using a 3 percent discount rate.

Costs. The High/Scope Perry Preschool Project was operated by the public school district.

[16]This included assistance from any of the following programs: AFDC, food stamps, General Assistance, protective services, Medicaid, and public housing.

The program costs were estimated from detailed budget reports originating from the school district for the years in which the program operated, including the costs of instruction (consisting of teacher salaries and fringe benefits), administrative support, overhead, supplies, and psychological screening. The estimated cost per participant for the program was $17,200.[17]

Benefits. Estimated savings were calculated in several categories, including education services, crime, and tax revenues. The amount of savings (or added revenues) was based on the difference in costs (or revenues) incurred by the program group and the control group.

The estimates of savings in elementary and secondary education costs were based on the actual experiences of most study participants and official public school records. The preschool's impact was calculated as the difference in costs between program and control children. The High/Scope Perry Preschool Project was estimated to save $9,565 per participant in education services through K-12 at the age twenty-seven and age forty calculation. These savings arose from reductions in special education placements and fewer costs associated with adult education.[18]
Increased educational attainment prompted some of the participants to enroll in post-secondary education which was estimated to cost $815 per participant in the age twenty-seven calculation and $1,283 per participant in the age forty calculation.

Barnett also estimated the High/Scope Perry Preschool Project's effect on earnings impacts, based on self-reports of annual earnings at ages twenty-seven and forty. He projected earnings through age sixty-five based on educational attainment. (This procedure resulted in a large earnings gain for females, but no gain for males, who showed no impact on educational attainment.) The estimated impact of the program on total compensation through age sixty-five was $35,184 per participant in the age twenty-seven calculation and $73,183 per participant in the age forty calculation. Assuming a 15 percent marginal tax rate, Barnett estimated savings to the government of about 15 percent of the earnings gain.

The largest savings in Barnett's analysis came from reductions in crime, including reductions in costs to victims, the criminal justice system, and for private security. The effects were based on the findings at ages nineteen and twenty-seven. The number of crimes was estimated using national data on the relationship between arrests and crimes committed. Barnett then combined these estimates with national estimates of victim costs. The savings in crime-related costs were projected beyond age twenty-seven using national data on arrests by age and sex. The lifetime savings to taxpayers associated with reductions in crime was estimated at

[17]The one year cost was considerably higher than the cost of Head Start. Barnett explained that this was due to the fact that teachers were paid public school salaries and the preschool's relative small class size.

[18]In some cases, the program increased costs, for example, by keeping some children in school longer or increasing the likelihood that the attend postsecondary schools.

$97,971 in the age twenty-seven calculation and $194,475 in the age forty calculation. In both calculations, most of the savings occurred through age twenty-eight and most reflect a reduction in victim costs.

The estimates of crime-related savings are the most questionable. The costs associated with criminal activity can be divided into two categories: tangible and intangible. Tangible costs, which include property loss, medical costs, and foregone earnings for those injured, were estimated using national data about the relationship between arrests, criminal activity, and costs. This procedure raises several concerns. First, as explained earlier, the program's impact on arrests was much larger than on convictions. If convictions are the more reliable indicator of criminal behavior, using arrest data would exaggerate savings. Second, the ratio of arrests to actual crimes may be greater, or less, in Ypsilanti than elsewhere in the nation. The costs of crime may also be different, although using national data may be the only practical procedure.

Barnett also included an estimate of the intangible costs of crime related to pain and suffering, although these costs are even more difficult to estimate. An "adaptation" of the High/Scope Perry Preschool benefit-cost analysis at age twenty-seven, conducted by Lynn Karoly and her colleagues at the RAND corporation, excluded intangible losses to victims because they are so difficult to measure.[19] In Barnett's analysis, these costs accounted for over 80 percent of the crime-related savings and about 65 percent of all savings. Excluding these costs reduced the overall benefit-cost ratio at age twenty-seven from 7.16:1 to 2.5:1. But intangible costs are not really zero, so the RAND adjustment produced a more conservative estimate of the benefit-cost ratio. A similar analysis that excludes victim costs has not been performed for the age forty cost-benefit analysis.

The final component of Barnett's benefit-cost calculation involved savings in welfare expenditures. Estimates of time spent on welfare were based on self-reported data from the age twenty-seven and forty interviews; costs were based on the state of Michigan's eligibility rules and payment levels for various public assistance programs. Barnett also estimated reductions in administrative costs and made projections of welfare savings beyond age twenty-seven, based on studies of welfare dynamics. Welfare savings through age sixty-five were estimated to equal $4,110 per participant in the age twenty-seven analysis and $3,104 in the age forty analysis, with most accruing through age twenty-seven. As with earnings (and taxes), alternative measures suggest somewhat different (generally smaller) impacts.

Benefit-cost ratio. Table 3 summarizes the benefits and costs of the program for the age twenty-seven analysis.

[19]Lynn A. Karoly, Peter W. Greenwood, Susan S. Everingham, Jill Hoube, M. Rebecca Kilburn, C. Peter Rydell, Matthew Sanders, and James Chisea, *Investing in Our Children: What We Know and Don't Know About the Costs and Benefits of Early Childhood Interventions* (Santa Monica, CA: RAND, 1998), 97–98.

Barnett suggests that High/Scope Perry Preschool Project at the age twenty-seven analysis saved $7.20 for each $1.00 it spent, counting savings to crime victims, and $2.50 for each $1.00 spent excluding these savings. At the age forty analysis, Barnett suggests that the High/Scope Preschool Project saved $12.90 for each $1.00 spent, counting savings to crime victims. (It is unclear the extent of the total savings excluding victim costs for this analysis.) Even excluding projected savings, the benefit-cost analysis suggests that the program has resulted in savings to taxpayers that exceed its costs, although the program did not break even until the children reached their twenties.

Despite Barnett's obvious care and attention to detail, his findings should be interpreted with caution, because, as Barnett acknowledges, the analysis has "important limitations" in that the estimates are "inexact" and "based on assumptions that may be debated or are uncertain." As Barnett also notes, however, "any one of the specific measured benefit estimates could be set to $0, and there would still be a net benefit by age 27."[20]

Karoly and her colleagues, by contrast, estimated the savings at about $25,736 per participant , a 2:1 benefit-cost ratio.[21] (The much lower savings figure is due to the fact that they excluded the intangible savings associated with crime.) They also caution that "the uncertainty caused by small sample sizes must be considered."[22] They estimated that the true savings had a two-thirds chance of being between about $24,000 and $36,000 per participant. Using the more traditional 95 percent confidence interval produced a range of approximately $18,000 to $42,000 in savings which translates to a 1.1:1 to 2.7:1 benefit-cost ratio interval. Such wide intervals suggest that caution should be used in making claims about the precise magnitude of the savings.

[20]Schweinhart, Barnes, and Weikart, 1980, 167.

[21]Karoly et al., 1998, 91.

[22]Karoly et al., 1998, 94.

Table 3. High/Scope Perry Preschool's Estimated Benefits and Costs

	Age 27		Age 40
	Savings for taxpayers/crime victims (Barnett)	**Savings for taxpayers (authors' calculations)**	**Savings for taxpayers/crime victims (Barnett)**
Savings measured at age of study			
K-12 education	$9,565	$9,565	$9,565
Adult/post secondary	-$815	-$815	-$1,283
Taxes	$5,885	$5,885	$13,193
Crime	$68,270	$12,364	$175,522
Welfare	$3,357	$3,357	$3,104
Total measured	$87,892	$30,356	$200,101
Projected savings			
Taxes	$6,428	$6,428	$2,773
Crime	$29,701	$5,448	$18,953
Welfare	$704	$704	$35
Total projected	$36,833	$12,580	$21,761
Total benefits (measured + projected savings)	$123,095	$42,936	$221,862
Program cost	$17,200	$17,200	$17,200
Net present value	$105,895	$25,736	$204,662
Benefit-cost ratio	$7.16/1	$2.50/1	$12.90/1

Source: Adapted from Schweinhart, Barnes, and Weikart; and Lawrence J. Schweinhart, Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Barfield, and Milagros Nores, *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40* (Ypsilanti, MI: The High/Scope Press, 2005).
Note: In 2005 dollars discounted at 3 percent.

Overall Assessment

The findings and evaluation methodology are described in considerable detail in a series of

publications by the High/Scope Educational Research Foundation,[23] thus allowing an in-depth assessment of the evaluation. Despite the obvious care the High/Scope team devoted to the evaluation, several problems related to the random assignment process and data inconsistencies limit the confidence that can be placed in these findings. Although the project has been replicated, its replications have not been rigorously evaluated.

Program theory. The High/Scope Perry Preschool Project was based on the hypothesis that preschool programs could have an effect on intelligence and improve the ability of participants to do well in school.[24] At the time of the project's implementation, this hypothesis was based on current psychological and child development literature:

> Weikart et al. derived the Perry study hypothesis from animal studies on environmental enrichment (Krech, Rosenzweig, and Bennet, 1960; Scott, 1962); from Boom's observation that "50 percent of [variance in intellectual] development takes place between conception and age 4" (1964, p. 88); and from the emerging work of Piaget on the development of the thinking process in young children (Hunt, 1961; Piaget and Inhelder, 1969).[25]

The evaluation tested outcomes that were appropriate within this context.

Program implementation. No implementation problems were reported.

Assessing the randomization. The study enrolled 123 at-risk black preschoolers ages three or four, assigning 58 to the program group and 65 to the control group. An assignment procedure that had elements of randomization was used to assign children to program and control groups. First, children were matched on the basis of initial IQ scores and then each child was randomly assigned to one of two undesignated groups. Second, some matched children were exchanged so that the two groups would be matched on *mean* socioeconomic status, *mean* intellectual performance, and percentages of boys and girls. Third, based on a coin toss, one group was randomly assigned to the program and the other to the control group. There were several departures from random assignment, which may have introduced some bias, although the magnitude and direction of such bias is uncertain.

---

[23]*See*, for example, Schweinhart and Weikart; John R. Berrueta-Clement, Lawrence J. Schweinhart, W. Steven Barnett, Ann S. Epstein, *Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19* (Ypsilanti, MI: High/Scope Educational Research Foundation, 1984); and Schweinhart, Barnes, and Weikart, 1980.

[24]Schweinhart, Barnes, and Weikart, 1980, 8.

[25]Schweinhart, Barnes, and Weikart, 1980, 8–9.

Sixty-four participants were initially assigned to each group, but "then, fearing overall sample attrition, staff transferred two children with single, employed mothers from the program group to the control group, because they were unable to participate in any of the program's classes or home visits."[26] As a result, the difference between the two groups in the proportion of children with a single employed mother was quite large (4 percent vs. 22 percent).[27] The High/Scope team indicated that in the absence of this transfer, the program versus no-program group difference in children with employed mothers would no longer have been statistically significant:

> If the 2 children of employed single mothers had not been transferred from the program group, the program versus no-program group difference in children with employed single mothers, although still noticeable, would have been neither significant nor nearly significant (7% vs. 19%, $p = .141$).[28]

The number of such transfers has been a subject of some debate and concern. The High/Scope team had earlier placed this number higher: "Between three and six children with single parents employed outside the home were transferred from the preschool group to the nonpreschool group because they were unable to participate in the classroom and/or home-visit components of the preschool program."[29] The researchers contend that a "thorough review" of the study's original data found only two such transfers.

Finally, siblings were assigned to the same group, "to prevent the preschool group from indirectly affecting siblings in the no-program group." The number of siblings in each group

---

[26]Schweinhart, Barnes, and Weikart, 1980, 31.

[27]The children in the two families transferred to the control group could be included with the program group for purposes of the analysis. Although this would dilute the effects of the intervention, since it would add nonparticipants to the program group, the effect is likely to be small and would remove the potentially biasing effect of an arbitrary movement of families from one group to another. Schweinhart indicated that when the two transferred children are returned to the program group, the magnitude of group differences increases. Larry Schweinhart, e-mail message to Peter Germanis, March 8, 2001.

[28]Schweinhart, Barnes, and Weikart, 1980, 48.

[29]Lawrence J. Schweinhart, John R. Berruta-Clement, W. Steven Barnett, Ann S. Epstein, and David P. Weikart, "The Promise of Early Childhood Education," *Phi Delta Kappan* 66 (April 1985): 551. Yet a third publication describes this differently, indicating that the children with an employed mother were *exchanged* with their control counterpart and that this occurred "approximately once in each of the five waves" over which children were randomly assigned. *See* Lazar and Darlington, 1982, 75. Similarly, an earlier evaluation report indicates, "in a slight deviation from random assignment techniques, five children were transferred from the study, because they were unable to attend preschool or to participate with their mothers in the home-visit component of the program. These children came from single-parent families in which the mother was employed." *See* Schweinhart and Weikart, 1993, 21.

differed: Of the fifty-eight program children, nineteen were siblings (six pairs, one set of three, and one set of four). Of the sixty-five control group children, twenty-four were siblings (twelve pairs). Thus, eleven of fifty-eight program group children (19 percent) were not randomly assigned nor were twelve of sixty-five control group children (18 percent). Although the High/Scope team pointed out that this treatment of siblings made the sampling unit the family, outcomes were not presented by family, but by child. Nearly 20 percent of the sample was not randomly assigned, possibly making the groups different on a number of characteristics. For example, the program group had one family with three siblings and one family with four siblings, which gave these bigger families disproportionate weight in the analysis. It is unclear whether this introduced any systematic bias, but is an unconventional practice among early childhood intervention evaluations.[30]

These deviations from strict random assignment raise substantial concerns about the comparability of two groups. For example, even though the groups were matched on the basis of gender, the program group had nearly 20 percent fewer boys than the control group (thirty-three vs. thirty-nine), which may be an important factor, considering the large differences in impacts by gender. (Some of this differential may have been due to attrition, described below.) This difference suggests that there may be other important differences as well.

Assessing statistical controls in experimental and nonexperimental evaluations. Despite the use of a random assignment-like procedure, there were a number of differences between the families at the time of entry, although most were not statistically significant differences. Nevertheless, table 4 shows the magnitude of some of the differences.

[30]To avoid these problems, a separate analysis could be undertaken of just the first child in each family assigned to the each group. This would limit the analysis to forty-seven Perry preschool children and fifty-three control children. Although it would reduce the sample size, it would remove the uncertainty created by the assignment process for siblings.

Table 4. Comparison of Selected Baseline Characteristics

Characteristic	Perry Preschool group	Control group
Single mother employed[a]	4%	22%
Father unskilled	28%	40%
Family on welfare	58%	45%
Family in public housing	40%	32%
Mother's birthplace-South	83%	75%
Single parent, female children	32%	58%

Source: Schweinhart, Barnes, and Weikart.
[a] Statistically significant at the 5 percent level.

An important difference was that children in the program group were less likely to have employed single mothers (4 percent vs. 22 percent). In addition, girls assigned to the program group were less likely to live in a single-parent family (32 percent vs. 58 percent), a difference that was significant at the 0.10 level. Although there were few statistically significant differences between the groups, there were others that were relatively large, though not statistically significant. Given the small sample size, this is not surprising, only large differences would be statistically significant. These differences raise some concern. In particular, the larger number of control children with employed mothers led Herman Spitz, former director of the Research Department at the E.R. Johnstone Training and Research Center in Bordentown, New Jersey, to observe that the potential lack of parental supervision may have made the control children "more susceptible to delinquent behavior."[31]

The research findings were "based on simple comparisons of the program group and no-program group, without statistical adjustments to compensate for the effects of background covariates."[32] The High/Scope team conducted supplementary analyses for selected outcomes, to examine whether there were changes in the level of statistical significance after controlling for gender, cognitive ability at age three, socioeconomic status, maternal education, and maternal employment (one-way analysis of covariance). Although the level of statistical significance declined somewhat for most outcomes, the adjustments did not materially affect the number of statistically significant findings, nor the size of the estimated impacts.

[31]Herman H. Spitz, "Were Children Randomly Assigned in the Perry Preschool Project?" *American Psychologist* 48, no. 8 (August 1993): 915.

[32]Schweinhart, Barnes, and Weikart, 1985, 44.

This issue was also addressed by an independent 1982 reanalysis of the High/Scope Perry Preschool Project's data, performed by Irving Lazar, then of Cornell University, and his colleagues as part of the Consortium for Longitudinal Studies (see chapter 4). This reanalysis also examined many of the early impacts of the program on IQ scores, achievement tests, and school performance as part of a larger meta-analysis of early childhood programs.[33] They controlled for a number of background characteristics, such as child's age, sex, and number of siblings, as well as maternal education and family structure. Adding these covariates had only marginal effects on the estimated impacts and level of statistical significance. (They did not, however, control for differences in maternal employment, the most serious difference between the two groups.)

Sample size. The total sample size was 123 children, but analyses by gender were about half that size. With a small sample, large impacts are needed to produce statistically significant findings. Thus, the absence of impacts in some areas does not mean that the program did not affect some outcomes, but that the impacts may have been too small to be detected given the sample size. A small sample also means that differences in baseline characteristics would also have to be very large to be statistically significant, making it more difficult to assess the comparability of the program and control groups.

Attrition. After the transfer of two children to the control group, the program group had sixty-two children and the control group had sixty-six children. Four of the program children moved out before completing the program and one child in the control group died. They were dropped from the sample, leaving fifty-eight children in the program group (thirty-three males/twenty-five females) and sixty-five in the control group (thirty-nine males/twenty-six females).

To avoid violating random assignment, all families originally assigned to a group should have been included in the analysis, regardless of whether they participated. As Robert G. St.Pierre, then vice president and principal associate at Abt Associates Inc., and his colleagues explain:

> This is the standard approach taken in all studies in which families are randomly assigned to alternative treatment groups—once the family is assigned to participate in the study, they are retained in the study and included in the analysis. This approach preserves the integrity of the study design; eliminating families from the analysis (due to a lack of participation) would leave the findings open to many interpretations.[34]

---

[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).

[34]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

Similarly, Frances Campbell, a senior scientist at the Frank Porter Graham Child Development Institute at the University of North Carolina at Chapel Hill, and her colleagues note that, in the Abecedarian Project (see chapter 1), they adopted a similar approach: "An intent-to-treat analysis plan was followed in which each individual who participated in the follow-up was analyzed according to his original preschool or school-age group random assignment, regardless of length of exposure."[35] They indicate that four children left the program before the age of one (that is, very early in the program), but removing them would have "violated random assignment."[36]

Robert Boruch, University Trustee Chair Professor of the Graduate School of Education and the Statistics Department at the Wharton School of the University of Pennsylvania, explains in a 1997 analysis that, even if some of those randomly assigned drop out or do not participate in the program, the relevant rule is: "*Analyze them as you have randomized them.*"[37] It is possible to adjust for dropouts by dividing the experimental impact by the participation rate. Using this approach would have avoided undermining random assignment.

Nevertheless, Schweinhart and his colleagues point out that attrition was "extraordinarily low," as the rate of missing data across all measures was just 4.9 percent.[38] The rate of attrition varied considerably, however, depending on the outcome and the time period studied. For example, with respect to comparison of achievement test scores at age fourteen, data were available for forty-nine (of fifty-eight) program group children, but only forty-six (of sixty-five) control children. Thus, for this test, attrition was 16 percent for the program group, but 29 percent for the control group. This difference creates the possibility of two types of bias: the missing children may have experienced different impacts than those who were measured; and if there was differential attrition, the comparability of the two groups could have been reduced.

How much of a difference do a few missing children make? This question can be crudely answered by comparing trends in some impacts over time. At age eleven, there were seven fewer children in the program group than there were at age ten and again at age fourteen. A comparison of impacts on achievement at ages ten, eleven, and fourteen shows wide differences in

---

Inc., June 1997), 3-3.

[35]Campbell et al., 2002, 14.

[36]An alternative to dropping the four children who moved before completing the program would have been to include them in the follow-up. This might dilute the impact of the program a bit, but at least it would not undermine the integrity of the random assignment. Alternatively, the matched pair from the original assignment could be dropped, although this would reduce sample size.

[37]Robert F. Boruch, *Randomized Experiments for Planning and Evaluation: A Practical Guide* (Thousand Oaks, CA: Sage Publications, 1997), 203. Emphasis in original.

[38]Schweinhart, Barnes, and Weikart, 1985, 37.

achievement between the program and control group at age ten (225.4 v. 199.3), smaller differences at age eleven (252.4 vs. 242.4), and wider differences at fourteen (122.2 vs. 94.5) (when one difference was seven fewer children in the age eleven group). This pattern occurred for the various other achievement tests as well. The data suggest that the achievement gap narrowed by about two-thirds at age eleven (from the year before), but it was possibly the missing seven children that explains most of the narrowing impact, since at age fourteen, the gap had widened again. Or, it may simply be a coincidence. The problem, however, is that such attrition increases uncertainty.

Attrition for most outcomes was quite low. In addition, for many outcomes, the High/Scope group presented data on the characteristics of the follow-up sample.[39] They also conducted statistical tests to determine whether there were any significant differences on various background variables, including sex ratio, initial IQ, socioeconomic status, proportion in single-parent families, and maternal education. There were few differences, leading them to conclude that "differential attrition was not likely to have distorted comparisons of outcomes for the intervention group and the control group for the instruments used at ages 14 and 15."[40]

Similarly, in an independent assessment of attrition conducted in 1982, during the earlier years of the program, Lazar and his colleagues also found little evidence of bias due to attrition. This reanalysis also controlled for some background characteristics, with similar findings.[41]

Data collection. The data collection relied on a wide range of standardized tests, school records, administrative data, and survey results. The data sources were appropriate for the questions being studied.

Measurement issues. Many of the early IQ and achievement tests used in the evaluation are nationally recognized assessment tools. School performance data were based mainly on school records. Most records were obtained from the Ypsilanti school district, but some were from other Michigan school districts, as well as from schools outside of the state. These outcomes may have been affected by school-specific reporting practices, so there could have been some bias due to the fact that children have different teachers, attend different schools, and so on. However, there is no reason to believe there would be any systematic bias that would affect the estimated impacts.

Spitz suggests in a 1986 analysis that participation in the project itself may have influenced teacher decisions about placement. For example, a larger percentage of control group children were classified as mentally retarded than preschool children. He argues, however, that "one must

[39]*See*, for example, Schweinhart and Weikart, 1985, 24–27; Berrueta-Clement et al., 186–188.

[40]Schweinhart and Weikart,1980, 27.

[41]Lazar and Darlington, 1982.

wonder whether the decision not to label these children as mentally retarded was influenced by the fact that they were in the well-known Perry Preschool Project."[42] Schweinhart notes that the project was not well-known at that time, and argues that it was unlikely that "educators would allow a historical event like that to affect their judgments about children they were serving."[43]

Measures of economic outcomes. The data related to employment, earnings, welfare receipt, and other outcomes were based on periodically conducted surveys of the children (or their parents). The interviewer for the age nineteen, twenty-seven, and forty studies was a long-time resident of Ypsilanti and the coach of the high school football team. His familiarity with the area and many of the children resulted in an extraordinarily high response rate of 95 percent. It may also mean that he was aware of which group they participated in, creating a potential source of bias.[44] Schweinhart commented: "Our interviewer sticks to the script and does not bring up whether people participated in a preschool program several decades earlier, nor do the study participants have any reason to do so."[45] Although the interviewer may have kept to the script, there is no way to know for certain how the interviews were conducted and the familiarity of the interviewer with the respondents is highly unusual. (This could also be an issue in regard to information on arrests the young people provided.)

Several issues arise with the use of survey data. Respondents may not have accurately recalled the details of the amount and source of their income, and they might have been more likely to offer socially desirable responses to a data collector they knew.[46] The survey asked about outcomes at the time of the survey as well as over longer periods in the past, ranging from one to ten or more years earlier. In general, people are likely to recall recent events more accurately. As a result, measurement error is likely to be higher for outcomes spanning a longer time frame. On the other hand, outcomes based on the month of the survey can be misleading, because some outcomes can fluctuate much more on a monthly basis. This possibility is suggested when respondents are asked about earnings over the previous month compared to the previous year. For example, at age twenty-seven, using the monthly estimate, the difference was large: $1,219

[42]Herman H. Spitz, *The Raising of Intelligence: A Selected History of Attempts to Raise Retarded Intelligence* (Hillsdale, NJ: Erlbaum, 1986), 106.

[43]Larry Schweinhart, e-mail message to Peter Germanis, May 30, 2001. It is said, however, that Weikart sat in the back of the classrooms of some former Perry Preschool participants, which could have affected both teacher and the student behavior.

[44]Earlier surveys were conducted by interviewers who had no knowledge of group memberships, so "their objectivity may be assumed." *See* Schweinhart and Weikart, 1980, 27.

[45]Larry Schweinhart, e-mail message to Peter Germanis, February 19, 2001.

[46]This would affect both the program and control group and probably did not impart a bias in favor of either group.

for the program group vs. $766 for the control group. But when asked about the previous year, the difference was smaller: $13,328 vs. $11,186. For the control group, the difference between the last month and year is odd, in that the monthly earnings suggest a much lower annual level of earnings. Such discrepancies across various measures raises questions about which is the most valid indicator.

Barnett raises questions about the earnings data for males:

> It is difficult to know exactly what to conclude from the earnings data for males. On the one hand, earnings and employment reports are likely to be more accurate for the current month than for past years. On the other hand, compared with earnings and employment over a 3-year period, a single month's earnings and employment are more likely to be influenced by transitory fluctuations in the economy and unusual personal circumstances. . . .[E]stimation of annual earnings for ages 25 to 27 based on educational attainment of the sample and national data on blacks' earnings by educational attainment produces results that are extremely close to the Perry study's self-reported annual earnings by preschool experience and gender for ages 25 to 27.[47]

Barnett concludes that, "On balance, the small estimated advantage for program males based on reported annual earnings from age 25 to age 27 seems to be the better indicator of cumulative differences through age 27, especially given the program males' lack of advantage in educational attainment."[48] This smaller difference (for males) is justified on the basis of better quality education, rather than more quantity.

The self-reported employment data could have been compared to administrative records. State Unemployment Insurance and welfare records are commonly used in evaluations and would have avoided problems related to respondent recall. Admittedly, administrative records have problems too, such as missing information on those in noncovered employment or for those who move to another state, but having such data available would have increased the credibility of the findings.

Similarly, the program group was slightly more likely than the control group to have used AFDC in the past five years at age twenty-seven (28 percent vs. 26 percent). At the time of the survey, however, the control group was more likely to be on AFDC (17 percent vs. 8 percent). Although it appears that, for welfare use, the High/Scope team consulted both administrative records and survey data, there is no comparison of the two that would help judge the comparability of the findings.

---

[47]Schweinhart, Barnes, and Weikart, 1985, 157.

[48]Schweinhart, Barnes, and Weikart, 1985, 157.

Measures of crime. A number of questions also arise concerning the validity of the criminal records data. First, there were some notable differences between the findings on the survey and data available from criminal records. For example, the number of self-reported arrests by age twenty-seven was considerably smaller than in data from criminal records. For the program group, the mean number of "times picked up or arrested by police by age 27" was just 0.5 (compared to 2.3 in criminal records) and for the control group, it was just 0.9 (compared to 4.6 in criminal records). Although the self-reported data still suggest that High/Scope Perry Preschool Project cut arrests in half, the overall magnitude was considerably smaller and the finding in the self-reported data was not statistically significant. If the survey data are correct, then the magnitude of savings associated with crime would be smaller as well.

Again, an issue may be that the interviewer was well known to the children. The High/Scope team, writing in 1993, offered this explanation for the possible underreporting of arrests on self-reports:

> Possible reasons for such underreporting are (a) memories that become imprecise as the years go by and, in some cases, as arrests accumulate; (b) cautiousness about revealing such information to the interviewer; and (c) the eagerness to please or appear better than one is to the interviewer. Also, the criminal records are cumulative, whereas the interviews represent memories at single points in time; age 27, in particular, may have been well after the peak of individual criminal activity.[49]

This explanation seems logical, but if correct, it raises questions surrounding all the outcomes reported on the age twenty-seven survey, including employment, earnings, and welfare receipt.

There were also some inconsistencies between the data sources. In the self-reported data, there was no difference in the mean number of months on probation (5.6 vs. 5.8 months for the program and control groups, respectively), whereas the criminal records showed a relatively large and statistically significant reduction (3.2 vs. 6.6 months for the program and control groups, respectively). It is unclear why the program group would exaggerate the number of months of probation. This example also illustrates how the data source for essentially the same variable can make a big difference in the size and statistical significance of the effect. These data issues make it difficult to know for certain the size of the effects.

Generalizability. The High/Scope Perry Preschool Project served low-socioeconomic status, black children with IQs between 70 and 85 at the time of entry into the program. The program's key emphasis was on services for mothers who stayed home with their children. Since then, there has since been a dramatic increase in work among all women, including full-time work among single mothers with young children. As Maris Vinovskis, Bentley Professor of History at

[49]Schweinhart, Barnes, and Weikart, 1993, 95.

the University of Michigan, concludes, "Given the small sample size and the particular location of the study in time and place, one wonders just how far we can generalize from this important study."[50]

Moreover, as Edward Zigler, director of the Yale Bush Center in Child Development and Social Policy and the Sterling Professor of Psychology at Yale University, points out in a 1991 analysis, the program was not typical of a public program:

> It is very unlikely that a preschool program mounted in the typical public school will be of the quality represented by the Perry Preschool Project. The program's experimental character ensured that it would be exceptionally well planned, monitored, and managed. Further, the very fact that staff members are participating in an experiment can stimulate and motivate them. For example, researchers worked extensively with the direct child care givers in analyzing and constructing the program, and visiting experts held weekly seminars for the entire preschool staff. Although the consequences of these aspects of the program were not analyzed, their potential effect on program outcome may well have been substantial.[51]

Zigler adds, "I would like to see the outcome of the High/Scope model when mounted by people with less expertise than those employed in the Perry project."[52]

Similarly, Ron Haskins, former administrative director of the Abecedarian Project, cautions against making recommendations based on the study's findings:

> First, none of the other model programs, including those in the Consortium for Longitudinal Studies, produced results as dramatic as those produced by the Perry project on teen pregnancy, crime, welfare, and employment. Second, even if there were stronger evidence from a number of model preschool programs, it would still be necessary to show that these could be generalized to the type of projects that would characterize a program of national scope.[53]

---

[50]Maris A. Vinovskis, "Do Federal Compensatory Education Programs Really Work? A Brief Historical Analysis of Title I and Head Start," *American Journal of Education* 107, no. 3 (May 1999): 196.

[51]Edward F. Zigler, "Formal Schooling for Four-Year-Olds? No," in *Early Schooling: The National Debate*, edited by Sharon L. Kagan and Edward F. Zigler (New Haven, CT: Yale University Press, 1987), 30.

[52]Zigler, 1987, 30.

[53]Ron Haskins, "Beyond Metaphor: The Efficacy of Early Childhood Education," *American Psychologist* 44, no. 2 (February 1989): 279.

Replication. There have been no random assignment replications of the High/Scope Perry Preschool Project. (One replication effort relying on a comparison group design found that children who attended Head Start classes that used the High/Scope educational approach rather than the traditional Head Start approach experienced improved school outcomes and reduced criminal activity.[54]) Any experiment with such highly touted findings should have been replicated several times over. Such replications should be carried out by independent evaluators and should include program modifications, such as testing the impact using staff that are more typical of those found in public preschool programs.

Evaluator's description of findings. The High/Scope Perry Preschool Project is widely regarded as one of the most successful early intervention programs ever tested. The High/Scope publications present all findings, but tend to highlight the positive ones. Often patterns could have been interpreted or presented differently. For example, Locurto wrote:

> Unkindly, we might marry the large number of nonsignificant and unfavorable findings into a different picture of the Perry Project's outcomes. We might argue that preschool training resulted in no differences in school motivation or school potential at the time of school entry, no lasting changes in IQ or achievement test performance. By age 15, preschool children placed no increased value on schooling; indeed, they were somewhat less certain that they would graduate high school. There were no differences in their average grades as compared to former control-group children, in their personal satisfaction with their school performance or their self-esteem. Their parents were no more likely to talk with teachers about school work or to attend school activities and functions than control-group parents. Preschool children were more likely to have been placed in remedial education. By age 19, they were unemployed at a rate equal to that of their control-group counterparts. The average income in both groups was the same as was the percentage of each group that had received public assistance.[55]

The point is that there are many findings, and the High/Scope team tends to limit its focus to the positive ones, rather than presenting a complete picture.

The High/Scope team identified a few highly significant findings in each area, but most were not statistically significant. For example, the relatively large and significant monthly earnings impacts at age twenty-seven were emphasized, whereas the smaller (and not statistically significant) longer-term impacts on annual earnings and five-year annual employment histories were downplayed. While monthly measures may have been subject to less measurement error, monthly estimates are more volatile. Indeed, Barnett, in his benefit-cost analysis, concluded that

---

[54]High/Scope Educational Foundation, "Head Start Study Find Long-Term Impact," August 20, 2000, http://www.highscope.org/research/HeadStartStudy.htm (accessed December 27, 2002).

[55]Locurto, 1991, 303–4.

the annual measure of earnings was the better one to use.

The High/Scope team appears to stretch the evidence. For example, it concludes that the project's findings "suggest that the preschool program affected society, and in particular the early childhood experiences of the next generation, in a basic way."[56] The data on childrearing experiences that the group presented, however, showed few statistically significant differences with respect to the next generation of children. Any conclusions about the effects on the next generation would be little more than speculation, because there were undoubtedly differences between the two groups in the characteristics of those who had children.

Vinovskis made a similar point about the selective presentation of findings in a 1997 analysis:

> In the most recent update of the Perry Preschool cohort, at age twenty-seven, however, the analysts acknowledge the existence of gender differences in the outcomes. Nevertheless, they still downplay the potential policy significance of those differences and emphasize the more positive results. For example, in the important executive summary of this lengthy and highly technical report, the positive results are repeatedly cited and emphasized. But the limitations and the failures of the Perry Preschool Program are not even acknowledged in the executive summary. Although it is understandable that the authors might want to highlight the achievements of their highly innovative and significant program, policymakers and the general public might have benefitted from a more balanced presentation of their overall results.[57]

Evaluator's independence. On the one hand, the analysts of the High/Scope team were employees of or closely associated with the High/Scope Educational Research Foundation. On the other hand, their findings have received extensive review from other sources. For example, the data and analysis were subject to an independent review by Lazar and his colleagues as part of the Consortium for Longitudinal Studies (see chapter 4).[58] The consortium researchers obtained the original data and examined the randomization process, checked for attrition bias, and reestimated many of the findings, with few differences. In addition, the age twenty-seven study was scrutinized, prior to funding, by analysts representing the Ford Foundation and the U.S. Department of Health and Human Services. The age twenty-seven study also had an advisory

[56]Schweinhart, Barnes, and Weikart, 1985, 141.

[57]Maris A. Vinovskis, *History and Educational Policymaking* (New Haven, CT: Yale University Press, 1999), 75-6.

[58]Lazar and Darlington, 1982.

panel of respected social scientists.[59] Finally, Schweinhart addressed many questions raised in the preparation of this study and provided many clarifications and corrections.

Nevertheless, the credibility of the High/Scope evaluation—like any—would be strengthened by an independent review. Indeed, without such a review, there is a danger that the findings could be perceived as biased. For example, Spitz, one critic of the program's research, observes that the foundation has become a business and strong proponent for early education. He cautions that "the High/Scope Foundation and others like it can no longer serve as unbiased sources for the assessment of preschool programs in which they have so large a vested interest."[60] He goes on to describe how the foundation "created" a news event to announce one of its publications, providing information to influential newspapers and even hiring a public relations firm to manage the event. He notes, "The usual safeguard of peer review was circumvented by the press's dissemination of uncritically accepted statements to a public that had neither the information nor the expertise to make an informed evaluation."[61] (See the "Commentary" section below for a reply to Spitz's critiques.)

Similarly, Sam Watson and Laura Black, writing for the Alabama Policy Institute, note in a 2001 analysis:

> The High/Scope curriculum is now one of the most widely used preschool curricula in the country. The High/Scope Foundation exists to promote the use of the curriculum and to train teachers in its use. . . . Relying on the High/Scope Foundation to research the effects of the Perry Preschool Project is somewhat akin to relying on the tobacco companies to research the effects of smoking. Those who consider cancer research funded by the tobacco industry to be unreliable should apply the same skepticism to research conducted by the High/Scope Foundation. This is not to say that either body of research is, in fact, unreliable. It is simply to say that a double standard should not be applied when it comes to evaluating scientific research. The same reasons for skepticism in evaluating the tobacco industry research are equally applicable as we evaluate the research on the Perry Preschool Project.[62]

As far as the current analysis could determine, the High/Scope researchers are honest and forthcoming about their research. These examples illustrate, however, the need for independent

[59]Schweinhart, Barnes, and Weikart, 1985, xi–xii.

[60]Spitz, 1993, 104.

[61]Spitz, 1993, 105.

[62]Sam Watson and Laura G. Black, *From Cradle to Kindergarten: An Analysis of Early Learning Programs* (Birmingham, Al.: Alabama Policy Institute, 2001), 23-24.

research on the effectiveness of the curriculum.

Statistical significance/confidence intervals. Statistical significance was measured and reported at the 1 percent, 5 percent, 10 percent, and 25 percent levels. Differences that were identified as "significant" were significant at the 5 percent level, "nearly significant" differences were significant at the 10 percent level, and "noticeable" differences were significant at the 25 percent levels.

Effect sizes. Effect sizes were calculated by dividing the difference between the program group mean and the no-program group mean by the standard deviation of the entire sample and were reported as standard deviation units.

Some effect sizes, such as those for IQ scores after two years of preschool, were as large as 1.1 SD. By the age of six, however, the effect size for Stanford-Binet IQ scores had fallen to 0.44 SD, and by age eight there was no statistically significant effect on this measure. At age fourteen, effect sizes for reading, math, and language achievement scores ranged from 0.5 to 0.76 SD. Additionally, at age twenty-seven, there was a statistically significant impact on the participants high school grade point average, with an effect size of 0.58 SD. Most reported statistically significant effect sizes fell within the range of 0.4 to 0.8 SD. (See Appendix 1 for a further discussion of effect sizes and their interpretation.)

Compared to effect sizes obtained by other studies included in this volume, these effect sizes were quite large, with impacts often lasting through the age forty follow-up. Thus, rather than focusing on the relative size of these effects, the authors expand their discussion to include impacts on other outcome measures that, under traditional measures, would not be considered statistically significant. For instance, in describing the program's effect on arrests, Schweinhart and his colleagues note, "Compared with no-program females, program females had nearly significantly fewer lifetime arrests, nearly significantly fewer adult arrests, and noticeably fewer arrests for crimes of drug making or dealing."[63] (See above, under "Statistical significance" for the definitions used.)

Sustained effects. The evaluation examined impacts through age forty, more than thirty years after the intervention ended.

Benefit-cost analysis. A benefit-cost analysis was conducted by the researchers, as well as in a separate analysis by researchers at the RAND corporation.

Cost-effectiveness analysis. Apparently not performed.

[63]Schweinhart, Barnes, and Weikart, 1985, 87–89.

Commentary

Lawrence J. Schweinhart and David P. Weikart*

While fair criticism is essential to the scientific method and the advancement of knowledge, unfair criticism undermines the scientific method and presents unhelpful information masquerading as knowledge. This review blends fair and unfair criticisms, a dangerous combination because the fair criticisms lend unwarranted credibility to the unfair criticisms. We do appreciate the opportunity to respond to these criticisms, however, particularly those that Besharov and his colleagues repeat from other sources to which we have never directly replied. In a good story that may even be true, B. F. Skinner is said to have retorted to a persistent questioner, "Science progresses by standing on the shoulders of those who have gone before. But you sir, are standing on my feet."

The High/Scope Perry Preschool Study began during the early 1960s as part of the great social ferment of the time, especially the civil rights movement. Young people—and those of us who began the project certainly were young then—greatly desired to contribute to the correction of injustice. We wanted to make a difference. To an education research team, that meant establishing a carefully designed research project to confirm the hypothesis that a new program contributed significantly to children's development. Today, because of research findings of the past few decades, that course of action may sound obvious; back then it was not. That a study begun by young idealists in a local school district forty years ago can still be found useful today speaks well of the strength of its early design and commitment to detail.

Over the years many professionals and professional organizations – among them, the Consortium for Longitudinal Studies, the American Psychological Association, and the National Mental Health Association—have reviewed the study's design and data. They have placed their confidence in the findings and held up the study as one of the successful longitudinal field-based studies undertaken with random assignment of children to program and no-program groups. This current review, applying ideas gained from four additional decades of development and experience in practical research design and statistical analysis, makes some modest criticisms that raise some interesting questions, but on the whole understates the value that this longitudinal study has offered.

*Lawrence J. Schweinhart is senior research scientist at High/Scope Educational Research Foundation in Ypsilanti, Michigan. David P. Weikart was founder and President Emeritus of the High/Scope Educational Research Foundation, Ypsilanti, Michigan.

The biggest problem with the review is that it repeats hostile criticisms, some based on loose speculation, uncritically, as though the fact that they are published makes them unqualifiedly true and fair. The review would have been stronger had it been more judicious in selecting what criticisms to repeat. Here are the specific criticisms and our responses to each. We also include our responses to several that Besharov and his colleagues make on their own.

Spitz: Reassignment of children of employed mothers to the no-program group

The fact that during their early years their mothers—often the only parent—were employed does raise the possibility that those children lacked supervision and, consequently, were more susceptible to delinquent behavior.[1]

The study's own data disprove this possibility. The mean number of study participant arrests through age 27 was 2.3 for the children of employed mothers, substantially *less* than the 3.8 for the children of mothers who were not employed. Rather than needlessly speculate about how maternal employment might have led to delinquent behavior, Spitz might more productively have asked us to check on the validity of his idea.

Spitz: High/Scope bias

The High/Scope Foundation and others like it can no longer serve as unbiased sources for the assessment of preschool programs in which they have so large a vested interest.[2]

The scientific method serves as an objective process for arriving at the truth despite the biases of the investigators. Indeed, a scientific hypothesis is the codification of the investigator's specific bias. We might well turn the criticism around and say that critics like Spitz could not serve as an unbiased source for the assessment of preschool programs because they seek to disprove the worth of these programs. As a country we have not had a good method of establishing and funding longitudinal studies. We can not blame Spitz if he or others were not available to undertake an evaluation of the High/Scope Perry Preschool Program in 1962, at a time when no one expected much from early childhood education. We are glad that a local public school district staff did find a way to conduct this evaluation to high scientific standards.

Spitz: Lack of peer review

The usual safeguard of peer review was circumvented by the press's dissemination of uncritically

[1]Herman H. Spitz, "Were Children Randomly Assigned in the Perry Preschool Project?" *American Psychologist* 48, no. 8 (August 1993): 915.

[2]Herman H. Spitz, *The Raising of Intelligence: A Selected History of Attempts to Raise Retarded Intelligence* (Hillsdale, NJ: Erlbaum, 1986).

accepted statements to a public that had neither the information nor the expertise to make an informed evaluation.[3]

At each stage of the study, staff prepared comprehensive research reports for dissemination and obtained independent expert reviews before releasing the information to the public. But Spitz ignores the universal logic of a random-assignment study. Part of the study's wide appeal is that people do not need specialized training to understand the idea of a fair comparison.

Locurto and Besharov and his colleagues: Not presenting a complete picture

Unkindly, we might marry the large number of nonsignificant and unfavorable findings into a different picture of the Perry Project's outcomes.[4] *The point is that there are many findings, and the High/Scope researchers tend to limit their focus to the positive ones, rather than presenting a complete picture.*

If we had not presented a complete picture, critics like Locurto would not have had access to the information to present a distorted one. Within our complete presentation, however, we do pay more attention to significant differences because they deserve more attention. They support a hypothesis, indeed in this study they reveal a logically consistent pattern of program effects across various measures over a long period of time. Non-significant differences, on the other hand, do not support the hypothesis for any number of reasons—the program may have had no specific effect, but the measure may have lacked the reliability, validity, or sensitivity to detect the program effect. So no firm conclusions can or should be drawn from them.

Vinovskis: Small sample size and particularity

Given the small sample size and the particular location of the study in time and place, one wonders just how far we can generalize from this important study.[5]

The purpose of statistical analysis is specifically to assess probabilities of general validity based on certain sample sizes. All studies are particular in time and place, and of course longitudinal studies have a particular past. However, these are the concerns a historian like Vinovskis brings to any social scientific study. On the face of it, this study focused on young

[3]Herman H. Spitz, *The Raising of Intelligence: A Selected History of Attempts to Raise Retarded Intelligence* (Hillsdale, NJ: Erlbaum, 1986).

[4]Charles Locurto, "Beyond IQ in Preschool Programs?" *Intelligence* 15 (1991): 295-312.

[5]Maris A. Vinovskis, "Do Federal Compensatory Education Programs Really Work? A Brief Historical Analysis of Title I and Head Start," *American Journal of Education* 107, no. 3 (May 1999): 187-209.

children living in poverty, so it is reasonable to generalize to young children living in poverty. The generalization is strengthened today by similar recent findings from the longitudinal studies of the Carolina Abecedarian Project and the Chicago Child-Parent Centers.

Zigler: Atypicality of the program and school

It is very unlikely that a preschool program mounted in the typical public school will be of the quality represented by the Perry Preschool Project. . . I would like to see the outcomes of the High/Scope model when mounted by people with less expertise than those employed in the Perry project.[6]

Study findings notwithstanding, the preschool program and its public school district *were* typical. Teachers were recruited locally, not nationally, at a time of teacher shortage when virtually no one outside of Michigan had ever heard of the Ypsilanti Public Schools. Others have found effects of the High/Scope model in ordinary program settings, both short-term effects[7] and the key long-term effect on crime[8] in a study that Zigler himself encouraged.[9] It is worth noting that Zigler's view on this question developed after further study. Zigler and Styfco state, "Some Head Start centers match [the Perry program's] level of excellence";[10] "they might very well have similar outcomes over time."[11]

Haskins: Uniqueness of some findings

First, none of the other model programs, including those in the Consortium for Longitudinal Studies, produced results as dramatic as those produced by the Perry project on teen pregnancy,

---

[6]Edward F. Zigler, "Formal Schooling for Four-Year-Olds? No," in *Early Schooling: The National Debate*, edited by Sharon L. Kagan and Edward F. Zigler (New Haven, CT.: Yale University Press, 1987).

[7]Anne S. Epstein, *Training for Quality: Improving Early Childhood Programs Through Systematic Inservice Training* (Ypsilanti, MI: High/Scope Press, 1993); and M.S. Smith, *Some Short-Term Effects of Project Head Start: a Preliminary Report on the Second Year of Planned Variation*, 1970-71 (Cambridge, MA.: Huron Institute, 1973).

[8]Sherri Oden, Lawrence J. Schweinhart, and David P. Weikart, *Into Adulthood: a Study of the Effects of Head Start* (Ypsilanti, MI: High/Scope Press, 2000).

[9]Edward F. Zigler and Sally J. Styfco, "Is the High/Scope Perry Preschool Better than Head Start?" *Early Childhood Research Quarterly* 9 (1994): 269-287.

[10]Zigler and Styfco, 278.

[11]Zigler and Styfco, 283.

crime, welfare, and employment.[12]

It is essential to distinguish between evidence of no effect and no evidence. There is little evidence of program effects on teen pregnancy, crime, welfare, and employment primarily because few preschool program studies, including those in the Consortium for Longitudinal Studies, have ever been in a position to examine them. Being in such a position requires a powerful experimental design, a study duration of over a decade, and an actual look for such effects. Had the High/Scope Perry study not found the long-term effects that it did, it is unlikely those who conducted the Carolina and Chicago studies would have ever looked; and when they looked for very long-term effects, they found them.

Haskins: Generalization to large programs

Second, even if there were stronger evidence from a number of model preschool programs, it would still be necessary to show that these could be generalized to the type of projects that would characterize a program of national scope.[13]

It is not necessary for every preschool program to demonstrate either very-long-term effects or an economic return on investment for it to be a worthwhile social investment. Virtually every civilization in the world considers public schooling a worthwhile social investment without experimental evidence. Every state in the U.S. now invests in public kindergarten without scientific evidence. Only a year of a child's life separates prekindergarten from kindergarten, yet Haskins would raise the bar for prekindergarten, already much higher than kindergarten, higher still. Even the demonstrated *potential* of clear financial return to society ought to be enough for us to strive for high-quality early childhood education for all young children living in poverty.

Besharov and his colleagues: Official versus self-reported arrest counts

If the survey data (collected in participant interviews) were correct, then the magnitude of savings associated with crime would be smaller as well.

The premise here is that study participants' responses to a single interview question are more accurate than the public record. But persons who are often arrested have little reason to precisely count how many times they have been arrested. We only added self-reported arrests into this number in four cases for whom we had not found the corroborating public record, based on the grounds that a study participant would not likely report an arrest that had not occurred; these additions did not substantially affect the findings of numbers of arrests. In general, the study uses

[12]Ron Haskins, "Beyond Metaphor: The Efficacy of Early Childhood Education," *American Psychologist* 44, no. 2 (February 1989): 274-282.

[13]Haskins, 1989.

the best evidence available—public records when they are available, self-reports when public records are not available. Neither source of evidence is perfect, but both are credible when they come together to support a consistent pattern over the years.

Besharov and his colleagues: Need for independent review

Nevertheless, the credibility of any evaluation would be strengthened by an independent review. Indeed, without such a review, there is a danger that the findings will be perceived as biased.

We agree that independent reviews are important and strengthen subsequent work on a study, and we have consistently taken advantage of them. Prior to this statement, Besharov and his colleagues list various independent reviews of the High/Scope Perry Preschool Study: an unusually thorough one by the Consortium for Longitudinal Studies, reviews of the age twenty-seven study proposal by researchers representing the Ford Foundation and the U.S. Department of Health and Human Services, and an advisory panel of four distinguished scientists (Andrew Billingsley of the University of Maryland, Harriette Pipes McAdoo of Michigan State University, Lyle Jones of the University of North Carolina at Chapel Hill, and Donald Campbell of Lehigh University) for the age twenty-seven study. Besharov, Germanis, and Higney might have added that each phase of the study had proposal reviewers and advisory panels and that the monographs include independent reviews of the study by distinguished scientists, published as submitted. Most surprisingly, he overlooks his own decidedly independent review of the study.

Final Comment

In 1962, some thought participation in educational programs at ages three and four would be harmful to disadvantaged children. The High/Scope Perry Preschool Study was carefully designed to insure that, at each point in the child's growth during the school years, the question of program effects could be tracked. We now know that the experience greatly benefitted the children, their families, and society at large. It is possible that some small change in analysis may make some small difference in reported findings. We try to respond to positive suggestions in subsequent phases of the study. We welcome considered criticism and Besharov, Germanis, and Higney have given us some useful ideas. But they also repeat a few hostile, unfair criticisms, giving them weight that they never deserved. We regret that and hope that readers will place them in proper perspective.

Commentary

W. Steven Barnett*

As is so often the case, I find myself in broad agreement with my friend and colleague Doug Besharov, but disagreeing with respect to important details and their implications. I concur that the benefit-cost analysis of the Perry Preschool Program is inexact and uncertain in some respects, and "caution should be used in making claims about the precise magnitude of the savings." These are fair caveats about any benefit-cost analysis and should be emphasized even more when generalizing from specific studies to broad policies. Thus, my benefit-cost analysis using data through age twenty-seven produced an array of benefit estimates under a wide range of assumptions. I would be comfortable stating that the most reasonable conclusion from these results is that: the benefits were large relative to the costs, and this investment had a high rate of return. However, for some purposes it is useful to be more specific. The uncertainty surrounding more specific estimates should be recognized, but it does not follow that this dictates specific estimates of benefits that are smaller than those I highlighted for use in policy analysis.

The benefit-cost analysis rests on the foundation of an analysis of the Perry Program's effects. Questions about the estimated effects thus raise questions about the benefit-cost analysis. Although Schweinhart and Weikart have commented on this part of the review, a few additional comments are in order from the perspective of the benefit-cost analysis. I believe that the basic pattern of program effects is consistent over time and across measures. As I have discussed at length elsewhere, the gradual decline in IQ effects has more to do with changes in what is assessed over time by IQ tests than by changes in the children's abilities.[1] Effect sizes indicate more stability in effects on achievement through the school years. Effects on delinquency and crime may have more to do with social and emotional development, and perhaps executive function, than with achievement or educational attainment. The state of the art in measuring the

*W. Steven Barnett is a professor at Rutgers University and the director of the National Institute for Early Education Research.

[1]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: SUNY Press, 1998), 11–44; and W. Steven Barnett, John W. Young, and Lawrence J. Schweinhart, "How Preschool Education Influences Long-Term Cognitive Development and School Success: A Causal Model," 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: SUNY Press, 1998), 167-184.

precursors of these later effects was limited in the 1960s, so that to some extent this conclusion rests on other, related studies[2].

It should be understood that the benefit-cost analysis depends on a somewhat different set of analyses than those reported in main studies of program effects. These analyses control for gender and utilize much more fine grained data than a simple mean difference between groups. For example, educational costs and benefits are estimated from the precise educational placements and services of each child each year rather than from estimated effects on more general outcomes such as classification as educationally mentally impaired. I also conducted analyses to examine the effects of the difference in mother's employment that may have resulted from a minor departure from random assignment. As it turns out, this difference tends to bias the estimated program effects downward and because it affects boys more than girls tends to reduce estimated effects for boys compared to girls on such outcomes as high school graduation.[3] However, I did not adjust for mother's education because I did not want to be accused of departing from a simple comparison in ways that would increase the estimated program benefits.

It also should be understood that the major benefits included in the analysis are based on effects that were (a) statistically significant in the Perry study and (b) replicated in other studies. An exception is the negative benefit, or cost, of additional higher education because the difference in higher education was not statistically significant in the Perry study. This is another example of the benefit-cost analysis bending over backwards to be conservative. Besharov and colleagues point out that the Perry study presents results for many measures for which no statistically significant differences were found. However, no effect would be predicted on some, and others are low frequency behaviors or events for which the study has little power to detect effects. The benefit-cost analysis incorporates this information by estimating the corresponding costs and benefits as zero. As I discuss below, there is evidence that this leads to underestimation of benefits.

Besharov and colleagues seem to say that the Perry Preschool program's results have not been replicated, but this is true only in a sense so narrow as to be irrelevant. No study can ever be replicated exactly. A replication is inevitably conducted in a different time with different people under different circumstances. Indeed, it may be highly desirable to introduce key differences in replication to increase generalizability and understanding with respect to dosage and other

---

[2]Lawrence J. Schweinhart and David P. Weikart, "The High/Scope Preschool Curriculum Comparison Study Through Age 23," *Early Childhood Research Quarterly*, 12 (2): 117-143; Adele Diamond, W. Steven Barnett, Jessica Thomas, and Sarah Munro, "Preschool Program Improves Cognitive Control," *Science*, 318 (5855): 1387–1388.

[3]W. Steven Barnett, *Lives in the Balance: Age-27 Benefit-Cost Analysis of the High/Scope Perry Preschool Program*, Monographs of the High/Scope Educational Research Foundation (Ypsilanti, MI: High/Scope Press, 1996), 71-121.

treatment parameters, populations served, and context. Many studies confirm that preschool programs can improve children's academic abilities. The randomized trial of the Abecedarian program is useful for what both differences and similarities in results reveal (for an informative comparison see Barnett and Masse, 2007).[4] The rigorous evaluation of the Chicago Child Parent centers can be viewed as testing a Perry Preschool "lite" operating on a large scale. It provides some confidence that the precise configuration of staff and exact curriculum employed by the Perry program in the 1960's are not necessary to produce results that are similar in kind, while indicating that teacher-child ratio does matter. Moving even farther a field, the crime results have been replicated as far away as Mauritius — with even more departures from the program model.[5]

The authors of the review note that some of the benefit-cost analysis assumptions may be considered "heroic." It does not follow that they are incorrect or that errors in them would have much effect on the results. In general, the longer the time horizon the more uncertain an assumption, but the less impact it has on the results as discounting sharply reduces the present value of benefits far into the future. Fortunately, the passage of time provides a direct test of the extent to which the assumptions on which an analysis is based were accurate and in what direction they might have been biased. Comparison of the estimates from the age twenty-seven benefit-cost analysis discussed here with the results of the more recent analysis based on follow-up through age forty indicates that the age twenty-seven analysis was overly conservative and tended to underestimate total benefits.[6]

Besharov and colleagues raise questions about the estimated earnings benefits. The difference between reported earnings in the last month and over the last year was one source of questions. My own view is that people more accurately report income for last month than for last year, especially if workers change jobs frequently and tend to supplement their income with multiple, sometimes casual, jobs. However, the benefit-cost analysis relied on earnings estimates for multiple years, rather than the latest month. It also assumed no earnings benefits for males, despite evidence to the contrary. At the age forty follow-up the longer job and earnings history added to the evidence of positive effects on earnings for males. Thus, the most recent benefit-cost analysis includes estimated earnings benefits for males as well as females. The explanation for the effects on male earnings could be increased academic skills, or improvements in executive function, social skills and behavior, or decreased involvement in delinquency and crime, or some

[4]W. Steven Barnett and Leonard N. Masse, "Early Childhood Program Design and Economic Returns: Comparative Benefit-Cost Analysis of the Abecedarian Program and Policy Implications," *Economics of Education Review* 26:113-125.

[5]Adrian Raine, Kjetil Mellingen, Jianghong Liu, Peter Venables, and Sarnoff A. Mednick, "Effects of Environmental Enrichment at Ages 3-5 years on Schizotypal Personality and Antisocial Behavior at Ages 17 and 23 years," *American Journal of Psychiatry* 160 (9): 1627-1635

[6]Clive R. Belfield, Milagros Nores, W. Steven Barnett, and Lawrence J. Schweinhart, "The High/Scope Perry Preschool Program," *Journal of Human Resources*, 41(1): 162-190.

combination of these.

Several questions are also raised about the estimated crime benefits. One is whether the benefits should have been estimated from arrests or convictions. I view convictions as a less reliable indicator of criminal behavior, given plea-bargaining and other vagaries of the criminal justice system. In addition, in the Perry study differences in sentencing and actual incarceration time correspond to the differences in arrests. Another issue concerns the "intangible" costs of crime, which should not be taken to mean that these are less direct or salient to the victims. Some analysts omit intangible victim costs, but I find this arbitrary and indefensible. A primary reason for concern with crime is the pain and suffering it imposes on victims. In addition, some crimes carry a risk of loss of life. Anyone can conduct a simple thought experiment to judge whether the assumptions of the benefit-cost analysis with respect to victim costs are unreasonably high by asking: "Would I agree to be the victim of a violent assault, robbery, rape, burglary, or other crime in return for my out-of-pocket costs and the estimated the intangible costs?" For example, in today's dollars the intangible cost estimates I used were about $500 per burglary and $8,000 per assault plus another $200 per burglary and $11,000 per assault when death rates are taken into account. I have yet to meet anyone who would take the money.

The reviewers suggest that "taxpayer" benefits provide a more appropriate (or at least a reasonable alternative way) to evaluate the return to preschool education. I may have contributed to confusion on this score by the labels I have used to communicate to a general audience the distinction that economists make between public and private benefits. I have described the public benefits as "taxpayer and crime victim" benefits and the private benefits as those to participants. This is, in fact, an abstract distinction about roles because all of these people are taxpayers and all of them are potential victims of crime. To clearly communicate about the nature and distribution of the various benefits (and to deal with issues like the excess economic burden associated with taxes) one can usefully separate benefits that affect tax revenues from other benefits. However, when assessing whether a program or policy is economically efficient (the ultimate goal of benefit-cost analysis) all benefits are relevant. It makes no sense to ignore the benefits to the program participants. Those benefits are the primary purpose of the program. Similarly, it makes no sense to ignore the benefits of reduced crime. Crime prevention's primary aim is to prevent crime, not to lower the costs of government.

In addition to the empirical evidence cited earlier that the benefit-cost analysis' assumptions have proven to be conservative, there is evidence that important benefits have been omitted from the analysis. From other studies, it may reasonably be expected that future generations would benefit from the improvements in their parents' lives and that future generations would have higher earnings as a result.[7] Benefits might also be expected from

[7]W. Steven Barnett and Clive R. Belfield, "Early Childhood Development and Social Mobility," *The Future of Children*, 16 (2): 73-96.

improvements in the health of the participants, and there is one important health benefit on which we have some evidence. There is not quite a statistically significant effect on smoking in either the Perry Preschool or Abecedarian studies, but the similarity in estimated effects between the two studies is striking. To further investigate the potential effect, I pooled data from the two studies and found that for the combined studies the estimated effect is statistically significant.[8] Even accounting for just the value of increased longevity produces a substantial economic benefit, over $17,000 per participant. If the value of decreased morbidity were added this benefit might be considerably higher. Yet, this is just one of many health benefits that have been linked with education in other studies.[9]

In sum, when judging whether one should be more concerned that the estimated benefits are too high or too low, I suggest that there is more reason to be concerned that they are too low. In evaluating the Perry Preschool program *per se*, this is not of real concern as the benefit-cost ratio exceeds one under even highly pessimistic assumptions. However, it is a concern when one starts extrapolating to other programs and populations and triangulating with estimates from other studies. Weaker programs should be expected to have weaker effects and smaller benefits. Programs for less disadvantaged populations should be expected to have smaller benefits. If economic benefits are roughly proportional to initial impacts, programs with effects that are only 1/10 the size of the Perry Preschool program's might have benefits that were 1/10 of Perry's. It would matter a great deal whether that figure is in excess of $25,000 per child (as I would argue based on recent estimates) or only about $2,300 per child as Besharov and colleagues' most pessimistic estimate might suggest. However, as I am sure they would agree, such extrapolations also require a great deal of caution and should consider a much broader array of evidence that the Perry Preschool study alone.

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/chapter16.html.

[8]Fisher's Exact Test, $p<.05$.

[9]David Cutler and Adriana Lleras-Muney, *Education and Health: Evaluating Theories and Evidence*, NBER Working Paper (Cambridge, MA: National Bureau of Economic Research, June 2006).